September 29, 2022

VIA EMAIL

National Institute of Standards and Technology
Att'n: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20899

RE: Call for Comments AI Risk Management Framework: Second Draft (Aug. 18, 2022)

AMERICAN CIVIL LIBERTIES UNION

National Office
125 Broad Street
18th Floor
New York, NY 10004
aclu.org

Deborah N. Archer
President

Anthony D. Romero
Executive Director

We write in response to the National Institute of Standards and Technology ("NIST")'s August 2022 publication of the "AI Risk Management Framework: Second Draft" ("the framework").[1] We applaud NIST's efforts to seek and incorporate feedback in the development of the AI Risk Management Framework ("RMF"), including changes that reflect feedback on the initial draft framework published in March 2022. We also appreciate NIST's efforts to seek feedback in the development of other resources related to "Trustworthy AI," including the draft report "A Proposal for Identifying and Managing Bias in AI," on which the ACLU submitted comments in September of 2021.[2] In the present comment, we highlight key areas of further improvement for the framework and issues that the ACLU believes NIST must address before the publication of the AI RMF 1.0, currently planned for January 2023.[3]

EXECUTIVE SUMMARY

As the AI RMF 1.0 is finalized in the coming months, the ACLU recommends that NIST address the following issues:

- Communities impacted by AI must play more than a "consultative" role in AI development processes. The framework should situate impacted communities as key decision-makers in the AI lifecycle with actionable protections and concrete resources when AI is deployed.
- Ensuring that AI systems are "trustworthy" requires interrogating and improving the broader structures in which AI systems may be developed and used.

---

[1] Nat'l Inst. of Standards & Tech., *AI Risk Management Framework: Second Draft* (Aug. 18, 2022), https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf [hereinafter "Second Draft Framework"].

[2] Reva Schwartz et al., Nat'l Inst. of Standards & Tech., Spec. Pub. 1270, *A Proposal for Identifying and Managing Bias in Artificial* Intelligence (June 2021), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf; *see also* Am. C.L. Union, *ACLU Comment on NIST's Proposal for Managing Bias in AI* (Sept. 10, 2021), https://www.aclu.org/letter/aclu-comment-nists-proposal-managing-bias-ai.

[3] Nat'l Inst. of Standards. & Tech., *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework (last visited Sept. 28, 2022).

- The framework should include more substantial guidance on evaluating whether to build or deploy AI systems at all, including guidance for considering non-AI alternatives.
- The framework should acknowledge and address competing incentive structures: among organizations that desire profits, among broader societies plagued by inequity and that feature differential tolerance of risk, and within the document itself, which may be overly invested in tech-solutionist ideals.
- Efforts to map, measure, and manage AI risks should include both technical and non-technical considerations and should not overstate or make unwarranted presumptions about AI systems' capabilities.

I. Communities impacted by AI must play more than a "consultative" role in AI development processes. The framework should situate impacted communities as key decision-makers in the AI lifecycle with actionable protections and concrete resources when AI is deployed.

The framework presents many of the communities, organizations, and decision-makers involved in the AI system lifecycle—referred to as "AI Actors"—as audiences for the framework.[4] It recognizes that communities impacted by AI are "AI Actors," whether such impact is direct or indirect.[5] However, in providing guidance to organizations considering building and deploying AI systems, NIST must not relegate end-users and affected entities to a merely "consultative" role.[6] Instead, it is crucial that the framework explicitly name the role that impacted communities play in the AI system lifecycle, and tailor the "actions" and "functions" in Part 2 of the framework to center the experiences and needs of impacted communities as key priorities in the AI system lifecycle. For example, while NIST rightfully emphasizes the importance of diversity in decision-making as part of the AI system lifecycle in the actions associated with Govern-3.1 of the framework, impacted communities must also be included as a core part of that decision-making.

In its description of AI Actors in Part 1 and associated "actions" and "functions" in Part 2, the framework recognizes the activities of "data collection and processing" as a crucial part of AI development.[7] In Figure 2, titled "AI actors across the AI lifecycle," the "Representative Actors" associated with data collection and cleaning include "data scientists, domain experts, socio-cultural analysts, human factors experts, data engineers, data providers, [and] TEVV experts."[8] This set of "representative actors" does not identify who and where data is gathered from, how it is processed, or who processes it.

Data used to train, fine-tune, validate, and evaluate AI and machine learning models often comes from marginalized communities, and neither the collection nor the processing of this data is a neutral act. Our Data Bodies, a research collective focused on data collection and human rights, highlights that data collection can be "dehumanizing," and people whose data is collected for analysis and processing

[4] Second Draft Framework, *supra* note 1, at 4.

[5] *Id.* 28–29.

[6] *Id.* 6 ("The People & Planet dimension of the AI lifecycle represented in Figure 1 presents an additional AI RMF audience: end-users or affected entities who play an important consultative role to the primary audiences. Their insights and input equip others to analyze context, identify, monitor and manage risks of the AI system by providing formal or quasi-formal norms or guidance.").

[7] *Id.* at 6, fig.2.

[8] *Id.* (The acronym 'TEVV' stands for "testing, evaluation, verification, and validation.").

"feel that technologies, people, and other entities are manipulating their narratives for their own ends, especially to criminalize them or their communities."[9] The framework does not adequately address these realities of data collection and processing, nor does it acknowledge that even well-meaning efforts to address algorithmic bias by focusing on diverse data collection or filtering of training data can produce harmful effects when the data collection process is not orchestrated thoughtfully and with care for those whose data is being collected.[10] Beyond that, efforts to address algorithmic bias through additional data collection are not always well-meaning, and can be actively exploitative.[11] The communities that are frequently targeted for AI-related data extraction are often also those most acutely impacted by AI deployment and associated externalities, including environmental impacts.[12] These same communities—disproportionately Black and Indigenous communities—are also disproportionately targeted by other discriminatory and harmful systems, including the criminal legal system and the family regulation system, which in turn increasingly deploy harmful and discriminatory AI systems against these communities.

The framework should identify these dynamics in Part 1, and provide guidance on data collection and processing in light of these concerns throughout Part 2, especially as "actions" in Govern-5, Map-1 and Map-5, Measure-1 and Measure-2, and Manage-2 and Manage-4. Such guidance should provide context on cultivating a "culture of care for the subjects of the datasets . . . throughout collection, development, *and* distribution."[13] This guidance should include consideration of the ways in which data collection may be extractive or exploitive,[14] identification of intentional or unintentional incentives to share data or participate in data collection systems,[15] and situations where compensation for those whose data is collected or used for commercial purposes should be considered.[16] Bender et al. (2021) advocate for more thoughtful and careful data curation efforts, and these kinds of practices, as well as practices to promote informed consent and respect of privacy in data collection, processing,

---

[9] Tawana Petty et al., Our Data Bodies, *Reclaiming Our Data* 28 (June 15, 2018), https://www.odbproject.org/wp-content/uploads/2016/12/ODB.InterimReport.FINAL_.7.16.2018.pdf [hereinafter, "ODB report"].

[10] Algorithmic Just. League, *The Algorithmic Justice League's 101 Overview* (2020), https://assets.website-files.com/5e027ca188c99e3515b404b7/5e332b739c247f30b4888385_AJL%20101%20Final%20_1.22.20.pdf (offering a brief overview of the ways in which data collection "in service of 'inclusion'" can be harmful) [hereinafter, "101 Overview"].

[11] *See, e.g.*, Sidney Fusssell, *How an Attempt at Correcting Bias in Tech Goes Wrong*, The Atlantic (Oct. 9, 2019), https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/.

[12] *See generally* Steven Gonzalez Monserrate, *The Staggering Ecological Impacts of Computation and the Cloud*, The MIT Press Reader (Feb. 14, 2022), https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/ (reviewing the ecological impacts of data centers used to power AI systems).

[13] Amandalynne Paullada et al., *Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning* Research, 2 Patterns 100388, 100394 (Nov. 12, 2021), https://www.sciencedirect.com/science/article/pii/S2666389921001847.

[14] Shakir Mohamed, Marie-Therese Png, & William Isaac, *Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence*, 33 Philosophy & Tech. 659 (Jan. 16, 2020), https://link.springer.com/content/pdf/10.1007/s13347-020-00405-8.pdf.

[15] ODB Report, *supra* note 8, at 28.

[16] Paullada et al., *supra* note 12, at 100396.

and use, should also be described in the framework.[17] These practices are not just important for the communities whose data is collected to power AI systems—they are also crucial to the performance and functioning of AI systems, including to prevent common issues that might hinder a system's reproducibility,[18] and to ensuring that AI developers understand the context in which a dataset operates.[19]

To center communities impacted by data collection and AI deployment, the framework should pair the current focus on explainability and interpretability (including in Part 4.6, Figure 4, and Measure-2.8)[20] with a commensurate focus on contestability and recourse. Explainability and interpretability are important principles, and techniques for explaining or interpreting AI systems have myriad use cases, including helping developers and engineers debug AI models and providing communities impacted by AI systems with information about how those systems work.[21] Though these concepts are frequently referenced in discussions about "Trustworthy AI," a growing literature has illustrated that explainability and interpretability are not panaceas for biased AI systems.[22] We understand that NIST has produced separate documents about this topic,[23] sought comment on one of

---

[17] Emily Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Proc. of the FAccT '21 Conf. on Fairness, Accountability, & Transparency 610 (Mar. 2021), https://dl.acm.org/doi/pdf/10.1145/3442188.3445922; *see also* Eun Seo Jo & Timnit Gebru, *Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning*, Proc. of the FAT* 20 Conf. on Fairness, Accountability, & Transparency 306 (Jan. 2020), https://dl.acm.org/doi/pdf/10.1145/3351095.3372829 (providing resources on data collection practices, including lessons drawn from archival efforts).

[18] Sayah Kapoor & Arvind Narayan, *Leakage and the Reproducibility Crisis in ML-based Science* (July 14, 2022), https://arxiv.org/abs/2207.07048.

[19] *See* Michelle Bao et al., *It's COMPASlicated: The Messy Relationship Between RAI Datasets and Algorithmic Fairness Benchmarks* (updated Apr. 28, 2022), https://arxiv.org/abs/2106.05498 (demonstrating how misguided conclusions can result from misunderstanding the context around a dataset).

[20] *See* Second Draft Framework, *supra* note 1, at 15, 24 tbl.4.

[21] Umang Bhatt et al., *Explainable Machine Learning in Deployment*, Proc. of the FAT* 20 Conf. on Fairness, Accountability, & Transparency 648 (Jan. 27, 2020), https://dl.acm.org/doi/abs/10.1145/3351095.3375624.

[22] *See, e.g.*, Marzyeh Ghassemi, Luke Oakden-Rayner, & Andrew L. Beam, *The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care*, 3 The Lancet Digit. Health J. e745 (Nov. 2021), https://www.sciencedirect.com/science/article/pii/S2589750021002089; *see also, e.g.*, Gabriel Lima et al., *The Conflict Between Explainable and Accountable Decision-Making Algorithms*, Proc. of the FAccT '22 Conf. on Fairness, Accountability, & Transparency 2103 (June 2022), https://dl.acm.org/doi/10.1145/3531146.3534628; Zachary C. Lipton, *The Mythos of Model Interpretability*, ACM Queue (May–June 2018), https://dl.acm.org/doi/pdf/10.1145/3236386.3241340; Amir-Hossein Karimi, Bernhard Schölkopf, & Isabel Valera, *Algorithmic Recourse: from Counterfactual Explanations to Interventions*, Proc. of the FAccT '21 Conf. on Fairness, Accountability, & Transparency 353 (Mar. 2021), https://dl.acm.org/doi/10.1145/3442188.3445899; Bhatt et al., supra note 21; Solon Barocas, Andrew Selbst & Manish Raghavan, *The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons*, Proc. of the FAT* 20 Conf. on Fairness, Accountability, & Transparency 80 (Jan. 2020), https://dl.acm.org/doi/abs/10.1145/3351095.3372830; Aparna Balagopalan et al., *The Road to Explainability is Paves with Bias: Measuring the Fairness of Explanations*, Proc. of the FAccT '22 Conf. on Fairness, Accountability, & Transparency 1194 (June 2022), https://dl.acm.org/doi/10.1145/3531146.3533179.

[23] *See* P. Jonathon Philips et al., Nat'l Inst. of Standards & Tech., NISTIR 8312, *Four Principles of Explainable Artificial Intelligence* (Sept. 2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf; David A.

these documents,[24] and referenced both separate documents in Part 4.6 of the framework.[25] However, neither these separate publications nor the context provided in Part 4.6 of the framework adequately reflect the limits of interpretability and explainability. The framework should address this gap. In particular, NIST should incorporate recognition of the ways in which explanations and interpretations of AI systems can be manipulated to harm impacted communities, especially when the decision-makers deploying AI systems are also those responsible for generating and communicating explanations.[26]

More broadly, impacted communities should always be able to access information about how an AI system works, but explanations must also be paired with the ability to opt out of data collection or algorithmic evaluation without penalty or punishment,[27] the ability to correct inaccurate information,[28] the ability to contest specific decisions or the use of an AI system altogether,[29] and other mechanisms for recourse (especially for those harmed by an AI system).[30] NIST should include in the framework guidance on instituting these principles, and in doing so, could draw on the growing literature focused on operationalizing contestability and recourse as core tenets of algorithmic accountability.[31] These protections relate both to technical aspects of AI systems, as well as the sociotechnical and non-technical structural policies and incentives shaping AI systems, which we discuss further in the next section.

Broniatowski, Nat'l Inst. of Standards & Tech., NISTIR 8367, *Psychological Foundations of Explainability & Interpretability in Artificial Intelligence* (Apr. 2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf.

[24] *See* Nat'l Inst. of Standards & Tech., *Comments Received on Four Principles of Explainable Artificial Intelligence*, NISTIR 8312 DRAFT (updated Apr. 5, 2022), https://www.nist.gov/artificial-intelligence/comments-received-four-principles-explainable-artificial-intelligence-nistir.

[25] Second Draft Framework, *supra* note 1, at 15.

[26] Sebastian Bordt et al., *Post-Hoc Explanations Fail to Achieve Their Purpose in Adversarial Contexts*, Proc. of the FAccT '22 Conf. on Fairness, Accountability, & Transparency 1194 (June 2022), https://dl.acm.org/doi/10.1145/3531146.3533153.

[27] *See* 101 Overview, *supra* note 10.

[28] *See id.*

[29] Henrietta Lyons, Eduardo Velloso, & Tim Miller, *Conceptualising Contestability, Perspectives on Contesting Algorithmic Decisions*, 5 Proc. of the ACM Conf. on Hum.-Comput. Interaction 1 (Apr. 22, 2021), https://dl.acm.org/doi/abs/10.1145/3449180.

[30] *See, e.g.*, Algorithmic Just. League, *Help Prevent, Report, & Redress Algorithmic Harms*, https://www.ajl.org/avbp (last visited Sept. 28, 2022) (describing the Community Reporting of Algorithmic System Harms, or "CRASH" Project of the Algorithmic Justice League).

[31] *See, e.g., id.*; *see also, e.g.*, Shalmali Joshi et al., *Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems* (July 22, 2019), https://arxiv.org/abs/1907.09615; Amir-Hossein Karimi et al., *A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects* (revised Mar. 1, 2021), https://arxiv.org/abs/2010.04050; Daniel N. Kluttz, Nitin Kohli, & Dierdre K. Mulligan, *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, in *After the Digital Tornado: Networks, Algorithms, Humanity* 137–52 (Kevin Werbach ed. 2020), available at https://www.cambridge.org/core/services/aop-cambridge-core/content/view/311281626ECA50F156A1DDAE7A02CECB/9781108426633c6_137-152.pdf/shaping-our-tools-contestability-as-a-means-to-promote-responsible-algorithmic-decision-making-in-the-professions.pdf.

II. Ensuring that AI systems Are "trustworthy" requires interrogating and improving the broader context in which AI systems may be developed and used.

The framework rightly identifies goals for AI development, deployment, and maintenance, but it focuses largely on the specific mechanical aspects of AI systems. Ensuring trustworthy AI will "take more than just setting technical standards,"[32] however, since contextual factors have a significant influence on the success of any AI system in achieving development, deployment, and maintenance goals. The framework currently provides insufficient oversight for contextual factors.

For example, consider an AI system trained on historical data related to one of the many economic or social factors affected by the COVID-19 pandemic (hours worked in office, commute patterns, purchase history, job retention, etc.). In the face of the pandemic, much of the underlying data shifted radically, invalidating many sophisticated predictive models.[33] If a system trained on pre-pandemic data is applied to pandemic-era data, it is unlikely to generalize robustly. For example, it might make invalid inferences, such as over-ordering or under-ordering specific inventory. Worse, if changes in the underlying data are unevenly distributed (as the social costs of the COVID-19 pandemic have been), the system's errors could fall more heavily on a subset of the population. In some cases, the AI system may not even be able to signal to the operators that the system is operating outside the bounds of safe use. Even if the AI system had worked perfectly before the pandemic, the change of context creates significant risks.

While Measure-2 of the framework ("Systems are evaluated for trustworthy characteristics") has a number of salient suggestions for identifying the representativeness and generalizability of AI systems, more attention should be given to ongoing measurement of the environment in which the system operates and its alignment with the expected range of suitable operation.[34] For example, Measure-2.3 recommends measuring system performance or assurance criteria "for conditions similar to deployment setting(s)" but does not describe measuring the deployment setting itself to ensure it is still within the justified environment.[35]

Additionally, decisions about the situations in which AI systems are deployed or not deployed may themselves reinforce a flawed or biased system. For example, many police departments across the country adopted location-based "predictive policing" AI systems targeting crimes such as violent crime and petty theft,[36] and those systems have led to overpolicing of poor, Black, and Brown communities. Meanwhile, the same approach has not been adopted for white-collar crime.[37] Examining which environments and contexts have *not* received investment in AI systems can illuminate social risks and

---

[32] Crystal Grant & Kath Xu, *Public Trust in Artificial Intelligence Starts with Institutional Reform*, ACLU (Sept. 17, 2021), https://www.aclu.org/news/national-security/public-trust-in-artificial-intelligence-starts-with-institutional-reform.

[33] *See* Jeffrey D. Camm & Thomas H. Davenport, *Data Science, Quarantined*, MIT Sloan Mgmt. Rev. (July 15, 2020), https://sloanreview.mit.edu/article/data-science-quarantined/.

[34] *See* Second Draft Framework, *supra* note 1, at 24 tbl.4.

[35] *Id.*

[36] Aaron Sankin, et al., *Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them*, The Markup / Gizmodo (Dec. 2, 2021), https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them.

[37] Brian Clifton, Sam Lavigne, & Francis Tseng, *Predicting Financial Crime: Augmenting the Predictive Policing Arsenal*, The New Inquiry (Apr. 26, 2017), https://whitecollar.thenewinquiry.com/static/whitepaper.pdf.

power structures that will likely make AI deployment problematic in a given context. The RMF should encourage this type of wider reflection.

Some uses of AI systems may also create a feedback loop with their surrounding environment: from the previous example, overpoliced neighborhoods yield more arrests, which would likely lead a "predictive policing" AI system to direct yet more law enforcement to those neighborhoods. For another example, a medical care AI system that prioritizes treatment of disease based on historical costs of medical care for similar cases might observe lower medical expenses for Black patients due to historic (and racist) patterns of medical resource allocation. The overall system's reliance on observations of cost in turn causes it to underestimate the needs of Black patients.[38] Consequently, if that AI system is used to make treatment decisions to prioritize the "sickest" patients, it will recreate the same pattern of underinvestment in care for marginalized communities, which will propagate bias into any future dataset based on subsequent medical histories. Understanding the possible feedback loops between the AI system and its deployed environment is critical in assessing the risks any AI deployment poses.

Another significant part of an AI system's context is the hidden workforce required to mark up data for training and testing. The framework does not sufficiently identify data annotators as AI Actors or represent their experiences and perspectives as part of the AI system lifecycle. Data collected for use in developing and deploying AI systems must often be labeled before it is used, and this labeling is often performed by human data annotators who often face low wages, exploitative working conditions, and retaliation for speaking up.[39] Data annotators should be considered AI Actors in the framework, including as "Representative Actors" associated with data collection and cleaning in Figure 2 of the framework, and as part of "third-party entities" in the AI system lifecycle.[40] To better guide organizations using AI to consider system context and structural factors, NIST should also include considerations related to the working conditions of data annotators in Part 2 of the framework, including in Map-1 (focused on context) and the various aspects of the Govern, Map, Measure and Manage functions focused on third-party data.[41]

Furthermore, the decision-making structures and power relationships in the context around any potential AI deployment can have an effect on the quality and risk of the deployment, regardless of what data sources, models, or training algorithms are proposed. As mentioned in the prior section of this response, if affected people and communities are given a merely "consultative" role at one or two points in the lifecycle of the system, then the decisions made during the actual deployment will fail to reflect those perspectives, potentially leading to harms as well as perceptions that the system is not accountable, even if transparency measures are in place.

These sorts of contextual factors require further consideration and documentation for any AI system with real-world impact. The RMF should encourage review and documentation of context and structures in which the AI deployment is embedded as an important factor in managing AI-related risks.

---

[38] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Population* 366 Sci. 447 (Oct. 25, 2019), https://www.science.org/doi/10.1126/science.aax2342.

[39] Mary Gray & Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (2019); Karen Hao & Andrea Paola Hernández, *How the AI Industry Profits from Catastrophe*, MIT Tech. Rev. (Apr. 20, 2022), https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/.

[40] Second Draft Framework, *supra* note 1, at 6 fig.2; *id.* at 28 (defining "third-party entities" to include "vendors of data . . . and/or systems and related services.").

[41] *Id.* at 17–25, 21 tbl.3.

III. The framework should include more substantial guidance on evaluating whether to build or deploy AI systems at all, including guidance for considering non-AI alternatives.

The RMF does a reasonable job identifying some places where organizations should consider halting AI development or deployment. For example, Part 5 refers to "go/no-go" decisions, Part 4.2 describes the "ability to shut down . . . systems," Part 6.3 references "[o]ptions [that] may include . . . removal of the system of production," and Manage 4.1 recommends elaboration on a "mechanism for . . . decommissioning."[42] However, the RMF in other places seems to imply that termination would be a drastic or extreme measure,[43] and there is insufficient focus on identifying and handling scenarios where abandoning the system is the correct choice. In reality, decommissioning an AI system may often prove to be simple or routine, as when an AI system proves to be unworkable, cost-ineffective, or just marginally more harmful or more opaque than a simpler solution. The RMF's guidance should accustom the reader to the decommissioning of AI systems as a natural development within ongoing efforts to tackle any problem space.

As the RMF already outlines, making the decision—at any stage of the AI lifecycle—to continue with an AI system deployment requires proper accounting and accountability for the risks and costs of the system itself. By the same token, if proper accounting and accountability for an AI system are themselves infeasible, risky, high cost, or non-functional, then a straightforward conclusion is that the system cannot be considered fit for production. It should be pulled, perhaps to be replaced by a simpler, more accountable, and less automated approach.[44] For example, an AI system intended to control the distribution of public benefits—say, by detecting and reducing fraudulent claims—could itself cost millions of dollars just to build and deploy, and then millions more on an ongoing basis to audit, and then still more millions to pay out to people whose benefits were unjustly denied (perhaps incurring legal costs along the way). Not using the system at all could be cheaper overall, even if the sum of benefits distributed ends up being higher.

One problematic factor here is the RMF's explicit "pro-innovation" stance,[45] which is easy to misinterpret as suggesting that high-tech solutions are inherently preferable to lower-tech ones. But this is clearly not the case: for some large classes of tasks, a simple linear regression using a handful of variables can outperform any more sophisticated AI approach.[46]

Throughout a system's lifecycle, regardless of the industry in which the system might be deployed, a careful record should be kept of alternate approaches considered (including AI and non-AI systems alike), their relative merits and flaws, and what rubric was used to decide between different choices. The AI RMF Playbook's Map-1.1 "Actions" list clearly states that AI system design should begin "after non-AI solutions are considered," but this recommendation (like others) should extend to

---

[42] *Id.* at 13, 16, 23, 26 tbl.5.

[43] *E.g. id.* at 9 ("[W]here negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present [ ] development and deployment should cease in a safe manner until risks can be sufficiently mitigated.").

[44] Ryan Calo & Dainielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 Emory L.J. 797 (2021).

[45] *See* Second Draft Framework, *supra* note 1, at 3 ("The AI RMF strives to[b]e . . . pro-innovation.").

[46] Arvind Narayanan, *How to Recognize AI Snake Oil*, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf (last visited Sept. 28, 2022).

ongoing evaluations, rather than applying only before the design phase. The RMF should encourage this kind of review as a persistent, ongoing practice.

There may be additional factors that discourage the use of specific types of AI systems or AI in certain contexts as well. For example, the Minneapolis City Council has banned the use of facial recognition throughout the city government,[47] and the European Commission has proposed a ban on AI in social scoring or children's toys, among other "unacceptable risk" categories.[48] Any organization considering an AI deployment should consider developing and publicly documenting bright lines that they refuse to cross regardless of the regulatory environment.

Encouraging contemplation of decommissioning AI systems is not out of step with industry. Rather, it helps to re-affirm the validity of some high-profile examples of responsible industry behavior. For instance, Amazon deliberately scrapped a hiring AI program that was both inefficient and biased against women.[49] And Microsoft removed from general public use AI image analysis routines that "infer emotional states, gender, age, smile, facial hair, hair, and makeup" because those tools "can be misused—including subjecting people to stereotyping, discrimination, or unfair denial of services."[50]

The fact that third-party vendors (not only Amazon and Microsoft, but also smaller vendors) may be critical suppliers of these systems to other organizations, and that these third parties can drop products, fail, or disappear, raises another point of risk, as acknowledged in Govern-6.2 of the framework.[51] However, the accompanying section of the AI RMF Playbook currently strongly implies that "redundancy" is the main remediating factor for such a failure, even though a third-party service may have a unique API or capability set that cannot be simply swapped out if the vendor goes away or declines to service the product. Consequently, parts of the RMF that deal with systems failure of this kind should also explicitly consider the possibility of decommissioning the AI system when a critical vendor fails.

Given the range of circumstances where abandoning, decommissioning, or replacing with a non-AI system are plausible and reasonable outcomes, the RMF as a whole should reflect these circumstances prominently and encourage AI actors to consider and plan for them. For example, in Figure 1, there appears to be no "off-ramp" to the AI system lifecycle, and in Figure 2 none of the activities include any activity related to decommissioning.

One responsible risk management practice for any engineering system with potential for failure is the "premortem",[52] an exercise where AI actors imagine a failure of the system and describe what could be the cause as part of the system's development and potential deployment. A premortem can also be coupled with a planning exercise that asks what steps would be necessary to tear down the

---

[47] City of Minneapolis, Minn. Ordinance No. 2021-006 (2021) (amending Minn. Code of Ordinances tit. 42, ch. 41), https://lims.minneapolismn.gov/Download/MetaData/20406/2021-006_Id_20406.pdf.

[48] European Comm'n, *Regulatory Framework Proposal on Artificial Intelligence* (June 7, 2022), https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[49] Jeffrey Destin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[50] Sarah Bird, *Responsible AI Investments and Safeguards for Facial Recognition*, Microsoft, https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/ (last visited Sept. 15, 2022).

[51] *See* Second Draft Framework, *supra* note 1, at 20 tbl.2 ("Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.").

[52] Gary Klein, *Performing a Project Premortem*, Harv. Bus. Rev. (Sept. 2007), https://hbr.org/2007/09/performing-a-project-premortem.

system safely when it becomes clear that it is no longer viable. The RMF should, as a pragmatic and responsible measure, encourage its audience to plan for identifying these situations and handling the aftermath. Concretely, an organization administering an AI system (at any stage of its lifecycle) should publicly enumerate any adverse conditions that, if realized, would make them pull the project. And the organization should make explicit plans for how such a decommissioning should play out.

Finally, the use case profiles developed in future versions of the RMF should include example case studies where the ultimate outcome is in fact to abandon the AI system. Including profiles of AI system failures provides a model for the audience of how responsible detection of failure and decommissioning can work. Additionally, if negative use cases are included, they can illustrate how failure to detect problems or failure to plan for decommissioning contributes significantly to the risk of the system. The Benefits Tech Advocacy Hub has collected a small set of case studies of failed algorithmic systems for distributing public benefits,[53] some of which have responsible resolutions and some of which are as of writing unresolved. Another recent report focuses on decisions to cancel deployed AI systems, with recommendations based on patterns observed in the cancellation decisions.[54] The AI Incident Database also includes examples of harms stemming from deployed AI systems.[55] These may all be fruitful sources for real-world example profiles.

The above scenarios illustrate why the RMF should center assessments of when and how systems should be responsibly decommissioned. Such an emphasis would make the RMF more effective in its goal of reducing the risks of AI systems.

IV. The framework should acknowledge and address competing incentive structures: among organizations that desire profits, among broader societies plagued by inequity and that feature differential tolerance of risk, and within the document itself which may be overly invested in tech-solutionist ideals.

Part 1 of the framework grapples with the broader impact of AI, including its possible risks and harms to society, acknowledging that "AI risks—and benefit—can emerge from the interplay of technical aspects combined with socio-technical factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context into which it is deployed."[56] While the framework includes mention of socio-technical factors, the breadth of these factors is not discussed, nor the competing incentive structures that often lead to an overly tech-solutionist viewpoint in the deployment of technological innovations. Given profit motives of companies focused first and foremost on improving shareholder gains, it is unrealistic to expect companies to decide, through voluntary adherence to the risk-management framework, whether they should attempt to seek profit from a tool in which they've invested. Absent a regulatory body or a binding third-party agreement ensuring AI tools' compliance with the framework, incentive structure will continue to favor profit and the resulting externalized harms of technology. These harms caused by untrustworthy technology are meant to be represented in Figure 3 of the framework, but even that figure omits the fact that harm to people, organizations, and systems can also represent harm to the to the very existence of trust in any person,

[53] Benefits Tech. Advocacy Hub, *Case Study Library*, https://www.btah.org/case-studies.html (last visited Sept. 28, 2022).

[54] Joanna Redden, Jessica Brand, Ina Sander & Harry Warne, Data Just. Lab, *Automating Public Services: Learning from Cancelled Systems*, https://www.carnegieuktrust.org.uk/publications/automating-public-services-learning-from-cancelled-systems/ (last visited Sept. 28, 2022).

[55] A.I. Incident Database, https://incidentdatabase.ai/ (last visited Sept. 28, 2022).

[56] Second Draft Framework, *supra* note 1, at 1–2.

organization, or system.[57] The "U.S. is the only established democracy to see a major decline in social trust.[58] Many factors contribute to this, unaccountable technology likely among them.

Another challenge the framework suffers from is its narrowly informed perspective. Part 1.2 describes the purpose of the RMF, mentioning that its use can "assist organizations, industries, and society to understand and determine their acceptable levels of risk."[59] The framework's references in its definitions and policy analysis almost exclusively represent Western perspectives. The policies analyzed are the EOs (U.S.), EU AI Act, and OECD recommendations, relying heavily on OECD and ISO definitions; even considering that the ISO has broader membership than the OECD or EU, there is still a need to reflect perspectives and analyze implications that are not primarily centered on Western or Global North views. If NIST's goal for the framework is that it be "universally" applicable to AI, this lack of analysis of frameworks or policies from the global South represents a major oversight amounting to a social hierarchy. In fact, the framework repeatedly mentions how its use can benefit society without grappling with the fact that "society" represents multifaceted groups and systems, often with competing and potentially even opposed needs. The reality that "society" as a player is inappropriate becomes especially apparent in the discussion of risk. For example, the framework includes that "[r]isk tolerance and the level of risk that is acceptable to organizations or society are highly contextual and application and use-case specific."[60] While the framework acknowledges that, in a society, risk tolerance is contextual and use-case dependent, it fails to acknowledge that who has power over the creation and deployment of a tool—and on whom different tools will cause increased risk—often do not factor as considerations in a tool's deployment. Because societies are heterogeneous, risks within them can be heterogeneous too.

Alluding to this, Part 3.2.2 discusses different types of risks and risk tolerances, "in some cases where an AI system presents the highest risk . . . development and deployment should cease in a safe manner until risks can be sufficiently mitigated. Conversely, the lowest-risk AI systems and contexts suggest lower prioritization."[61] While some systems are inherently high risk, such as those that impact people directly in their livelihood, liberty, family, it is unclear how "highest" and "lowest" risk will or should be decided. This is especially important when data that may be considered "low-risk" in isolation, when combined with other forms of information, becomes "high-risk" data. For example, researchers in one study were able to triangulate the exact identity of subjects who had contributed their DNA to publicly available "deidentified" databases using "low-risk" metadata like the person's state of residency and age.[62] It may also be possible that a tool poses a low risk to users in some locations and in some contexts but higher risk to others. This may especially be true about tools used to predict who will need services related to abortion care in the wake of the *Dobbs* decision when, for some, such care has been outlawed. Additionally, a tool may be considered low-risk at first because of narrow deployment or low market share, but this risk can increase as more faith is placed in the tool to make decisions for more people and market share increases. In this case, the high risk is that tools, perhaps trained on the same types of data or even the same datasets, may consistently be biased against

[57] *Id.* at 8 fig.3.

[58] Kevin Vallier, *Why Are Americans So Distrustful of Each Other*, Wall St. J. (Dec. 17, 2020), https://www.wsj.com/articles/why-are-americans-so-distrustful-of-each-other-11608217988.

[59] Second Draft Framework, *supra* note 1, at 2.

[60] *Id.* at 9.

[61] *Id.* at 9–10.

[62] Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 Sci. 321 (2013), https://www.science.org/doi/10.1126/science.1229566.

certain impacted groups. If there is no alternative to algorithmic decision-making, those groups have effectively no recourse.[63]

V. Efforts to map, measure, and manage AI risks should include both technical and non-technical considerations and should not overstate or make unwarranted presumptions about AI systems' capabilities.

The framework should take care not to overstate the capabilities of AI systems in light of the available evidence. In its current form, the framework manages this only inconsistently. For example, the framework states that "[s]ince AI systems can make sense of information more quickly and consistently than humans, they are often deployed in high-impact settings as a way to make decisions fairer and more impartial than human decision-making, and to do so more efficiently."[64] Elsewhere, though, it states that

> Perceptions about AI system capabilities can be another source of risk. One major false perception is the presumption that AI systems work—and work well—in all settings. Whether accurate or not, AI is often portrayed in public discourse as more objective than humans, and with greater capabilities than general software. Additionally, since systemic biases can be encoded in AI system training data and individual and group decision making across the AI lifecycle, many of the negative system impacts can be concentrated on historically excluded groups.[65]

These statements are largely inconsistent—while the latter excerpt makes an important point about AI systems that should be a central part of the RMF, the former point advances a false narrative that AI works[66] and is, on net, better than human decision-making.

There are many real-world examples of AI being used in ways that evidence does not support. In April of this year, the FDA noted that clinicians were using a tool "intended to aid in prioritization and triage of time-sensitive suspected findings of" a patient suffering from a type of stroke to provide diagnostic information. To curb this off-label use, the FDA issued a letter reminding health care providers of the intended use of this medical AI.

The case of a commercial algorithm that some hospital systems used to allocate care—discussed in Part II, above—provides another example. There, an algorithm trained to predict how much money patients are likely to spend on health care costs associated with a hospital visit was inappropriately used to infer patients' underlying health needs.[67] This erroneous assumption may have had serious consequences for Black patients when hospitals began using the tool to predict which patients would need extra care, and to direct that care to them. The hospitals' assumption about the AI's technical capabilities proved to be false because of socio-technical factors—simply put, the tool was trained on

---

[63] Kathleen Creel & Deborah Hellman, *The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems*, 52 Can. J. of Phil. 26 (2022), available at https://www.cambridge.org/core/journals/canadian-journal-of-philosophy/article/algorithmic-leviathan-arbitrariness-fairness-and-opportunity-in-algorithmic-decisionmaking-systems/3AA0ECA77F8622488E9DB0834287215B.

[64] Second Draft Framework, *supra* note 1, at 12.

[65] *Id.* at 31.

[66] Inioluwa Deborah Raji et al., *The Fallacy of A.I. Functionality*, Proc. of the FAccT '22 Conf. on Fairness, Accountability, & Transparency 959 (June 2022), available at https://doi.org/10.1145/3531146.3533158.

[67] Obermeyer et al., *supra* note 38.

data that encodes sociological factors that make expected medical expenditures a poor proxy for underlying medical needs. Past health data reflects inequities in access to health care for marginalized communities: Of particular relevance here, it registers the reality that Black patients have had less money to spend on health care and less access to this care compared to White patients. The AI's predictions of how much Black patients would spend were therefore lower than predictions for White patients with the same health needs. When these predictions on *costs* were inappropriately used to infer patients' degree of health care *need*, Black patients were interpreted as "less sick" than White patients.

The motivation behind the off-label uses described in the previous examples could be predicted and addressed by encouraging clarity as to the targeted application of an AI tool, and how and why this application may have narrowed in development—perhaps documenting the other applications for the tool found to be improper.[68] Making this documentation available to users of the tool could help them better understand the tool's proper use, including its intended purpose, prospective settings, types of users, and limitations.[69] While those using a tool may be tempted to push the boundaries of its use, clear documentation of the tool's limits, which should be made available to end-users and impacted communities, and be made part of marketing materials, will help potential buyers and users assess the true business or other value of a tool and decide whether it is appropriate for them.[70] Assessing or re-evaluating this value would be perhaps best if done or confirmed by a third party, as a vendor may be incentivized to overstate what their own product can and should do, and the purchaser might be tempted to experiment with off-label use. The framework mentions that a product should "fail safely and gracefully if it is made to operate beyond its knowledge limits."[71] Additionally, the product should inform the user in what way they were attempting to operate the tool beyond its limits. The development of this user interface would force developers to proactively anticipate limitations[72] of the tool but, more importantly, imagine ways in which real-world users will misuse the product and discourage this use.

Another way that inappropriate uses of AI could be foreseen and avoided could be through more of an inclusion of end-users as well as impacted communities incorporated into all stages of the framework, as we discuss above, and especially the "Measure" stage in the framework. While the framework mentions that "the risks or trustworthiness characteristics that will not be measured are properly documented"[73] and that "[t]est sets, metrics, and details about the tools used during" testing, evaluation, validation, and verification ("TEVV")[74] be documented as well, there is no mention of how end-users and impacted communities will be empowered to decide which risks or characteristics will be selected to be measured and which will not be measured. It is mentioned that "[d]omain experts, users, and external stakeholders and affected communities are consulted in support of assessments."[75] This is simply insufficient. Were these groups actively involved in the design of a tool, they could point out the reality that, for example, health care workers, when presented with a triage tool that appears to give diagnostic information, will come to use the tool to aid diagnosis. Or, in the previous example,

---

[68] *See* Second Draft Framework, *supra* note 1, at 22 tbl.3 (encouraging that the "[t]argeted application scope is specified, narrowed, and documented based on established context and AI system classification").

[69] *Id.* (providing that the "[i]ntended purpose, prospective settings in which the AI system will be deployed, the specific set or types of users along with their expectations, and impacts of the system are understood and documented.")

[70] *Id.* at 21 tbl.3 ("MAP 1.3: The business value or context of business use has been clearly defined.").

[71] *Id.* at 24 tbl.4.

[72] *Id.*

[73] *Id.* at 23 tbl.4 (MEASURE 1.1).

[74] *Id.* (MEASURE 2.1).

[75] *Id.* (MEASURE 1.3).

that hospital administrators may be tempted to use a tool that predicts projected money spent at their institution to assume which patients will be most needy. Moreover, actively involving domain experts—including from non-technical fields—in both the design and measurement of AI could help avoid embarrassing and harmful instances of bias. This occurred when Microsoft, whose voice recognition tool was discovered to not function as well for Black users compared to White users, attempted to address the issue by incorporating the feedback of a sociolinguist.[76] The framework alludes to assessing this kind of bias at the "Measure" stage,[77] but it is not entirely clear what the term "computational bias" is intended to mean in this context; and if it is tied to social bias and/or discrimination, then it inherently cannot be a purely computational evaluation.

The TEVV framework should include policy considerations and incorporate perspectives of impacted communities. When performing risk/benefits analyses, technical performance metrics should not automatically be presumed most important. The "AI Deployment" definition in Appendix A should recognize that deployment includes many multifaceted policy decisions that affect performance and outcomes. These may include decisions in which important parameters or thresholds may be altered or created as AI systems move from development to production as well as effects stemming from interactions between AI systems and human decision-makers. One such example occurred in Allegheny County, Pennsylvania, where the County deployed an algorithmic tool in 2016 to help child welfare call screeners make decisions about whether to investigate calls made to the county's hotline. The tool estimates the probability a child will be removed from their home within a certain time period after a referral to child welfare. Recently, researchers found that call screeners sometimes mitigate racial disparities in decision-making that the tool would have created had the call screeners strictly followed its recommendations.[78] As this example illustrates, deployment is not just about technical aspects of moving a technical system into production.

In conclusion, to improve the AI RMF, we recommend that NIST focus on better centering communities impacted by AI systems, interrogating the broader structures around AI systems, including guidance in the framework on considering non-AI alternatives, and recognizing both technical and non-technical aspects as crucial to understanding AI systems' risks. In addition, NIST should address the competing incentive structures presented by these issues throughout the framework. Thank you for considering our views.

Sincerely,

American Civil Liberties Union

---

[76] Natasha Crampton, *Microsoft's Framework for Building AI Systems Responsibly* (June 21, 2022), https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/.

[77] Second Draft Framework, *supra* note 1, at 24 tbl.4 (MEASURE 2.6).

[78] Cheng et al., *How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions*, Proc. of the 2022 CHI Conference on Human Factors in Computing Systems (Apr. 2022), https://dl.acm.org/doi/abs/10.1145/3491102.3501831.