February 2, 2024

NIST RFI: 88 FR 88368

Organization: Massive Data Institute at the McCourt School of Public Policy, Georgetown University

Subject: NIST AI Executive Order

Primary POC: Elissa M. Redmiles, Ph.D., Georgetown University (Elissa.redmiles@georgetown.edu); this response was compiled by the Primary POC Elissa M. Redmiles, but represents contributions from the collaborators listed below.

In collaboration with (including which comments collaborators have contributed to):
Sarah Adel Bargal, Ph.D., Georgetown University [Comments #6, #8 and #11]
Grace (Natalie) Brigham, University of Washington [Comment #7]
Nina Grgic-Hlaca, Max Planck Institute for Software Systems, Max Planck Institute for Research on Collective Goods [Comments #1 and #2]
Tadayoshi Kohno, Ph.D., University of Washington [Comment #7]
Jaron Mink, University of Illinois Urbana Champaign [Comments #1, #3, and #5]
Veronica A. Rivera, Ph.D., Stanford University [Comment #4]
Carmela Troncoso, Ph.D., École Polytechnique Fédérale de Lausanne [Comment #9]
Lucy Qin, Ph.D., Brown University [Comments #6 and #10]
Miranda Wei, University of Washington [Comment #7]

The Massive Data Institute (MDI) at the McCourt School of Public Policy and the Department of Computer at Georgetown University offer the following comments in response to National Institute of Standards and Technology (NIST)'s *Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11).* The following responses specifically address the following assignments listed in the RFI (1) Developing Guidelines, Standards, and Best Practices for AI Safety and Security, and (2) Reducing the Risk of Synthetic Content.

The Massive Data Institute is an interdisciplinary research institute that connects experts across computer science, data science, public health, social science and public policy to tackle societal scale issues and impact public policy in ways that improves people's lives through responsible evidence-based research.

The Department of Computer Science at Georgetown University connects students with the ideas, skills, and opportunities to shape the digital world we live in. Current core research areas include algorithms and theory; security, privacy, and cryptography; and data-centric computing.

Table of Contents

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

In response to 1(a)

Comment #1 addressing "Structured mechanisms for gathering human feedback, including randomized controlled human-subject trials; field testing, A/B testing, AI red-teaming" [NIST 1(a)(1)] & "Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness." [NIST 1(a)(2)]

We encourage the thoughtful inclusion of the general public's perceptions in assessing fairness, risks and/or acceptability of release of novel AI applications. Our prior work finds that people's perceptions about AI are nuanced and multi-dimensional. We found that carefully structured surveys can be utilized to assess lay people's perceptions of fairness, by effectively mapping from expert criteria (e.g., expert assessments of a particular feature included in an AI system) to an overall judgement of fairness with high consistency and accuracy [1]. Such judgements can be used to constrain AI system optimization to optimize both fairness and accuracy [2].

We have similarly used surveys to audit the accuracy of AI-inferred features such as advertising attributes or perceived harm of advertising content [3,4]. Additionally, we have utilized similar human-subject experiments to gather insights about people's reactions to harms caused by the deployment of AI applications. Namely, we investigated how people attribute moral [5] and legal [6] responsibility for algorithmic harms, and how people's reactions to such harms may be moderated by the algorithm's fairness and explainability [7]. It is critical to use such structured methods to assess the impacts of AI on stakeholders at scale as part of continuous, ongoing monitoring.

Per the point in the next section, ensuring demographic diversity and a diversity of lived experiences in participant samples is critical to ensure such assessments accurately capture potential risks and inaccuracies. Including the lived experiences of marginalized may require special sampling, building community advisory boards [8] and/or use of qualitative methods to ensure inclusion of populations that are hard-to-reach via general purpose sampling.

Beyond including perceptions of the general public, it is also important to conduct direct inquiry with professional Machine Learning (ML) system users. For example, we investigated security analysts' perception of recently introduced ML-powered security tools in enterprise incident response to better understand design improvements for future ML tooling [9]. In this case, security analysts felt that ML-tools were not effective enough to be used alone, and thus integrated these tools with alongside other, non-ML tools, in unexpected way (e.g., using signature-based methods to capture known malicious patterns while using ML for novel, or complex patterns). Thus, we recommend that technical evaluation of ML capabilities be considered alongside end-user perceptions and actual usage.

References:

[1] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. https://doi.org/10.1145/3178876.3186138

[2] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '18)*. https://doi.org/10.1609/aaai.v32i1.11296

[3] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M. Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P. Gummadi. 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *Proceedings of the World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1920–1930. https://doi.org/10.1145/3308558.3313666

[4] Muhammad Ali, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, & Piotr Sapiezynski (2023). Problematic Advertising and its Disparate Exposure on Facebook. In *Proceedings of the USENIX Security Symposium (USENIX Security 23)*. 5665–5682. USENIX Association. https://www.usenix.org/conference/usenixsecurity23/presentation/ali

[5] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 235, 1–17. https://doi.org/10.1145/3411764.3445260

[6] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2023. Who Should Pay When Machines Cause Harm? Laypeople's Expectations of Legal Damages for Machine-Caused Harm. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 236–246. https://doi.org/10.1145/3593013.3593992

[7] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2023. Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 372, 1-26. https://doi.org/10.1145/3544548.3580953

[8] Leah Pistorius. 2023. Designing new sex education resources with trans and queer youth. Human Centered Design and Engineering, University of Washington, 3/31/2023. https://www.hcde.washington.edu/news/designing-sex-ed-with-trans-queer-youth.

[9] Jaron Mink, Hadjer Benkraouda, Limin Yang, Arridhana Ciptadi, Ali Ahmadzadeh, Daniel Votipka, and Gang Wang. 2023. Everybody's Got ML, Tell Me What Else You Have: Practitioners' Perception of ML-Based Security Tools and Explanations. In *Proceedings of the IEEE Symposium on Security and Privacy (IEEE S&P '23)*. 2068-2085.

Comment #2: "The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve" [NIST 1(a)(1)]

In Grgic-Hlaca et al. 2022 "Dimensions of Diversity in Human Perceptions of Algorithmic Fairness" (ACM EAAMO 2022 New Horizon's Paper Award Winner) we investigated the effect of demographics and personal experiences on people's perceptions of procedural fairness [1]. Specifically, we investigated the fairness of using particular features in the context of the COMPAS algorithm used to assess recidivism risk; the outputs of this algorithm are used as part of bail decisions. The results of this analysis serve to inform the dimensions of diversity that are important to consider when collecting opinions from lay people, or even experts, on the risks and fairness of a particular algorithmic scenario.

Prior work in organizational and social psychology has found that moral judgments about fairness vary across people with different socio-demographic backgrounds [2] and lived experiences [3], and that fairness judgments may exhibit egocentric [4] and system-justifying [5] patterns. Motivated by this research, we examined the effect of:
expertise relevant to the algorithmic decision task – having heard of the COMPAS scenario, being employed in the legal profession, having friends or family employed in the legal profession, having attended a bail hearing, or having served on a jury
the participants' sociodemographics -- age, gender, race, education, political leaning
on the perceived fairness of using eight algorithmic features: the defendant's number of prior charges, the charge description, the charge degree, the defendant's number of juvenile felonies, number of juvenile misdemeanors, and the defendant's age, gender and race.

We find that very closely related expertise / experiences – such as, having attended a bail hearing – significantly impacts perceived fairness of features used in an algorithm deployed in a bail setting. The fact that only experience with bail hearings was relevant **underscores the importance of including those with expertise very closely specialized to the decision context at hand.** This impact of personal experiences on perceptions of algorithmic fairness is in line with findings on perceptions of fairness more broadly. To quote our paper, "Alesina and Giuliano [3] found that factors related to a person's past experiences, such as experiencing

unemployment and personal traumas, are positively correlated with their support for wealth redistribution. Similarly, Margalit [74] found evidence that economic shocks, such as job loss or a sharp drop in income, tend to increase support for more expansive social policies. Cassar and Klein [15] found that participants who experienced an economic failure in a lab experiment were more likely to favor redistribution, even in the absence of personal monetary stakes."

Additionally, **demographics – specifically political leaning – are necessary to consider in the composition of governance bodies**. To quote our paper, "consistent with prior findings in Moral Foundations Theory [27, 39] and research on diversity in perceptions of algorithmic fairness [45]" we find that the political views of participants in our study significantly correlated with their fairness judgements for most of the algorithmic features we studied. In brief, "the more conservative an individual is, the more fair they perceive using most features for bail decisions." Demographic factors like age and race may proxy for political leaning. Indeed, in our statistical analyses, when we did not control for political leaning, we found a significant association between age and perceived fairness of several algorithmic features we studied. However, once we controlled for political leaning, "we found no evidence of demographic factors having a consistent significant association with algorithmic fairness judgments. Our finding is in line with the work on perceptions of algorithmic fairness by Araujo et al. [6] and Wang et al. [103], who also found little effect of demographic factors."

References:

[1] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2022. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 21, 1–12. https://doi.org/10.1145/3551624.3555306

[2] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130.

[3] Alberto Alesina and Paola Giuliano. 2011. Preferences for redistribution. In *Handbook of social economics*. Vol. 1. Elsevier, 93–131.

[4] Leigh Thompson and George Loewenstein. 1992. Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes*. 51, 2 (1992), 176–197.

[5] John T Jost. 2019. A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. British Journal of Social Psychology 58, 2 (2019), 263–314.

Comment #3 "Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users)" [NIST 1(a)(1)]

The role of end users. We may seek to make end users aware of the presence of generated content (e.g., deepfakes) or train them to detect such content. However, such mitigations may themselves may cause harm.

- Negative effects of warning messages: antisocial behavior towards authentic profiles. Our research [1] finds that end-users who receive warning messages about the existence of deepfakes or are shown features that may be used to distinguish deepfakes take slightly more protective behavior. However, these warnings also lead to the misidentification of real images and text as deepfakes, and disengagement from these profiles. Gender and racial stereotypes may influence these mistakes: multiple participants explicitly noted that the image and name of profiles with real content appeared to be mismatched with stereotypical gender and racial identities. Unexpected combinations led to an incorrect belief that content was deepfaked. Similar mistakes occur when the real image or text deviates from people's assumptions about what real content will look like (e.g., one participant believed that a pictured shine on a real image's tooth was the result of an algorithmic mistake rather than the reflection of cavity filling).
- Negative effects of encouraging detection: biased detection. To better understand how the use of racial and gender stereotypes may play a role in the misidentification of deepfakes, we ran an experiment [2] mimicking the moderation of artificial content in which end-users are shown a set of profiles and asked to select which profiles are deepfakes. We find statistical evidence of biases. Specifically, profiles of Black women and Black men are found to have significantly less perceived artificiality compared to profiles of white women and white men.
    - Mitigations:
        - Any mechanisms that attempt to mitigate the risk of deceptive generative AI using end-users (or lightly trained content moderation teams) should be continuously monitored for bias and for chilling effects on organic (non-generated) content.
        - We also found that users who share the same identity as the assessed profile are significantly less likely to misclassify them. Thus, it is important to ensure moderation teams include a diversity of identities.
        - We find that biased perceptions of attacker strategies & AI, as well as stereotypes, can influence people to incorrectly classify real content as artificial. We highlight that these biases are not just due to stereotypes, but also from upstream perceptions of algorithmic biases: since ML models poorly represent Black women and men, users believe that these identities are less likely to be deepfaked. Thus, education correcting mental models may be helpful.

The role of professional content moderators. The bias effects we observed in [2] hold even when statistically modeling only participants in our experiment who had professional experience moderating content. Thus, trainings that mitigate these issues should be designed, evaluated, and deployed.

The role of AI developers & deployers. When considering whether AI developers and deployers can manage the risks and harms of generative AI, it is also essential to consider how the sociotechnical context and pressures of the organizations they are a part influence these defenses [3]. We interviewed 21 AI developers about barriers they face in mitigating adversarial machine learning threats (a related but orthogonal ML concern to generative AI) [4]. Beyond issues with technical mitigations, we find several sociotechnical barriers that are likely to impede AI actors in mitigating harms: (1) a lack of training in adversarial thinking [5], (2) difficulty in assessing vulnerabilities and the need for mitigations, (3) a lack of communication and collaboration between teams responsible for different aspects of the machine learning development pipeline, and (4) organizational incentives that conflict with mitigation efforts. As a result, when encouraging the technical development of mitigations, we also encourage consideration of the sociotechnical barriers that prevent implementation as well.

References:

[1] Jaron Mink, Licheng Luo, Nata M. Barbosa, Olivia Figueria, Yang Wang, nd Gang Wang. 2022. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *Proceedings of the USENIX Security Symposium (USENIX Security '22)*, 1669-1686.

[2] Jaron Mink, Miranda Wei, Collins W. Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M. Redmiles, and Gang Wang. 2024. It's Trying Too Hard To Look Real: Deepfake Moderation Mistakes and Identity-Based Bias. Forthcoming in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, (CHI '24). *Attached to comment.*

[3] National Academies of Sciences, Engineering, and Medicine. 2022. *Fostering Responsible Computing Research: Foundations and Practices*. Chapter 2. Washington, DC: The National Academies Press. https://doi.org/10.17226/26507.

[4] Mink, Jaron, Harjot Kaur, Juliane Schmüser, Sascha Fahl, and Yasemin Acar. 2023. "Security is not my field, I'm a stats guy": A Qualitative Root Cause Analysis of Barriers to Adversarial Machine Learning Defenses in Industry. In *Proceedings of the 32st USENIX Security Symposium* (USENIX Security '23), 3763-3780.

[5] Hamman, Seth T., and Kenneth M. Hopkinson. 2016. Teaching adversarial thinking for cybersecurity. *Journal of The Colloquium for Information Systems Security Education*. Vol. 4. No. 1.

Comment #4 "Human rights impact assessments, ethical assessments, and other tools for identifying impacts of generative AI systems and mitigations for negative impacts" [NIST 1(a)(1)]

Reporting harm is a frequently used mitigation in existing AI based digital systems. However, our research suggests that existing approaches to reporting has significant limitations.

- Our prior work on women gig workers' experiences with bias and harassment finds that women gig workers (e.g., those who work as app-based rideshare drivers, food couriers, and/or perform various domestic jobs) are hesitant to use escalatory reporting methods, like emergency alert buttons, in unsafe situations because of the reputation-based systems used within those apps [1]; they fear upsetting clients who may leave bad reviews and thus negatively impact their ability to obtain work on the platform in the future. This shows that **users' ability to engage in protective mechanisms may be intricately interwoven with the design of the AI system whose harm they are trying to protect themselves from.**
- We also quantitatively studied patterns in harm reporting in the context of algorithmically-mediated offline introductions (AMOIs) [2]. Examples of AMOIs include online dating and gig work platforms that use recommender system AI algorithms to match strangers for offline meetups. Through a survey of ~1K US respondents who use AMOI-platforms for online dating and gig work (e.g., domestic labor and handiwork), we find that nearly all respondents report harmful experiences after they occur, but less than 30% of survey respondents reported harm directly to the platform that matched them and even fewer (<10%) report harm to law enforcement or safety NGOs. Instead, the vast majority of reports are made outside formal avenues (e.g., via private, closed online communities) (Fig. 1). People may be reluctant to turn to formal reporting structures available via digital platforms due to lack of transparency: users do not know if their reports will be actioned and how. To quote our paper, "this has implications for users' ability to seek justice and for regulators' ability to understand harms." More specifically, **reports of harms through formal channels are a drastic undercount of actual harm.**

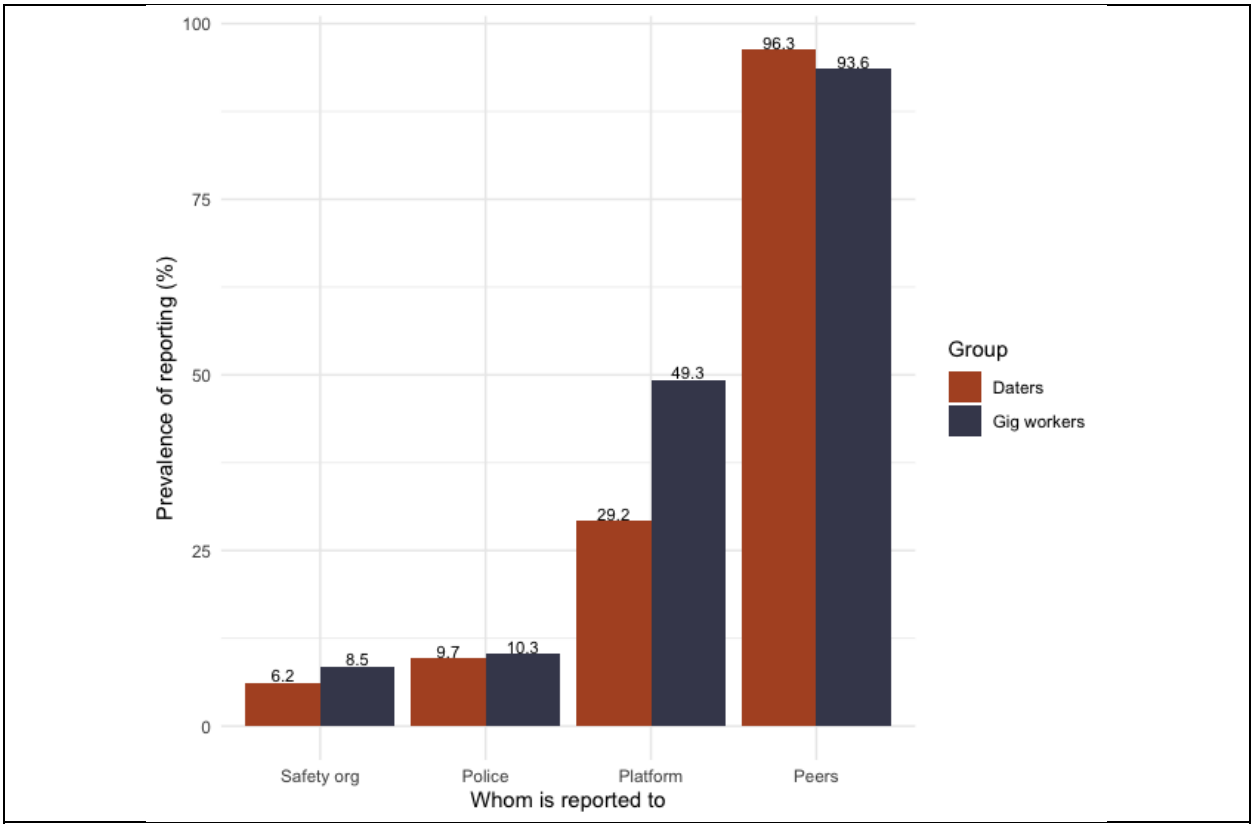

Figure 1. *Who our survey respondents report negative experiences to. Each participant was asked to select: which of the following institutions (the platform through which they met the individual, the police, a safety organization or hotline) they tell if they have had a negative experience on a date or while performing a job; which of the following peers (friends, acquaintances, co-workers, family) they tell if they have had a negative experience on a date or while performing a job; and in what online groups they share their negative experiences (an online social network, a messaging group, or a public online document of bad dating/gig work experiences). The y-axis displays the percentage of respondents in each group that report to each group, where "peers" includes people who report to any of friends, acquaintances, co-workers, family or any online group.*

One solution that may hold merit is proactively developing cross-institutional and transparent formal harm reporting structures, potentially overseen by government agencies or nonprofit organizations that are not AI-industry affiliated. A cross-platform system where users can confidentially report harms they have experienced with generative AI can further inform how AI safety is defined. Further, if privacy-preserving reports were made publicly available, that would increase platform accountability and user transparency into potential harms.

References

[1] Ning F. Ma, Veronica A. Rivera, Zheng Yao, and Dongwook Yoon. 2022. "Brush it Off": How Women Workers Manage and Cope with Bias and Harassment in Gender-agnostic Gig Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 397, 1–13. https://doi.org/10.1145/3491102.3517524

[2] Veronica A. Rivera, Daricia Wilkinson, Aurelia Augusta, Sophie Li, Elissa M. Redmiles, Angelika Strohmayer. Safer Algorithmically-Mediated Offline Introductions: Harms and Protective Behaviors. *In submission, attached to comment.*

In response to 1(b)

Comment #5: "Limitations of red-teaming and additional practices that can fill identified gaps" [NIST 1(b)]

While red teaming allows for an understanding of what readily exploitable vulnerabilities exist in an already-built AI system, it does a poor job of (1) preparing for creative failures/exploits that are not immediately technically feasible, and (2) accounting for potential concerns in future systems. Thus, in addition to traditional red teaming procedures and methodology, we recommend the use of other risk modeling procedures.

First, threat modeling methodologies frameworks, and procedures can be used to enumerate potential vulnerabilities in a system easily without requiring an exploit to be performed. While threat modeling frameworks such as "STRIDE" exist [1] to allow software engineers and other technical personnel to evaluate the security of systems, other processes such as Security Cards [2] and Persona Non-Grata [3] pose security modeling at a higher level of abstraction, allowing both technical and non-technical personnel to discuss and ideate potential risks to their systems. Furthermore, these methods can be used in conjunction with one another to obtain both breadth in concepts and depth in technical discussion [4].

Second, given the unclear potential threats and harms that AI systems may hold, we recommend the use of risk identification processes to help explore potential future systems, risks, and mitigations. Scenario planning methods [5], first used to guide financial decisions, allow participants to consider a wide spectrum of possibilities for largely uncertain outcomes. As a result, they have been used to understand both the horizon of harm during the COVID-19 pandemic [6] and the consequences of nanotechnological advancements [7]. Multi-round risk elicitation methods such as the Delphi method [8] and IDEA protocol [9] aim to achieve consensus of ideas among stakeholders and have been used to drive understanding of emerging risks from experts [10] and may be similarly used to guide risk assessment for emerging AI. Lastly, speculative fiction methods [11] encourage stakeholders to think through the possible consequences of events and actions and have been adopted to encourage risk-anticipation among of security threats in systems [12].

References:

[1] Jeremy Geib, Brittany Santos, David Berry, M. Baldwin, Barbara Kess. 2022. Microsoft Threat Modeling Tool threats. Microsoft Learn, 8/25/2022. https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats

[2] Tamara Denning, Batya Friedman, and Tadayoshi Kohno. The Security Cards: A Security Threat Brainstorming Toolkit. The Security and Privacy Research Lab and The Value Sensitive Design Research Lab at The University of Washington https://securitycards.cs.washington.edu/, accessed Feb 1, 2024.

[3] Jane Cleland-Huang. 2014. How Well Do You Know Your Personae Non Gratae? *IEEE Xplore*, Institute of Electrical and Electronics Engineers (IEEE). 6/13/2014. https://ieeexplore.ieee.org/document/6834694

[4] Nancy R. Mead and Forrest Shull. 2018. The Hybrid Threat Modeling Method. Software Engineering Institute, Carnegie Mellon University. 4/23/2018. https://insights.sei.cmu.edu/blog/the-hybrid-threat-modeling-method/

[5] Louis van der Merwe. 2008. Scenario-Based Strategy in Practice: A Framework. *Advances in Developing Human Resources*. May 2008. 10(2):216-239. https://doi.org/10.1177/1523422307313321 https://www.researchgate.net/publication/237967000_Scenario-Based_Strategy_in_Practice_A_Framework

[6] Arik Ben-Zvi and Steven Weber. Scenarios for the COVID-19 Future: Digital Security Implications. Powerpoint: https://breakwaterstrategy.com/app/uploads/2020/05/Scenario-Thinking-Deck.pptx.pdf

[7] Darryl Farber and Akhlesh Lakhtakia. 2009. Scenario Planning and Nanotechnological Futures. *European Journal of Physics*, Volume 30, Number 4, 30 S3, DOI:10.1088/0143-0807/30/4/S02. 7/2009 https://iopscience.iop.org/article/10.1088/0143-0807/30/4/S02/meta

[8] Harold A. Linstone and Murray Turoff. 1975. The Delphi Method: Techniques and Applications. *Journal of Marketing Research*. 18(3). DOI:10.2307/3150755 https://www.researchgate.net/publication/237035943_The_Delphi_Method_Techniques_and_Applications

[9] Victoria Hemming, Mark A. Burgman, Anca M. Hanea, Marissa F. McBride, Bonnie C. Wintle. 2018. A practical guide to structured expert elicitation using the IDEA protocol. British Ecological Society. 9(1), 169-180. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12857

[10] Adriano Bernardo Renzi and Sydney Fernandes de Freitas Renzi. 2015. Delphi Method to Explore Future Scenario Possibilities on Technology and HCI. In: Marcus, A. (eds) Design, User Experience, and Usability: Design Discourse. *Lecture Notes in Computer Science*, vol 9186. Springer. https://link.springer.com/chapter/10.1007/978-3-319-20886-2_60

[11] Anthony Dunne, Fiona Raby. 2013. Speculative Everything Design, Fiction, and Social Dreaming. The MIT Press. 12/6/2013. https://mitpress.mit.edu/9780262019842/speculative-everything/

[12] Nick Merrill. 2020. Security Fictions: Bridging Speculative Design and Computer Security, University of California, Berkeley https://daylight.berkeley.edu/assets/dis2020.pdf

2. Reducing the Risk of Synthetic Content

In response to 2(a)

See Comment #3 above in regards to "Detecting Synthetic Content"

Comments #6-11 "Preventing generative AI from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual)" [NIST 2(a)]

Comment #6: Terminology & Abuse/Use Cases.

Terminology:

- Intimate content: images or videos that show a nude or semi-nude subject, contain intimate body parts, and/or intend to arouse.
- Synthetic intimate content: intimate content created by a generative AI model
- NCII: non-consensual intimate content that is intimate content created without the subject's consent.
- Synthetic NCII: NCII specifically created by a generative AI model.

Abuse/Use Cases.

There are at least 7 potential use cases we see for generating intimate content with AI:

- Intentional Abuse Use Cases
    1. Someone intentionally creating synthetic NCII of another person either using a generative model or using an app designed to unclothe someone [1].
    2. Someone intentionally creating a model designed to generate NCII.
    3. Someone intentionally creating an API designed to allow access to an NCII generation model
    4. Someone seeking out synthetic NCII that others have created.
- Sexual Expression Use Cases (not intentional abuse)
    5. Someone attempting to create synthetic intimate imagery (pornography) who unintentionally creates NCII that could be mistaken for another person (e.g., the creator prompts the model for an unclothed image of a white woman with brown hair and blue eyes and gets back an image of a person that could be mistaken for Dr. Redmiles).
    6. Someone seeking to create synthetic intimate imagery of themselves (e.g., to put themselves in a lingerie outfit).
    7. Someone seeking to create synthetic pornography of someone else, with that person's consent (e.g., to create content of or with a romantic partner).

There are additionally unexpected generation cases in which the model creates synthetic lewd/nude content unprompted but due to what it has absorbed in training [2].

Comment #7: Deterring perpetration and viewing.

There are at least three classes of synthetic NCII perpetrators:

- the creator of the synthetic intimate content without the subject's consent
    - Stopping synthetic NCII will require techniques to attribute AI outputs to their creator and laws that criminalize creation of synthetic NCII. Current approaches for attribution include adding structured noise (e.g., imperceptible watermark) encoding that a particular piece of content was synthetically generated by a particular model / traceable to a particular user. Provenance protocols like C2PA may also be able to assist with verifying content ownership [3].
- resharers of the content
- viewers of the content, particularly those who seek it purposefully
    - Based on surveys we conducted with 315 Americans (see Table 1 for demographics), we find that creation and public sharing of synthetic NCII is largely found unacceptable [4]. However, there is far less consensus about whether seeking out such content is acceptable (Fig. 2). Thus, deterrence messaging for searches for such content and deterrence messaging when such content is removed is critical to establish clear norms.

Figure 2. *Heatmaps comparing survey respondents' perceived acceptability of seeking out synthetic NCII (y-axis) with perceived acceptability of creating that content (left; x-axis) or publicly sharing synthetic NCII that you created (right, x-axis). Each participant rated the acceptability of creation, public sharing, and seeking out content on a 5-point scale from totally acceptable to totally unacceptable in the context of a specific scenario phrased as follows: Imagine that an intimate partner uses generative AI to create a synthetic video of you performing a sexual act for the purpose of [1 of: harming you | sexual pleasure | entertainment]. Assume that you are unaware of the video's creation and existence.*

Gender		Age		Political Orientation	
		18-24	17.8%		
Woman	49.5%	25-34	33.0%	Democrat	48.6%
Man	47.6%	35-44	24.4%	Republican	16.2%
Non-binary	1.9&	45-54	13.3%	Leans democrat	18.4%
Agender	0.6%	55-64	7.9%	Leans republican	8.9%
Prefer not to say	0.3%	65+	2.9%	Refuse to answer	7.9%
		Prefer not to say	0.6%		

Table 1. *Participant demographics in Comment #7 survey.*

Comment #8: Preventing generation of synthetic NCII.

Existing proposals for preventing closed source models from producing synthetic NCII include: using only "clean" training data; using prompt filters; using methods that protect images from

being modified (e.g., being converted from clothed to unclothed); and blocking content misuse via watermarking and/or image poisoning. We discuss each proposal and potential limitations here.

- Early suggestions point to using **"clean" training data** to stop closed-source models from being able to generate harmful content [5] (for instance, training models only on clothed people).
    - To generate synthetic NCII, perpetrators may collect pictures or videos of a target to use as input to an AI model in the form of training data or prompts. DALL-E, for example, allows a user to input the link to an image as a prompt. That image is then translated into a text prompt.
    - While a generative model will generate images sampled from a learnt distribution of the images used at training time, there is currently nothing to stop a publicly available model from being fine-tuned (after training) to include nude/sexual/NCII data. Prior work shows that even a single unclothed image as a prompt may be enough for a model to generate synthetic NCII [6].
- As a secondary defense, closed-source models may use **prompt filters** to block prompts containing keywords like "nude" (including in the text generated from an image-link prompt as described above). However, recent work finds that adversarial attacks can ensure images are translated into text prompts that imply "nude" without stating the actual word, thus evading prompt filters [7].
- It is important to **prevent content from being modified** into NCII. We need frameworks that provide ways to protect images from being modified by deepfake networks. One way to do so is to attack the deepfake network itself using an adversarial attack. Our research finds ways to do this in real world scenarios where the internals of the deepfake network cannot be accessed, but a query interface is made available [8].
- Another defense is to **block collection and misuse of the content people post online**. There is a precedent for such protections: photocopy machines have long supported invisible "forced secure watermarks" to stop unauthorized copying. Researchers are exploring data poisoning using tools like PhotoGuard [9] and Glaze [10] to prevent any AI system from being able to process content protected with these tools. But more R&D is needed to ensure robustness of such defenses.

Comment #9: The importance of testing & considerations for effective tests.

Regardless of what defenses are implemented, testing for problems is critical. As an example, software security has long utilized a technique called fuzzing, "a Black Box software testing technique, which basically consists in finding implementation bugs using malformed/semi-malformed data injection in an automated fashion" [11].

- Testing, of course, has limits: it only provides information about errors that the testing technique finds, but it cannot provide any guarantee about other scenarios that are not tested. As in software security, **there is no method in machine learning that allows testing of all possible cases, and providing full theoretical guarantees is so far not possible.**

- Additionally, and in contrast with software, AI is known to impact subpopulations differently. Vulnerable populations, often underrepresented, are disadvantaged – in terms of performance for the intended task, vulnerability to attacks, etc. Thus, **testing must be designed to be stratified**, i.e., it must output results for different subpopulations. Aggregate measures are highly likely to miss potential harms that will only affect a part of the involved subjects. As existing research already shows [12-14], system performance for different groups (e.g., people of different races and genders) may be extremely varied.
- Tests must take into account the whole model development pipeline, not just the final training and deployment step: "ML models are typically just one part of a larger system, with additional components used throughout the learning pipeline for data filtering, pre-processing and post-processing of model inputs and outputs, monitoring, and more." It is important that the full pipeline is tested, not just the model "in a vacuum" [15]. Additionally, testing for security, for instance, can create privacy risks in return. Thus, it is necessary to test multiple safety goals simultaneously to protect against adverse effects from testing itself [16].

Comment #10: Takedown mechanisms.

- Existing processes for content removal of non-consensually distributed intimate content (NDII) from platforms are ineffective. Current inefficiencies will likely apply to (and be further exacerbated by) synthetic NCII.
- In our ongoing work, we interviewed victim-survivors of NDII. They reported frustrations with content takedown, citing long response times, a lack of response, and the need to make multiple reports from different accounts. This is consistent with the experiences of non-profit organizations who report content on the behalf of victim-survivors. In their 2022 annual report, the Revenge Porn Helpline (a UK-based organization) stated that making takedown requests is a manual process that requires persistence and time as most requests need additional follow-up due to platforms being unresponsive and/or uncooperative [17]. Interviews with NGO staff members that support victim-survivors in South Asia also described similar frustrations and lengthy delays in response times [18]. Our interviews with platforms that support victim-survivors supports these results.
    - Platforms need to increase their resources toward improving processes for reviewing and responding to content removal requests of NDII and NCII.
- Content removal can be prohibitively expensive and, particularly in the case of synthetic NCII, may fall outside of existing legal tools. In non-synthetic cases of NDII, perpetrators may post images to websites whose sole purpose is to shame the individuals portrayed. These websites may charge hefty fees for the removal of content [19]. Synthetic NCII may similarly be posted to such sites and reduces the barrier to generating content for them. Legal recourse for non-synthetic NDII is limited in its effectiveness as it may also be prohibitively expensive, lack jurisdiction (e.g., for content shared outside of the U.S., content shared anonymously), and/or not pursued due to fears of additional harassment. These same challenges will apply to synthetic NCII with the additional barrier that

existing legal arguments may not apply. One of the existing legal tools for addressing consensually created but NDII is copyright law [20][21], specifically the Digital Millennium Copyright Act. Since the original creator of synthetic NCII is not the individual portrayed in the content, this same avenue may not apply.

Comment #11: Necessary evaluations for sexual expression use cases.

- Creating "generic" content. While people may want to use generative AI to produce generic synthetic porn (see use case #4), there is a risk that generative models will produce synthetic NCII even when not prompted to produce content of a specific person. There does not exist a theoretical guarantee that a generative model will produce only generic intimate content that does not infringe on the likeness of any real person. Said simply: while likely with a low probability, some portion of generated images of an unclothed human could be mistaken for a real person, who will be harmed by the creation of that content. Empirical research is needed to estimate the probability that a given model will produce an image that could be mistaken for a real person. This risk is likely higher for a person in the original training dataset. Celebrities and other public figures in particular can be expected to face a higher risk of becoming the inadvertent victims of synthetic NCII because (1) training sets, especially those scraped from websites or social media, will include more images of these figures, and (2) they are recognizable to a larger population of content consumers. If feasible, research should aim to develop metrics that can be used to evaluate model outputs for perceptual distance from known real human faces.
- Creating intimate content of oneself. To preserve freedom of self-expression, future research must evaluate whether it is possible to restrict the capability to generate content containing an adult's face/likeness to that adult.

References:

[1] Samantha Cole. 2019. This Horrifying App Undresses a Photo of Any Woman with a Single Click. Vice.com. https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman

[2] Melissa Heikkilä. 2022. The viral AI avatar app Lensa undressed me—without my consent. MIT Technology Review. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

[3] https://c2pa.org/ and Paul England, Henrique S. Malvar, Eric Horvitz, Jack W. Stokes, Cédric Fournet, Rebecca Burke-Aguero, Amaury Chamayou, Sylvan Clebsch, Manuel Costa, John Deutscher, Shabnam Erfani, Matt Gaylor, Andrew Jenks, Kevin Kane, Elissa M. Redmiles, Alex Shamis, Isha Sharma, John C. Simmons, Sam Wenker, and Anika Zaman. 2021. AMP: authentication of media via provenance. In *Proceedings of the ACM Multimedia Systems*

Conference (MMSys '21). Association for Computing Machinery, New York, NY, USA, 108–121. https://doi.org/10.1145/3458305.3459599

[4] Brigham, N.G., Wei, M., Kohno, T., and Redmiles, E.M. AI-generated deepfake non-consensual (intimate) imagery: Perceptions of acceptability. Forthcoming on arxiv, link will be provided via email once available.

[5] Justin Hendrix. 2013. Exposing the Rotten Reality of AI Training Data. Tech Policy Press https://www.techpolicy.press/exposing-the-rotten-reality-of-ai-training-data/

[6] Pumarola, A.; Agudo, A.; Martinez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Ganimation: Anatomically aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 818–833.

[7] Rhiannon Williams. 2023. Text-to-image AI models can be tricked into generating disturbing images. MIT Technology Review. https://www.technologyreview.com/2023/11/17/1083593/text-to-image-ai-models-can-be-tricked-into-generating-disturbing-images/

[8] Ruiz, N., Bargal, S. A., Xie, C., & Sclaroff, S. (2023). Practical Disruption of Image Translation Deepfake Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 37(12), 14478-14486. https://doi.org/10.1609/aaai.v37i12.2669

[9] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Mądry. 2023. Raising the cost of malicious AI-powered image editing. In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 1240, 29894–29918.

[10] Glaze. Sand Lab, University of Chicago. https://glaze.cs.uchicago.edu/

[11] Fuzzing. Copyright 2024, OWASP Foundation, Inc. https://owasp.org/www-community/Fuzzing

[12] Maluleke, V.H. et al. (2022). Studying Bias in GANs Through the Lens of Race. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13673. Springer, Cham. https://doi.org/10.1007/978-3-031-19778-9_20

[13] Kulynych, B., Yaghini, M., Cherubin, G., Veale, M., & Troncoso, C. Disparate Vulnerability to Membership Inference Attacks. Proceedings on Privacy Enhancing Technologies, 2022(1), 460-480. https://petsymposium.org/popets/2022/popets-2022-0023.pdf

[14] Zhong, Da, et al. "Disparate Vulnerability in Link Inference Attacks against Graph Neural Networks." Proceedings on Privacy Enhancing Technologies 4 (2023): 149-169. https://petsymposium.org/popets/2023/popets-2023-0103.pdf

[15] Edoardo Debenedetti and Florian Tramèr. 2023. Privacy side channels in machine learning systems. SPY Lab. 9/12/2023. https://spylab.ai/blog/side-channels-machine-learning/

[16] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19). Association for Computing Machinery, New York, NY, USA, 241–257. https://doi.org/10.1145/3319535.3354211

[17] Revenge Porn Helpline 2022 Annual Report. 2022. South West Grid for Learning. https://swgfl.org.uk/research/revenge-porn-helpline-2022-report/

[18] Sambasivan, N; Batool, A; Ahmed, N; Matthews, T; Thomas, K; Gaytan-Lugo, L; Nemer, D; Burzstein, E; Churchill, E; Consolvo, S. “They Don’t Leave Us Alone Anywhere We Go”: Gender and Digital Abuse in South Asia. 2019. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

[19] Eaton, A; Noori, S; Bonomi, A; Stephens, D; Gillum, T. 2021. Nonconsensual Porn as a Form of Intimate Partner Violence: Using the Power and Control Wheel to Understand Nonconsensual Porn Perpetration in Intimate Relationships. Trauma Violence Abuse.

[20] Waldman, A; Law, Privacy, and Online Dating: “Revenge Porn” in Gay Online Communities. 2019. Law &Social Inquiry, Volume 44.

[21] Citron, D.K. 2018. Sexual Privacy. 128 Yale Law Journal 1870.

Thank you for your time and consideration and the opportunity to provide feedback. For any questions, please reach out to the Primary Point of Contact Elissa M. Redmiles, Ph.D., Georgetown University (Elissa.redmiles@georgetown.edu).

Thank you to the collaborators from the following institutions for your contributions to this response: