GETTING-Plurality
Research Network

Date: July 7, 2023
Docket ID: OSTP-TECH-2023-0007
From: GETTING-Plurality Research Network, Edmond & Lily Safra Center for Ethics, Harvard University
Submitted via regulations.gov

GETTING-Plurality Research Network Comment on the White House OSTP National Priorities for Artificial Intelligence: Protecting Rights, Safety, and National Security

GETTING-Plurality is a multi-disciplinary research network at the Edmond & Lily Safra Center for Ethics at Harvard University led by Professor Danielle Allen and Professor Allison Stanger. The network pursues foundational analysis and theory, field-building, and policy development to foresee and mitigate potential harms to democracy and to strengthen the public benefit and democracy-supportive effects flowing from technology innovation. More information available at: https://gettingplurality.org/

To protect rights, safety, and national security, in light of the emergence of generative foundation models, will require focus on five things: (1) an analytical framework defining key concepts and hazard tiers for different categories of AI; (2) organizational frameworks including both voluntary industry bodies to set standards for the lower hazard tiers and governmental capacity to regulate and monitor work in the highest hazard tier category or categories of AI; (3) development of human capacity to staff and lead both the governmental offices involved and the voluntary industry bodies; (4) achieving shared definitions and regulatory frameworks with both allied and competitive nations; and (5) structures for implementing all of the above. This memo addresses these five themes in the course of answering the Protecting Rights, Safety, and National Security RFI questions.

What specific measures – such as standards, regulations, investments, and improved trust and safety practices – are needed to ensure that AI systems are designed, developed, and deployed in a manner that protects people's rights and safety? Which specific entities should develop and implement these measures?

1A. Key Concepts that Define Governance Landscape

Determining which standards, regulations, investments, and improved trust and safety practices are necessary will start from establishing a shared vocabulary for key concepts in the field and hazard tiers for different categories of AI. That will then permit recommendations about organizational structures for governance. Additional definitions may be found in the appendix.

Generation 1.0 AI	AI models and tools that pre-date the arrival of GPT and other foundation models. For these models, algorithmic bias is a leading and also well-understood problem.
Generation 2.0 AI	Already existing massively scaled foundational models (both big corporate models and open source tools).
Frontier AI	“We loosely define the ‘frontier’ as models that are both (a) close to, or exceeding, the average capabilities of the most capable existing models, and (b) different from other models, either in terms of scale, design (e.g. different architectures or alignment techniques), or their resulting mix of capabilities and behaviors. Accordingly, frontier models are uniquely risky because (a) more capable models can excel at a wider range of tasks, which will unlock more opportunities to cause harm; and (b) novel models are less well-understood by the research community.” Source: Extreme Risks 2023
Capabilities Risk	These are risks that flow from how the AI models themselves function. The contrast is to Systemic Risk, which flows from how AI models interact with other social systems.
Individual Risk	These are risks that put individuals at risk of serious degradation to their mental, physical, or economic health, including risk of death, but with low frequency of occurrence, so that they do not create systemic, structural, or extreme risks. These risks are nonetheless important as they are in any consumer protection or biomedical context.
Existential Risk	Existential risk is the “‘permanent destruction of human potential.’ Actual human extinction is existential, but so would be the irreversible collapse of civilization.” (Toby Ord, Precipice.)
Extreme Risk	Extreme risks flow from the capabilities of the models themselves or alignment failures with impacts “that would be extremely large in scale (even relative to the scale of deployment). This can be operationalized in terms of the scale of impact (e.g. damage in the tens of thousands of lives lost, hundreds of billions of dollars of economic or environmental damage) or the level of adverse disruption to the social and political order. The latter could mean, for example, the outbreak of inter-state war, a significant erosion in the quality of public discourse, or the widespread disempowerment of publics, governments, and other human-led organisations (Carlsmith, 2022).” Source: Extreme Risk 2023

Systemic Risk	These are risks that depend on how the AI system interacts with other social, political, and economic systems—including legal structures, social practices, and cultural norms—and that entail either a high degree of disruption to ordinary functioning (if on a slower timetable or with somewhat lower amplitude than extreme risks) or a significant exacerbation of existing patterns of injustice or barriers to human flourishing. This can include transformative impacts on the functioning of political institutions or economic systems that take a decade or more to unfold, or exacerbation of domination-based disparities. Systemic risks can involve either low or high impacts on individuals but will involve many individuals. In cases where systemic risks also entail high individual risk, a challenge will be to determine whether the appropriate response is redress of many individual harms or a structural response that would prevent further harms but not necessarily redress harms already suffered. In general, it is important both to develop structural responses to prevent future harm and methods of redressing the individual harms that lie beneath the systemic phenomenon.
Genocidal Risk	Whereas genocide is defined by the UN Article II Convention as a crime committed with the intent to destroy a national, ethnic, racial, or religious group in whole or in part, genocidal risk is (a) the foreseeable capacity of a model to destroy a national, ethnic, racial, or religious group in whole or in part, or (b) the observable emergence of an unforeseen capacity of a model to destroy a national, ethnic, racial, or religious group in whole or in part. Genocidal risk should be added to the categories of risk for which tool developers are responsible. Challenges with algorithmic bias and entrenched discrimination are relevant here.

1B. AI Risk Profiles and Hazard Tiers

Using these definitions, we can identify categories of risk that are applicable to different categories of AI.

Type of Risk	Existential	Extreme	Systemic	Genocidal	Individual
Generation 1.0			X	X	X
Generation 2.0	X	X	X	X	X
Frontier AI	X	X	X	X	X

This risk framework leads us to propose the following Hazard Tiers:

Hazard Tier 1	Generation 1.0	Low capabilities risk	Deployed in low-risk contexts
Hazard Tier 2	Generation 2.0	Low capabilities risk	Deployed in low-risk contexts
Hazard Tier 3	Generation 2.0	Moderate to extreme capabilities risk	Deployed in low-risk contexts
Hazard Tier 4	Generation 1.0	Low capabilities risk	Deployed in contexts of systemic, high individual, or existential risk
Hazard Tier 5	Generation 2.0	Moderate to extreme capabilities risk	Deployed in contexts of systemic, high individual, or existential risk
Hazard Tier 6	Frontier AI	Extreme to existential capabilities risk	Deployed in low-risk contexts
Hazard Tier 7	Frontier AI	Extreme to existential capabilities risk	Deployed in contexts of systemic, high individual, or existential risk

Using these hazard tiers will require, of course, a more fully developed set of definitions and lexicons for legibility. We expect to publish a Roadmap for AI Governance later this summer that will provide more fully elaborated definitions and lexicons.

2. Recommended Organizational Structures

Different organizational structures should be deployed for different hazard tiers.

Hazard Tiers 1-3: Encourage the development of a voluntary industry body to certify quality for the lowest-risk apps, modeled on the Vitamin Board (U.S. Pharmacopeia). This Board, established and sustained by industry entities, could set standards and issue certifications to guide consumers toward high quality products, and to permit consumers, whether individual or organizational, to distinguish between high and low quality applications. This will require clarifying that antitrust rules permit (narrow) collaboration on safety and security.

Hazard Tiers 4-5: All government agencies need offices to monitor and evaluate the integration of AI tools by entities operating within their jurisdiction, and need to prepare to use existing enforcement mechanisms from consumer protection to anti-discrimination mechanisms, and including litigation, to ensure integration of AI in modes consistent with existing protections for rights and safety. These offices must be staffed with teams capable of auditing the self-evaluations of labs and app-builders for lower-risk models and diverse contexts of deployment.

Hazard Tiers 6-7:

- Frontier AI research and development should be conducted only in federally registered Labs that are regulated by the Department of Energy. We recommend the Department of Energy for this role given the Department's existing experience and expertise with regulating and monitoring national Labs, as well as the existing integration of the Department of Energy with the National Security infrastructure. The Department of Energy should require that these federally-registered Frontier AI Labs meet well-thought out standards for continuous evaluation of training, pre-deployment decisions, deployment, and security.
- Deployment decisions for new technologies emerging from frontier AI labs should be regulated by a new Office within the Department of Homeland Security, established to function in a fashion similar to the FDA. The Department of Homeland Security is the appropriate agency because the evaluations needed for new AI technologies will require whole systems analysis, and cross-jurisdictional analysis. While the Department of Homeland Security may not yet have the necessary capacity, it is best situated among existing federal agencies to acquire the necessary capacity. Assignment here would, however, require a standard for agency leadership where the agency Secretary combines capacity to lead in conventional security work with capacity to lead build-up of broader systems-oriented analytic capacity. An interesting example of the type of person who could lead in this way is James Manyika.
- Either the Department of Energy or the Department of Homeland Security would need to be in a position to conduct independent evaluations of Frontier AI technologies. This will require public sector provision of compute sufficient to conduct that work. This will require a significant investment of public funds.

3. Human Talent Pipeline

Operationalizing any of the above organizational structures will require significantly increasing the talent pipeline of engineers, scientists, and AI ethicists into public sector service. We recommend the following actions:

- Investment in public workforce with technical capabilities to monitor compute/evaluate and audit models via use of Federal scholarship funds, perhaps drawing on the model of the Fulbright to require service in exchange for scholarship support;
- Engagement of all colleges and universities with high levels of graduates in technical fields to build an ethos of an expectation of national service at some point over the career life course; collaborate with the American Association of University Presidents and other such national consortia to achieve that;
- Engage all National Service agencies and civil society bodies in incorporating AI-related service

in their mandates;

- Investment in ensuring that all of these offices and agencies will include on their teams people trained to do work on ethics. We can ask those AI ethics teams to think about whether models and model-based apps support or harm human flourishing, support or harm democracy and political stability, support or harm innovation, creation, and the integration of all people in the productivity of the economy.
- Fund research into (1) hardware security features on high-end computing hardware, e.g. verify compliance with regulations, allow for remote shutdown of noncompliant training runs (also relates to next question); (2) research in explainability and public education/disclaimers and (3) provenance markings on AI-generated content.

4. National Security Framework

What are the national security benefits associated with AI? What can be done to maximize those benefits?

- Generative AI is a powerful dual use technology that conveys a significant advantage to the United States. The supply chain is intricate, with choke points that are largely controlled by the US and its allies.
- Continue to incentivize TSMC's opening of a second North America office in Phoenix
- Continue to assure ASML and the Dutch that lending their technological expertise to the Chinese is a bad idea
- Can AI enable lower-trust verification of treaties?

We recommend a national security framework in which we secure shared key concepts and hazard tiers with allies with large labs (UK and Canada), seek to achieve a cooperative regime with China, and develop capacity in the State Department for monitoring and evaluating implications of AI for nations without significant AI capacity.

- Engage the Department of State to secure shared conceptual frameworks with allies.
- Work to build a global accord around a commitment to (1) no first use of generative foundation models against digital infrastructure or civilians; and (2) no unregulated release to open source.
- More specifically, develop international agencies or organizations for global standards-setting on weaponized AI. An IAEA for AI that would allow for the sovereignty of states over the use of AI within their borders if member states pledge no first use of AI weaponry that targets the civilian population and its related critical infrastructure in other sovereign countries. Sanction violators.
- Strengthen cooperation among the free world countries through the NATO Alliance, the Five Eyes countries alliance and a US-European CERN for AI.
- Track the stocks and flows of high-end computing hardware (analogy to nuclear material); potentially expand October's export controls
- Conduct rigorous oversight of the U.S. Intelligence Community's incorporation of artificial intelligence. Benchmark the IC's activities against human rights principles as well as their

own Artificial Intelligence Ethics Framework.

5. Implementation Vehicle: National AI Task Force

To implement all of the above, we recommend elevating the National AI Advisory Committee (NAIAC) to the level of a task force similar to the National Climate Task Force and with the National Artificial Intelligence Initiative Office providing administrative support, the National AI Research Resource Task Force coordinating research support, and the Select Committee on AI as a liaison on matters pertaining to the National AI Initiative, that would comprise government officials from relevant agencies and research institutes (Departments of Justice, Commerce, Labor, Federal Trade Commission, Homeland Security (TSA, FEMA, cybersecurity & infrastructure), NIST, National Science Foundation, National Academies/American Academy of Arts & Sciences, State for international collaboration) together with AI experts, civil society and private sector representatives to inform Executive Branch decision-making and work with relevant agencies and partners on specific measures and regulations. NAIRRTF's mandate includes a narrow focus, which can be expanded, on assessments of privacy, civil rights and civil liberties requirements of AI applications. The SCAI consists of the most senior R&D officials across federal agencies and represents a whole-of-government approach to AI R&D planning and coordination.

The NAIAC as a task force could reorient its focus from AI R&D toward work with the Department of Energy and Department of Homeland Security to design and implement domestic standards, regulations, investments,and improved trust and safety practices for AI to be deployed in a manner that protects people's rights and safety while enabling continued US leadership in AI and prepare present and future US workforce for integration of AI systems across all sectors of the economy and society. This would necessitate greater involvement of the National Academies and social science research on specific measures and regulations.

Thank you for this opportunity to comment on this important topic. The GETTING-Plurality Research Network welcomes any further discussion and can be reached at contact@gettingplurality.org

EDMOND & LILY SAFRA
Center for Ethics

JUSTICE, HEALTH & DEMOCRACY
IMPACT INITIATIVE

Appendix

Additional terminology:

Individual Opportunities	These are opportunities that individuals could directly experience as a result of the integration of new AI tools into social processes, as characterized from the individuals' perspective.
Systemic Opportunities	These are opportunities to improve the functioning of major social systems—whether economic, political, educational, recreational, or otherwise pertaining to civil society.
Labor Share	The share of national income received in wages.
Economic Integration	The incorporation of individuals and population sub-groups in the productive structure of the economy
Alignment	This is the process of working to ensure the AI model functions in alignment with the safety, ethics, and impact parameters that designers are seeking to impose.