OUR **SECURE** FUTURE

Women Make the Difference

WPS MESSAGE GUIDE

A Gender Perspective on AI Risks to National Security

"Artificial intelligence (AI) holds extraordinary potential for both promise and peril … irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. ***Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society."***

EXECUTIVE ORDER 14110 ON THE SAFE, SECURE, AND TRUSTWORTHY DEVELOPMENT AND USE OF ARTIFICIAL INTELLIGENCE

CHECKLIST FOR DEVELOPING GENDER-SENSITIVE AI POLICY

- The interests and rights of women, men, boys, girls, and other genders who increasingly use, interact with, or purchase AI and AI-enabled products in their daily lives are protected.
- People's privacy and civil liberties are protected by ensuring that the collection, use and retention of data is lawful, secure and promotes privacy.
- AI policies are consistent with the advancement of equity and civil rights.
- Strategies, policies and programs take into account the risks, including gender-sensitive risks, from the government's own use of AI and increase its internal capacity to regulate, govern and support responsible use of AI to deliver better results for Americans.

UNDERSTANDING GENDERED RISKS IN AI IS CRITICAL TO NATIONAL SECURITY[1]

- A rules-based order is the foundation of national security.
- Applying a gender lens to AI ensures alignment with national security objectives.
- Nonalignment with human rights standards undermines security priorities and national interests in promoting democratic and free societies.
- AI security risks are gender-blind, ignoring the disparate roles, needs and impacts for women and girls.
- A gender perspective provides early warning signals of AI risks to national security.

Five Major AI Risks and their Gender Impacts

MILITARY AI RACE

Competition by government and corporate entities to rapidly develop new AI innovations raises risks of premature deployment, accidents, cyberattacks or intensified conflict.[2]

One major arena of AI competition is the development of Lethal Autonomous Weapon Systems (LAWS), which independently identify and engage targets. Algorithmic bias, malfunction or improper deployment in LAWS can lead to fatal errors. International protocols like the Convention on Certain Conventional Weapons (CCW) have been passed to keep humans in the loop of LAWS decision-making, but changing national definitions of autonomous weapons and levels of human involvement can potentially decrease human control over lethal weaponry.[3]

GENDER BIAS INCREASES POLITICAL INSTABILITY

Autonomous weapons trained on incomplete data that is not representative of their targets can lead to fatal errors.[4] AI can also ingrain harmful norms based on biased data-sets, especially when predominately men develop this technology.[5]

DISINFORMATION

AI or Human Driven Disinformation

AI-generated content can become increasingly pervasive in dis- and mis- information campaigns as it grows more realistic and persuasive, manipulating public understanding of reality and contributing to increased polarization.

The spread of AI will enable personalized targeting that could skew shared societal perceptions of reality.[6] AI-generated images and videos, already a feature of the 2024 presidential race, are helping to spread disinformation.[7] The creation and dissemination of disinformation can be orchestrated by either human malicious actors or rogue AI.

AI Driven Deception

As AI systems generate increasingly realistic content and seek to obtain human-designated objectives, the capacity of rogue AI to deceive humans grows.

Deception thrives in areas like politics and business. ChatGPT lied to a TaskRabbit worker, telling them that the AI was a real, visually-impaired person who needs assistance unlocking a Captcha.[8] In 2023, a lawyer asked ChatGPT to find relevant legal briefs. The model, using deceptive language and word prediction, created legal briefs that appeared real. A later judge deemed them "legal gibberish."[9]

GENDER BIAS INCREASES PHYSICAL INSECURITY & CHILLS DEMOCRATIC PARTICIPATION

Women are disproportionately targeted by gendered and sexualized disinformation campaigns, seeking to frame female leaders as inherently incompetent or overly emotional to hold office positions.[10] Deepfakes, an example of AI deception, also predominantly target women.[11] The advent of more realistic synthetic content increases the likelihood that bad actors will create sexualized, demeaning imagery of more women. Both forms of manipulation lead to increased physical insecurity and chill women's participation in the democratic political process.

AI POWER-SEEKING BEHAVIOR

AI systems may be incentivized to acquire more power and resources to fulfill power- seeking objectives, making them difficult to control.

A recent study shows that reinforcement learning teaches machines to develop behavior that avoids shutdown, displaying power-seeking behavior.[12] AI could learn to seek power through legitimate, deceitful or forceful tactics, by hacking computer systems, acquiring financial or computational resources, influencing politics, or controlling factories and physical infrastructure.[13]

GENDER BLINDNESS EXPLOITS INEQUALITIES BASED ON BIASED AND LIMITED DATA

Power seeking behavior can come from socially constructed gender norms and conceptions of masculinity. If AI systems are based on gender-blind data sets, they are more likely to be instilled with harmful norms that exploit inequality.[14] Power-seeking behavior could also make AI resistant to efforts to re-train or deactivate them.

POWER IMBALANCE

Competent AI systems can concentrate power among a small group of people or states, while excluding groups with less resources, thus creating a lock-in of oppressive systems.

Human-made rules about AI, as well as AI's autonomous decisions regarding resources, have the potential to disrupt the balance of power. Groups and countries with better resources for AI development can exacerbate inequalities between groups, as well as increase the wealth gap between rich and poor countries in AI development.[15] The AI global workforce lacks gender, racial and ethnic diversity; powering these technologies also requires sophisticated infrastructure, concentrating AI investment and development in advanced economies. Power imbalances also appear in limited datasets that are biased.

GENDER BIAS INCREASES STRUCTURAL INEQUALITY

AI gender bias is more pronounced than real life. If AI models are trained from biased and synthetic data and developed by mainly one demographic, they can lock automation systems in past norms unrepresentative of contemporary realities.[16] Countries with more advanced AI infrastructure also possess a greater quantity of data centers, further entrenching structural inequalities. Biases in AI decision-making can also tilt the balance of power.

Women, Peace and Security is Critical to AI National Security Interests

WPS should be a critical element of engaging with foreign allies and partners when developing a policy framework to manage AI risks.

> *"The Federal Government will seek to promote responsible AI safety and security principles and actions with other nations, including our competitors,* ***while leading key global conversations and collaborations to ensure that AI benefits the whole world, rather than exacerbating inequities, threatening human rights, and causing other harms."***
>
> EXECUTIVE ORDER 14110 ON THE SAFE, SECURE, AND TRUSTWORTHY DEVELOPMENT AND USE OF ARTIFICIAL INTELLIGENCE

KEY MESSAGE

Ensure AI Alignment with Existing Human Rights Frameworks

- Global and national commitments to human and women's rights must be integrated into technology and security policy. AI should be aligned with established human rights frameworks.
- Gender analysis on technology policy and development increases situational awareness and democratic civilian engagement, and is required by the WPS Act and the Women's Entrepreneurship and Economic Empowerment Act.
- Gender inequality in technology is an early warning sign of broader political and social instability, and inherent issues of gender bias and inequalities should be a main focus in the development and adoption of new technologies within the national security and defense spaces.

WOMEN, PEACE AND SECURITY RECOMMENDATIONS

- Ensure AI alignment in policy and programs with existing international human rights frameworks, policies, and norms.
- Use the Women, Peace and Security policy framework to reinforce established human rights standards and support national security objectives.
- As per the WPS Act mandate, consult regularly with civil society with specific attention to including women leaders and women-led civil society organizations.

OUR SECURE FUTURE
Women Make the Difference

Gender perspectives improve security outcomes

Washington, DC 20036, USA
oursecurefuture.org | @OurSecureFuture

Endnotes
1. Based on an upcoming report, "Women, Peace and Security, Technology, and National Security: What World Are We Building?," by Sahana Dharmapuri and Jolynn Shoemaker
2. Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," Center for AI Safety, October 9, 2023, https://arxiv.org/pdf/2306.12001.pdf
3. United Nations Institute for Disarmament Research, "Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies," 2018, https://unidir.org/files/publication/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf
4. Zachary Arnold and Helen Toner, "AI Accidents: An Emerging Threat: What Could Happen and What to Do," Center for Security and Emerging Technology," July 2021, https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/
5. Ray Acheson, "Gender and Bias: What does gender have to do with killer robots?," Stop Killer Robots, 2021, https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Gender-and-Bias.pdf
6. Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," Center for AI Safety, October 9, 2023, https://arxiv.org/pdf/2306.12001.pdf
7. Nina Jankowicz, "The threat from deepfakes isn't hypothetical. Women feel it every day," the Washington Post, March 25, 2021, https://www.washingtonpost.com/opinions/2021/03/25/threat-deepfakes-isnt-hypothetical-women-feel-it-every-day/
8. Beatrice Nolan, "The latest version of ChatGPT told a TaskRabbit worker it was visually impaired to get help solving a CAPTCHA, OpenAI test shows," Business Insider, March 16, 2023, https://www.businessinsider.com/gpt4-openai-chatgpt-taskrabbit-tricked-solve-captcha-test-2023-3?IR=T
9. Benjamin Weiser and Nate Schweber, "The ChatGPT Lawyer Explains Himself," the New York Times, June 8, 2023, https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html
10. Lucina Di Meco and Kristina Wilfore, "Gendered disinformation is a national security problem," the Brookings Institution, March 8, 2021, https://www.brookings.edu/articles/gendered-disinformation-is-a-national-security-problem/
11. Nina Jankowicz, "The threat from deepfakes isn't hypothetical. Women feel it every day," the Washington Post, March 25, 2021, https://www.washingtonpost.com/opinions/2021/03/25/threat-deepfakes-isnt-hypothetical-women-feel-it-every-day/
12. Victoria Krakovna and Janos Kramar, "Power-seeking can be probable and predictive for trained agents," DeepMind, 2023, https://arxiv.org/abs/2304.06528
13. Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," Center for AI Safety, October 9, 2023, https://www.safe.ai/ai-risk#Deception
14. Ray Acheson, "Gender and Bias: What does gender have to do with killer robots?," Stop Killer Robots, 2021, https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Gender-and-Bias.pdf
15. Cristian Alonso, Siddharth Kothari, Sidra Rehman, "How Artificial Intelligence Could Widen the Gap Between Rich and Poor Nations," IMF Blog, December 2, 2020, https://www.imf.org/en/Blogs/Articles/2020/12/02/blog-how-artificial-intelligence-could-widen-the-gap-between-rich-and-poor-nations
16. Leonardo Nicoletti and Dina Bass, "Humans are Biased. Generative AI is Even Worse," Bloomberg, 2023, https://www.bloomberg.com/graphics/2023-generative-ai-bias/