DATA&
SOCIETY

Written Statement of Janet Haven
Executive Director, Data & Society

US Senate AI Insight Forum:
Risk, Alignment, and Guarding Against Doomsday Scenarios

December 6, 2023

Leader Schumer, Senators Rounds, Heinrich, and Young, and other members of the US Senate, thank you for the opportunity to speak with you all today.

I am the executive director of Data & Society, an independent, nonprofit research institute studying the social implications of AI, automation, and other data-centric technologies. Through empirical research and policy and media engagement, our work illuminates the values and decisions that drive these systems and helps shape futures grounded in equity and human dignity. I am also a member of the US National AI Advisory Committee (NAIAC), the federal advisory committee that submits expert recommendations on AI to the president.

Addressing the theoretical risks of AI in the future begins with addressing the ways AI is harming Americans now.

Evidence already demonstrates AI's negative impacts on workers' jobs and economic opportunity,[1] excessive use of scarce resources such as water and energy,[2] racially biased outcomes in medical treatment,[3] arbitrary decisions in social and medical benefits,[4] and civil rights abuses in policing and the justice system,[5] among other areas.

Legislation to address these harms, grounded in existing regulatory powers of the federal government, is not only the right thing to do for the many Americans suffering from these AI impacts; such legislation would also build the enduring structures of AI evaluation, transparency, and refusal (proactively choosing not to use or discontinuing use of AI) that allow us to better identify and safeguard against novel emerging risks.

---

[1] https://datasociety.net/library/explainer-algorithmic-management-in-the-workplace/; https://datasociety.net/library/challenging-worker-datafication/.
[2] https://www.ted.com/talks/sasha_luccioni_ai_is_dangerous_but_not_for_the_reasons_you_think?language=en; https://arxiv.org/abs/2304.03271.
[3] https://www.nature.com/articles/s41746-023-00939-z; https://doi.org/10.1126/science.aax2342.
[4] https://www.statnews.com/2023/03/13/medicare-advantage-plans-denial-artificial-intelligence/; https://www.google.com/books/edition/Automating_Inequality/pn4pDwAAQBAJ?hl=en&gbpv=0.
[5] https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html; https://www.brookings.edu/articles/algorithms-and-sentencing-what-does-due-process-require/.

1. **AI policy should be evidence-based and grounded in the urgent, real world harms affecting people today.**

It may be enticing to imagine AI's popular emergence as a signal from the future, a marker that ideas from science fiction are nearly upon us. But technological forecasting is uncertain at best, and becomes more uncertain the further out we try to predict. AI policy should be grounded in the here and now, drawing on the policy tools we have at our disposal, not in a theoretical future that exists outside of human controls.

The failure point in focusing exclusively, or even primarily, on hypothetical AI threats is that they are, by definition, unprovable and unfalsifiable.[6] Potential doomsday scenarios are not *sui generis*, warranting an altogether different, alternative approach to regulation. They are one of many possible risks along a spectrum of risk deriving from advanced technologies.

While the possibility of a doomsday scenario is important to take into account, it must also be contextualized. Currently, there is no empirical evidence that warrants policymakers focusing primarily on hypothetical scenarios like a sentient superintelligence (aka artificial general intelligence or "AGI") that is uncontrollable by humans or a hyper-powerful machine that finds and attacks vulnerabilities in critical infrastructure.[7] These kinds of scenarios are the realm of thought experiments and future-casting. They are not rooted in the reality that everyday Americans occupy, nor are they tractable safety engineering problems.[8] They should not distract Congress from putting in place legislative controls to address and alleviate the harms from AI that Americans are suffering today.

When politicians and policymakers of the prior century took steps to avert nuclear proliferation and avoid a nuclear apocalypse, their concerns made sense. Clear evidence existed about what nuclear weapons do to human beings, to human societies, and to the planet. There is no equivalent real world evidence that AI's existential threat warrants lawmakers' undivided attention.

A possible doomsday situation is, to be sure, *a risk.* But policymaking should focus on the urgent issues impacting Americans today, not the hypothetical risks of an unknown future.

Critically, responding to real, material harms is what the American people want. The majority of Americans are concerned about issues like how employers use AI to hire and manage workers, how healthcare providers' use of AI may degrade patient outcomes, and how AI's widespread deployment will

[6] https://techpolicy.press/artificial-intelligence-and-the-ever-receding-horizon-of-the-future/.
[7] https://paperswithcode.com/paper/a-review-of-the-evidence-for-existential-risk; https://onlinelibrary.wiley.com/doi/full/10.1111/rati.12320; https://www.technologyreview.com/2020/02/25/906083/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/.
[8] https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf; https://dl.acm.org/doi/pdf/10.1145/3544548.3581407.

affect their privacy and freedoms.[9] Federal lawmakers should respond to the urgent concerns of their constituents rather than to unverifiable risks.

2. **By passing federal legislation to address current harms now, Congress can establish the expectation and culture of accountable and governed AI—and create the frameworks of protection for future risks as they unfold.**

We are not starting from scratch. Congress should draw from a rich body of research[10] and leading policy approaches[11] to legislate ways not only to curb the current harms of AI but to create an ecosystem of accountability and control needed to concretely identify, mitigate, and avoid emerging risks.

Increasingly, a set of core principles are guiding best practice to address AI harms. **Principles like safe and effective systems, protection from discrimination, data protections, transparency, and human fallback have formed the foundation of landmark policy frameworks.**[12] Bipartisan legislation applying these principles with the force of law is the right direction for congressional action to protect Americans now.

Just as critically, these principles build a robust foundation for tangibly identifying future risks as they emerge. The difficulty with hypothetical AI futures is that *we don't know what we don't know*. The simplest way to seriously interrogate risks and build a grounded understanding of the unintended consequences of AI systems is to **mandate rigorous testing, evaluation, and reporting regimes.**

We're already seeing this principle have positive effects. Because researchers were able to test the ability of generative AI to create new poisons and generate novel bioweapons, they were able to better understand the AI model's capabilities and *how* such a model might present an existential threat to safety.[13] Strong transparency obligations around the data sources used by LLMs, for example, would enable policymakers and AI developers to better understand what information to exclude from datasets so as not to inadvertently generate blueprints for chemical weapons.

Further, principles of pre-deployment testing, algorithmic impact assessments,[14] and requiring a "human in the loop" are evergreen across the range of algorithmic and AI harms. They are useful not just to mitigate the risks of present day issues like algorithmic discrimination, but would also enable

---

9 https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/; https://www.axios.com/2023/11/07/ai-regulation-chat-gpt-us-politics-poll.
10 https://facctconference.org/2023/harm-policy.
11 https://www.nist.gov/itl/ai-risk-management-framework; https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.
12 https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
13 https://arxiv.org/abs/2304.05332.
14 https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

accountability regimes and human interventions to discover and safeguard against novel threats, such as the risk of AI hijacking critical infrastructure systems.

Critically, leading policy frameworks[15] and research[16] center the importance of not using AI where the system, after undergoing testing and assessment of its impacts, is found unsafe, ineffective, or violative of civil or human rights. **The option to not use AI** is important for systems that might arbitrarily deny people housing, decent elder care, or other key services or benefits. It's also an irreplaceable policy lever for those worried about AI's existential threat.

3. **Aligning AI with societal values should invite a broad, participatory dialogue on the purpose and role of AI.**

As today's AI Insight Forum indicates, "alignment" has increasingly become a north star for many developing AI. Generally, and especially within the tech industry, it refers to the goal of ensuring that AI systems behave in a way that *aligns* with how the developer intended the system to behave, reflecting the values that the developer imbued in the system.[17]

By focusing on the technical outputs of a system, current alignment work is designed to mitigate unintended, unanticipated, or harmful system behavior. However, the technical framework of alignment does not give us tools to ask whether the *intended* purposes of AI systems are just or desirable. Systems can be perfectly aligned with malign purposes of developers or governments, and better technical measures for "alignment" cannot solve that. Systems can also be aligned with corporate goals—for instance, to drive profit—at the expense of societal goals, such as the elevation of democratic practice. We have seen this dynamic play out in the lack of regulatory controls over social media platforms, and the subsequent impacts on electoral safety and trust in government institutions.

Congress, accountable to democratic principles and to constituents, should take a more expansive view. The call to "align AI systems" should be a call to consider the purpose and place of AI systems in society more broadly, solicit democratic participation in technology governance,[18] and work toward shared visions of a better future. Human-centric alignment practices should be asking: *whose* risks are foregrounded, *whose* safety is protected, and *whose* values are we aligning AI systems to?[19]

---

15 https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence; https://www.nist.gov/itl/ai-risk-management-framework; https://www.federalregister.gov/documents/2023/11/03/2023-24269/request-for-comments-on-advancing-governance-innovation-and-risk-management-for-agency-use-of; https://artificialintelligenceact.eu/the-act/.

16 https://dl.acm.org/doi/pdf/10.1145/3630107

17 https://www.nsf.gov/pubs/2023/nsf23610/nsf23610.htm.

18 https://datasociety.net/library/democratizing-ai-principles-for-meaningful-public-participation/.

19 https://www.science.org/doi/10.1126/science.adi8982.

Societal values, of course, are highly contextual. "Alignment with the best outcomes for humanity" may sound compelling, but humanity is diverse and vibrant and does not have one simple "best outcome" with which to align.

As a first step to align AI systems with pluralistic societal values, Congress should legislate ways for Americans to have a say in the technology in their lives. AI researchers increasingly have identified public participation in technology design, deployment, and oversight as a critical safeguard,[20] and multiple leading AI developers and civil society organizations have invested in methods to solicit public feedback.[21] Research indicates that public participation, when done well, improves decision-making by incorporating the viewpoints of those most likely to be impacted by technologies.[22]

Congress should require developers to sustain engagement with communities about their values, preferences, and experiences with AI throughout the AI lifecycle of design,[23] development,[24] integration with real-world use,[25] and retirement of systems.[26] In certain cases, this may require companies to subsume their business interests to democratic public input.

Conclusion

The most achievable way to anticipate the risks of an uncertain future is to begin with what we can control now. Congress should draw on best practices and recent landmark government approaches, mandating the enforceable structures to mitigate AI's current and known harms through robust accountability, transparency, protection of civil rights, data protection requirements, and the option not to use AI.

These approaches help solve the problems of today, and they position us to have more control in shaping the wide range of futures before us—and to interrogate novel risks as they arise with expertise, experience, and empirical rigor.

[20] https://dl.acm.org/doi/10.1145/3551624.3555290; https://arxiv.org/pdf/2310.00907.pdf.
[21] https://arxiv.org/abs/2310.13798; https://arxiv.org/abs/2310.13798; https://dl.acm.org/doi/abs/10.1145/3491102.3502004; https://arxiv.org/abs/2303.08177.
[22] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4266250.
[23] https://dl.acm.org/doi/pdf/10.1145/3531146.3533132.
[24] https://dl.acm.org/doi/pdf/10.1145/3359283.
[25] https://datasociety.net/library/ai-in-context/.
[26] https://arxiv.org/abs/2206.03275.