FROM: FUTURE OF LIFE INSTITUTE
DATE: MAY 16, 2023
RE: AI RESEARCH, INNOVATION, AND ACCOUNTABILITY ACT OF 2023

SUMMARY OF RECOMMENDATIONS

- Require external, independent assessment of compliance with prescribed TEVV standards for critical-impact AI systems.
- Include *developers* of critical-impact AI systems in definition of "critical-impact AI organization."
- Include advanced general-purpose AI systems in definition of "critical-impact AI system."
- Remove the right to cure.
- Authorize the Secretary to prohibit a critical-impact AI organization from making a critical-impact AI system available to the public on the basis of reported information if the system presents unacceptable risk of harm.
- Increase the maximum civil penalties to scale with gross receipts of organization.
- Clarify that the limitation on disclosure of trade secrets, IP, and confidential business information applies only to the extent that information is not necessary to comply with requirements of the bill.

BILL SUMMARY

The "Artificial Intelligence Research, Innovation, and Accountability Act of 2023 (AIRIAA)," authored by Sen. Thune and co-authored by Sens. Klobuchar, Wicker, Hickenlooper, Luján, and Capito, is, to date, the most comprehensive legislative framework for regulation of high-risk AI systems that has been introduced in Congress. The AIRIAA has several major provisions that seek to address a range of potential harms from advanced AI systems through transparency, recommended standards and best practices, mandatory assessments, and self-certification.

These provisions: (1) commission research and development of standards for verifying provenance of AI-generated and human-produced content, including through watermarking; (2) commission research and development of standardized methods for the detection and understanding of anomalous behavior by AI systems, and safeguards to mitigate "potentially adversarial or compromising anomalous behavior"; (3) commission a study into the barriers and best practices for the use of AI by government agencies; (4) require notice to users when a covered internet platform uses generative AI to generate content that the user sees; (5) require

transparency reports for “high-impact AI systems” to be submitted to the Secretary of Commerce prior to deployment that detail design and safety plans for the system; (6) require the National Institute on Standards and Technology (NIST) to provide sector-specific recommendations to individual Federal agencies on the oversight of high-impact AI systems within their domain, and require each agency to submit a formal written response to the Director of NIST indicating whether or not those recommendations will be adopted; (7) require an organization that deploys a “critical-impact AI system” to perform a risk management assessment at least 30 days before the system is made publicly available and at least biennially thereafter, and to submit a report to the Secretary of Commerce detailing that assessment; (8) establish an advisory committee to provide recommendations on standards for testing, evaluation, validation, and verification (TEVV) of critical-impact AI systems; (9) require the Secretary of Commerce to establish a 3-year implementation plan for self-certification of compliance with established TEVV standards by critical-impact AI organizations; and (10) establish a working group for furthering consumer education related to AI systems.

We applaud the Senator on his effort to develop a detailed framework toward AI governance infrastructure in both the public and private sector. The framework recognizes that AI systems vary in their potential risk, and takes the reasoned approach of stratifying requirements on the basis of impact. Provisions for development of digital content provenance and risk mitigation standards, detailed guidance on responsible government AI use, and transparency, assessment, and certification requirements for the highest-risk AI systems are commendable, and have the potential to reduce the risks presented by advanced AI systems.

Unfortunately, **several shortcomings in the bill in print significantly limit the effectiveness of these provisions, rendering the proposed bill insufficient to adequately address the major risks these high-risk AI systems present**. These shortcomings, discussed in the following section, can be easily resolved by relatively minor amendments to the bill’s provisions, which would render the bill a promising avenue toward thoughtful oversight of this promising yet risky technology. Failure by Congress to effectively mitigate these risks could lead to large-scale harm that could shutter the broader AI industry and stymie future innovation in AI, depriving the public of the anticipated benefits of the technology. The catastrophic nuclear meltdown at Three Mile Island presents an illustrative analogy here, as it ultimately precipitated the foreclosure of advancement in nuclear technology.

We encourage the Senator to address the following concerns to ensure the bill accomplishes its laudable objective of mitigating the risks of advanced AI systems and realizing their considerable benefits.

COMMENTS AND RECOMMENDATIONS

Definitions

Critical-impact AI systems

The bill in print defines "critical-impact AI systems" to mean non-defense/intelligence AI systems that are "used or intended to be used […] to make decisions that have a legal or similarly significant effect on" the collection of biometric data without consent, the management or operation of critical infrastructure and space-based infrastructure, or criminal justice "in a manner that poses a significant risk to the rights afforded under the Constitution of the United States."

- *"Used or intended to be used"* — Demonstrating intent can be extremely challenging, and even if a system was not specifically designed for a particular use, developers and deployers should be cognizant of likely uses outside of their intended scope and evaluate accordingly. These are largely general-purpose technologies, so their intended use isn't always apparent - GPT-4, for instance, has revealed hundreds of use cases that developers did not intend or anticipate prior to deployment, and future generations are likely to demonstrate even more emergent capabilities. The definition should thus include reasonably expected uses in addition to intended uses in order to avoid a loophole allowing circumvention of the bill's requirements. This recommendation is also applicable to the analogous language in the definition of "high-impact AI system."

- *"In a manner that poses a significant risk to rights […]"* — The standards and assessments prescribed by the bill are intended to evaluate the risks to rights and safety posed by critical-impact systems. The requirement that the systems initially meet that criterion in order to be subject to the bill's standards and assessments raises the risk non-compliance if the organization lacks (or alleges to lack) a prima facie expectation that their system poses a significant risk to rights or safety, whether or not it actually does. This provision should be struck, as systems involved in these critical functions should be presumed to pose a significant risk to constitutional rights or safety, with compliance with standards and assessment then evaluating the scale and scope of that risk. This recommendation is also applicable to the analogous language in the definition of "high-impact AI system."

- *Advanced general-purpose AI systems are not included in definition* — Any general-

purpose AI system (GPAIS) that exceeds a certain threshold of capability poses an inherent risk of dangerous emergent behavior and should undergo the requisite scrutiny to ensure safety before deployment. We believe systems more capable than GPT-4 should fall into this category, based on the positions of various experts and existing evidence of emergent capabilities in that class of systems. We recognize that the bill prefers a use-based approach to regulation, but GPAIS do not fit neatly into such a paradigm, as they are by definition multi-use, and can exhibit capabilities for which they were not specifically trained. While their use for, e.g., collection of biometric data, management of critical infrastructure, or criminal justice may not be intended or anticipated, such uses may nonetheless arise, necessitating preliminary evaluation of their effects on rights and safety. Accordingly, inclusion of these systems in the category of "critical-impact" is imperative.

Significant risk

The bill in print defines "significant risk" to mean "*a combination of* severe, high-intensity, high-probability, *and* long-duration risk of harm to individuals." Requiring a combination of these factors would exclude many major risks from AI, which tend to be high-intensity and/or long-duration, but relatively low-probability. For instance, failure or compromise of a critical infrastructure management system may be low-probability, but would have catastrophic, high-intensity harmful effects. Any of these characteristics in isolation should qualify as a significant risk. We thus recommend striking "a combination of" and changing the final "and" to an "or" to capture the spectrum of major risks AI is likely to present.

Critical-impact AI organization

The bill in print defines "critical-impact AI organization" to mean a non-governmental organization that serves as the *deployer* of a critical-impact AI system. However, the requirements placed on critical-impact AI organizations under this bill should also extend to developers of critical-impact AI systems. This would improve transparency by allowing developers to more readily provide requisite information to deployers, and would ensure that developers are accounting for safety by design in the process of developing the system itself. Developers are most directly exposed to the mechanics of system development and training, and as such are in the best position both to assess risk and to mitigate risks early in the value chain to ensure they do not proliferate. Presently, developers are also significantly larger and better-resourced organizations than most deployers, as the resources necessary to train critical-impact AI systems can be prohibitive for small companies who may nonetheless be equipped to

use those systems. Evaluation and transparency from developers pursuant to the bill's requirements for critical-impact AI systems would reduce the compliance costs for downstream small-business deployers, as more of the necessary information can be carried over from assessments and reports produced by well-resourced developers. We recommend including developers and deployers of critical-impact AI systems under the definition of “critical-impact AI organization,” and further recommend striking the exclusion of deployers from the definition of “developer,” since this would otherwise absolve developers of these systems who also act as deployers from any obligations under the bill.

Transparency and Certification

Self-certification

Perhaps most concerning is that the bill in print does not require *any* independent evaluation of compliance with prescribed standards, and relies entirely on self-attestation to the Secretary that critical-impact AI organizations are compliant with the prescribed safety standards. Self-certification is mainly useful in that it can be adopted quickly, but self-certification as a long-term solution for ensuring the safety of critical-impact AI systems is woefully insufficient. It could be extremely dangerous to have those producing and deploying the AI systems managing our critical infrastructure, who may be driven by a profit motive and have failed in the past to adequately predict the capabilities of their systems, determining whether or not their own assessments were sufficient to comply with standards and determine if their systems are safe. These companies grading their own homework provides no external assurance that compliance has actually been achieved, and in this high-risk of a circumstance, that can lead to catastrophic outcomes. We do not rely on self-certification for assessing the safety of our airplanes, and we should not rely on self-certification for assessing the safety of these critical systems.

The three year period allotted for the development of TEVV standards provides a reasonable window to iron out any ambiguities in how to structure an independent mechanism for certifying compliance with the issued standards. To the extent that reasonable standards can be developed in a three year period, so too can a mechanism for independently verifying compliance with those standards.

Limitations on information disclosure requirements

The bill in print specifies that it shall not be construed to require deployers of high-impact AI systems or critical-impact AI organizations to disclose information relating to trade secrets, protected IP confidential business information, or privileged information. These limitations

severely limit the usefulness of the reporting requirements and hamper effective assessment of the systems, as most pertinent information relating to the AI system can be construed to fall into these categories, and omissions on this basis would not be apparent to the Secretary.

In other words, if the deployer cannot be required to disclose confidential business information/trade secrets/intellectual property to the Secretary, the Secretary has no way to evaluate completeness of the assessment since the information they can evaluate on is substantially limited. These sensitive types of information are routinely disclosed to government agencies for a number of functions, and, to the extent that information is necessary to effectively evaluate the system and meet the bill's requirements, should be disclosed here as well. Accordingly, we recommend amending the rules of construction to specify that the limitations on compelled information disclosure apply only to the extent that the information is not necessary for satisfying the bill's requirements, and to reiterate the confidentiality of information disclosed to the Secretary in this manner.

Enforcement

Right to cure

The bill in print empowers the Secretary to initiate an enforcement action and recover penalties for violation of the bill's provisions only if the violator does not "take sufficient action to remedy the non-compliance" within 15 days of notification. The inclusion of this right to cure provides little incentive for those subject to the bill's requirements to comply, unless and until they receive notice of non-compliance from the Secretary. To the extent that compliance entails any cost to the company, it would not be in the financial interest of the company to comply prior to receiving notice, which in practical terms means a long latency before companies actually comply, additional administrative costs to the Secretary, and the possibility that more companies out of compliance fall through the cracks with no incentive to comply. The companies at issue here, which are predominately large and well-resourced, generally have the legal resources to remain up to speed on what laws are being passed, so a right to cure serves little benefit with significant cost.

Maximum civil penalty

The bill in print caps civil penalties for violation of the bill at the greater of "an amount not to exceed $300,000; or […] an amount that is twice the value of the transaction that is the basis of the violation with respect to which the penalty is imposed." Because the latter may be difficult to quantify - not all violations will occur within the context of a transaction - the former will likely

often apply, and seems insufficient to incentivize compliance by large companies. While $300,000 may have some effect on the budget of a very small company, the AI industry is presently dominated by a small number of very large companies, most of which would be virtually unaffected by such a penalty. For reference, $300,000 would constitute 0.00014% of Microsoft's annual revenue, making such a penalty an extremely small line-item in their annual budget. Rather, the maximum civil penalty should be based on a percentage of the annual gross receipts of the organization such that the impact of the penalty scales with the size and resources of the organization. Additionally, it is not clear whether failure to, e.g., perform a required assessment at all would constitute a single violation, or whether each required assessment item not reported to the Secretary would be considered a separate violation. In the latest consensus draft of the European Union's AI Act, civil penalties are capped as the higher of a set monetary amount or a percentage of the organization's worldwide turnover in the previous year, ranging from 1.5% - 7%, depending on the severity of the violation. A similar model could be employed here.

Prohibition of deployment

The bill in print explicitly deprives the Secretary of the authority to prohibit a critical-impact AI organization from making a critical-impact AI system available to the public based on review of a transparency report or any additional clarifying information submitted to the Secretary. This undermines the bill's intent to ensure that the systems being deployed are safe. If an assessment reveals material harms that are likely to occur from the release of the system, it is critical that the Secretary be able to take action to ensure that the unsafe system is not put on the market. For essentially all consumer products, if there is reason to believe the product is unsafe, there is an obligation to remove it from the market. In this case, these assessments are the most comprehensive evaluation of the safety of a system, so barring the Secretary from acting on the basis of that information seems inconsistent with the safety standards we otherwise expect of consumer products, which generally have significantly less potential for catastrophic harm relative to these systems. While in some cases there are existing laws that could penalize resulting harms irrespective of whether AI was involved, these laws only apply once the harm has already occurred. In critical-impact circumstances, the scale of harms would be enormous, and would likely entail significant infringement on rights, property damage, suffering, or loss of human life that could be avoided if the Secretary were empowered to prevent deployment until sufficient safeguards are adopted.

The bill does permit the Secretary to prohibit the deployment of a critical-impact AI system if the Secretary determines that the organization *intentionally* violated the act or any regulation issued

under the act. However, intent would be remarkably difficult to prove in order to actually enforce this. It would effectively require internal documentation/communications that explicitly communicate this intention, which are unlikely to be available to the Secretary in determining whether to take such action. Additionally, violations of the act would generally be failures to perform certain assessments or submit reports, but so long as the assessments are conducted and reports are submitted, the actual contents of those assessments and reports cannot be acted upon. This significantly hampers the Secretary's capacity to protect the public from dangerous systems.

To put it more succinctly, if a system is being implemented in a critical-impact context (e.g. supporting the power grid, ensuring clean water, protecting biometric information, etc.), a single failure can be catastrophic. If there is reason to believe the system is likely to fail, it seems it should be well within the remit of the Secretary to prevent those risks from being realized.

Advisory Committee/Working Group

Exclusion of civil society

As it is currently drafted, the Artificial Intelligence Advisory Committee established to provide advice and recommendations on TEVV standards and certification of critical-impact AI systems would specifically require representation of two different types of technology companies, but no representation of civil society groups focused on ethical or safe AI. Similarly, while the working group relating to responsible consumer education efforts for AI systems established by the bill in print includes representation of nonprofit technology industry trade associations, it excludes independent civil society representation. Given the valuable, public-interest perspective civil society organizations can provide, we recommend including required representation of nonprofit organizations with a substantial focus on the safe and ethical use of AI in both the Advisory Committee and the working group.