Seven Provisions in the National Defense Authorization Act with High Potential to Accelerate AI Security

The U.S. House of Representatives passed the 2026 National Defense Authorization Act (NDAA), sending it for consideration by the U.S. Senate before the end of the year. Several provisions of the NDAA concentrate on the security of the most advanced AI models and systems. Based on a preliminary reading, **seven provisions hit the bull's eye on AI security hazards that, when occurring in high-stakes environments such as defense or intelligence agencies, could escalate into national security threats**. We review these provisions briefly below.[1] Next, we explain why these provisions have high potential to accelerate AI security and how they intersect with Apollo Research's efforts to address the threat that misalignment in AI models and systems could undermine their trustworthiness and efficacy, as well as human command.

- **Section 224** enables the Secretary of War to establish **National Security and Defense Artificial Intelligence Institutes** within eligible host institutions. The vision for these Institutes is to focus on foundational science for AI systems in the national security and defense sector, including by developing sector-specific test beds for pre-deployment evaluation of AI systems.
    - National Security and Defense AI Institutes could contribute to the science of evaluation by, among other things, undertaking national security threat modeling (including threats arising from misaligned behavior), designing national security benchmarks, as well as building prototypes for real-world honey pots in defense deployment contexts.

- **Section 1512** requires the Secretary of War, in consultation with other relevant federal agencies, to develop and implement Department-wide **policy for the cybersecurity and associated governance of AI models and systems used in defense applications**. Among other things, this policy must address how the U.S. Department of War (DoW) protects against AI security threats (e.g., model tampering, data leakage, adversarial prompt injection, and model jailbreaks), including by developing AI security best practices inspired by industry-recognized frameworks, and developing standards for the governance, testing, auditing, and monitoring of AI systems' integrity and resilience.

- **Section 1513** requires the Secretary of War to develop a risk-based and risk-proportional framework for the implementation of **physical and cyber security standards and best practices** against, among others, insider threat risks, data poisoning risks, and adversarial tampering. The frameworks must also include security management practices (e.g., continuous monitoring, and incident reporting procedures) as well as an evaluation of commercially available platforms for continuous monitoring and assessment of AI systems.

[1] The following is not an exhaustive review of the NDAA, nor an analysis of all the ways in which AI models and systems are impacted by the NDAA. Additional provisions addressing the application or security of AI models and systems include, for instance, Sections 245, 347, 547, 1007, 1019, 1234, 1513, 1532, 1534, 6602, 6604, and others.

- The DoW-wide policy and the standards and best practices under Sections 1512 and 1513 could help build DoW preparedness against outsider and insider threat. Among other things, outsider and insider malicious actors could induce misalignment in AI models and systems deployed in high-stakes contexts (such as defense and intelligence) by deliberately designing a model that is misaligned, or poisoning an AI model's training data to 'trigger' misalignment in AI models and systems of any size. In other words, malicious actors could cause an AI model's goals and, therefore, its behaviors to deviate from what humans (e.g., commanders) intended. As AI models and systems become more capable and entrenched within the national security apparatus, inducing misalignment could potentially lead to loss of control threats. In a recent research paper, we explained why and how AI misalignment could act as a catalyst for loss of control in high-stakes deployment contexts such as the military, and outlined technical and governance tools that decisionmakers can adopt to mitigate this risk.

- **Section 1533** requires the Secretary of War to establish a **Cross-Functional Team** for AI model assessment and oversight, led by the Chief Digital and Artificial Intelligence Officer (CDAO). Among other things, this Cross-Functional Team is tasked with developing a "**standardized assessment framework and governance structure**" **to evaluate AI models *currently* used by the DoW** as well as **guidelines for evaluating *future* AI models** being considered by the DoW, including "procedures" and "methodologies" for testing and assessing AI models.

  - Section 1533 closely reflects one of the priorities set forth by the AI Action Plan, as well as one of our current policy and research focus areas. The AI Action Plan tasks DoW, the Office of the Director for National Intelligence (ODNI), the National Institute of Standards and Technology (NIST), and the Center for AI Standards and Innovation (CAISI) to refine DoW's responsible AI frameworks and issue an Intelligence Community (IC) standard on AI assurance. To support DoW, ODNI, NIST, and CAISI in implementing the AI Action Plan, earlier this year we published a policy memorandum outlining recommendations for the testing and evaluation of today's and tomorrow's frontier AI systems developed for national security. Specifically, we drew from Apollo Research's experience in evaluating frontier AI systems for misaligned behavior, including scheming, and put forward actionable recommendations on *what*, *when* and *how* DoW or the IC should thoroughly vet when they acquire frontier AI capabilities. To summarize, we suggested that DoW and the IC leverage existing developmental and operational testing & evaluation (T&E) pipelines to perform a suite of scheming and control evaluations, and make informed decisions by comparing the results of these evaluations against acceptable failure rates pre-defined in testing and evaluation master plans. For more information, please read this policy memorandum or the relevant blog post.

- **Section 1535** requires the Secretary of War to establish an **Artificial Intelligence Futures Steering Committee**. The Steering Committee is tasked with developing a "**proactive policy**" for the DoW's evaluation, adoption, governance, and risk mitigation of AI systems that are **more**

advanced than any existing advanced AI systems, including advanced AI systems that "approach or achieve **artificial general intelligence**." The Steering Committee will also be responsible for developing a strategy for the risk-informed adoption, governance, and oversight of advanced or general-purpose AI by the DoW, including identifying "**guardrails** to maintain, to the extent practical, appropriate **human decision making**," and assessing "potential effects on commanders of operational commands, including effects related to **maintaining oversight of mission command** when using artificial intelligence and the capability for humans to **override** artificial intelligence through technical, policy, or other operational controls."

- **Sections 6601** and **6603** amend sections of the 2023 and 2025 Intelligence Authorization Act to require the Director of National Security Agency (NSA) to identify vulnerabilities in advanced AI systems, with a focus on security and cyber security risks related to theft or sabotage by nation-state adversaries, and the Chief Artificial Intelligence Officer (CAIO) of the IC to "establish standards for testing of artificial intelligence models in proportion to risk, including benchmarks and methodologies for … **trustworthiness**."

    - Section 1535 could be highly consequential for AI security, as it creates a Committee that is tasked with *proactively* addressing future, highly advanced, systems that "approach or achieve **artificial general intelligence**." Building preparedness for the potential risks posed by more advanced, future AI systems, including artificial general intelligence, is of crucial importance, as human- or above-human- level AI capabilities, combined with affordances, permissions and high-stakes deployment contexts could unlock national security threats and risks of loss of control.
    - Sections 1535, 6601, and 6603 could be highly impactful for AI security for a second reason. In a policy memorandum published earlier this year, we concentrated on why and how the DoW and the IC should strengthen the principles of **AI model reliability** (i.e., an AI model dependably performs its intended functions without undesirable behaviors) and **AI model governability** (i.e., an AI model demonstrating unintended behaviors can be disengaged or deactivated) by thoroughly vetting AI models and systems that the DoW and the IC acquires. These principles, which are cornerstones of previous responsible AI policies by the DoW and the IC (e.g., DoD AI Ethical Principles, DoD DT&E Guidebook, and DHS Directive 139-08), are now closely reflected in Sections 1535, 6601, and 6603 of the NDAA. Specifically, Section 1535 tasks the AI Futures Steering Committee to identify guardrails that can maintain "human decision making" and "oversight of mission command" and ensure the ability to "override" AI systems if needed. Section 6603 tasks the IC CAIO to establish AI testing standards for trustworthiness. Therefore, Sections 1535, 6601, and 6603 corroborate the importance and urgency of strengthening and future-proofing AI model reliability and AI model governability.

If you are interested in learning more, please contact us at matteo@apolloresearch.ai and charlotte@apolloresearch.ai.