
Cognitive Insights for Artificial Intelligence

CIfAI on Recent Announcement of Voluntary Commitments by Leading AI Companies to Manage AI Risks
Comment by Monica Lopez, PhD
July 25, 2023

Upon the heels of last week's announcement from the Biden-Harris Administration on Securing "Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI", we at CIfAI provide several remarks.

First and foremost, this is a significant step forward in ensuring AI-enabled products and services are safe, secure and trustworthy prior to entering end users' hands. AI technologies pose multiple risks, and they necessitate adequate guardrails in place. The voluntary commitment of the seven leading AI companies in the U.S. –Amazon, Anthropic, Inflection, Google, Meta, Microsoft, and OpenAI– comes at a critical moment in the development of the nation's stance on responsible AI governance. These commitments gain further relevance as international cooperation becomes ever more important with the European Union's Artificial Intelligence Act in its final phase of inter-institutional negotiations since the parliament's vote on June 14th. As an AI Governance, Risk and Regulatory Compliance platform with a mission to empower enterprises to adopt and scale AI with confidence, we uphold this initiative in the U.S. and stand ready and available to participate in efforts moving forward by the administration and said companies.

Second, in acknowledgment of the U.S.'s position to forge a balance between supporting innovation and promoting responsible AI governance across the nation, we underscore the following which was committed to for each principle of safety, security, and trust:

(1) Safety - Ensuring the safety of AI systems prior to deployment
- Companies commit to internal and external security testing prior to releasing their system(s).
- Companies commit to sharing information on managing AI risks within their system(s) across industry and with governments, civil society and

academia.

(2) Security - Building AI systems with security as a priority
- Companies commit to investing in robust cybersecurity and insider threat safeguards for their system(s).
- Companies commit to third-party discovery and reporting of system vulnerabilities.

(3) Trust - Earning the public's trust through transparency of AI systems
- Companies commit to developing robust technical mechanisms for their system(s) to ensure users know when content is AI-generated.
- Companies commit to publicly reporting their system(s)' capabilities and limitations, and the areas of appropriate and inappropriate use.
- Companies commit to developing and deploying their system(s) to address society's grand challenges and benefit humanity.

Third, given our deep practical experience in auditing AI systems and supporting in-house academic research work, we advance further suggestions to guarantee the fulfillment of the above stated commitments:

- A commitment to robust long-term research partnerships and collaborations between industry, academia and the public sector.

Cross-communication between innovative research and real-world solutions is paramount. From bias and discrimination and model effectiveness and validity to data protection and privacy and transparency and explainability, to name a few areas of debate, pressing questions around the practice of AI assurance and their needed implementations abound. Innovations necessitate thorough study, testing, evaluation and standardization. This can be achieved through a collective-driven and consensus-bound manner.

- A commitment to the development of robust human-in-the-loop systems.

Clarity on the level of automation of an AI system is fundamental. End users need to know with what type of system they are interacting, and the measures taken to mitigate any identified risks. This can be achieved through a human-centered approach that advocates human well-being and enhancement of the human condition through knowledge building and provision of information such as through multistakeholder transparency of the system's operation and interpretability of its output, including a straightforward mechanism for opting out, and adverse incident reporting that is prompt, continuous, and straightforward.

- A commitment to convening and working with diverse stakeholders, particularly those outside large and already well-established technology leaders.

The ubiquity of AI systems and the depth of impact on users is too large to ignore. Small and medium-sized enterprises (SMEs) offer unique value from the direct frontline product-to-consumer relationship they have with their customers and the day-to-day management of customer understanding and experience. No voice is too small. This engagement can be

achieved through active participation with SMEs to address positive and negative outcomes for affected parties.

- A commitment to increase the adoption of risk management protocols.

The identification and evaluation of an AI system's risks and safety issues is now imperative throughout the research and development process and beyond. AI systems need to be reliable and end users' trust needs to be gained. This can be achieved through the responsible validation of the AI system's behavior through standardized documentation, database reporting, and continuous review and updating throughout the entire lifecycle of the system.

We commend the Biden-Harris Administration on this initiative, and as this agenda advances, we underscore the importance of working together with our allies and partners to establish a robust international framework to govern the development and use of AI. It is through this international collaboration that we will ensure the benefits of AI for generations to come.