Research Agenda

One Size Does Not Fit All

Assessment, Safety, and Trust for the Diverse Range of AI Products, Tools, Services, and Resources

Author
Heather Frase, PhD

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

February 2023

Executive Summary

The spread of artificial intelligence across business, defense, and security sectors has the potential to improve the speed of operations, provide new capabilities, and increase efficiencies. Along with the integration of AI comes an upsurge in risk and potential harm from AI accidents, misuse, and unexpected behavior. The growing concern about AI having unforeseen negative impacts on U.S. commercial, social, infrastructure, and national security highlights the need for AI assessment that can help reduce potential harm from AI and ensure that AI applications and technologies are safe and trustworthy.

The Center for Security and Emerging Technology has published studies related to AI safety, accidents, and testing. Building on this work, CSET has launched a new line of research titled “AI Assessment” to investigate the development and adequacy of current AI assessment approaches, along with the availability and sufficiency of tools and resources for implementing them. Specifically, the research will:

1. Understand and contribute to the development and adoption of AI standards, testing procedures, best practices, regulation, auditing, and certification.
2. Characterize the wide variety of AI products, tools, services, data, and resources that influence AI assessment.
3. Understand the needs for additional infrastructure, academic research, tools, or budgetary resources to support demonstration and adoption.
4. Explore the global differences and similarities of AI assessment, standards, and testing practices among various sectors and government entities.

There is no simple one-size-fits-all assessment approach that can be adequately applied to the diverse range of AI. AI systems have a wide variety of functionalities, capabilities, and outputs. They are also created using different tools, data types, and resources, adding to assessment diversity. A collection of approaches and processes are needed to cover a wide range of AI products, tools, services, and resources. Additionally, because the number and frequency of AI creation will greatly increase, resources need to include techniques and tools for scaling assessment, handling the variety and quantity of AI systems. With new AI innovations, assessment needs may change. This research will provide a foundation for assessment that can be adapted to future needs. It will also provide a better understanding of the current U.S. needs and capabilities for AI assessment, and support decisions on AI policy, resourcing, research, and national security.

Introduction

AI systems are becoming more capable, more pervasive, and easier to create. Over the last few years, AI has made large capability leaps in the fields of language processing, autonomous vehicles, medicine, and image creation. AI is being integrated, often invisibly, into more and new aspects of daily life, critical infrastructure, and our government institutions. Governments and organizations are experiencing strong internal and external pressures to take advantage of AI's new capabilities. AI is becoming easier to create, because of open-source software and code along with software packages that accelerate and democratize its development. These software packages enable the nontechnical general public to create, develop, and deploy AI without code or a technical understanding.

The combination of more capable, more pervasive, and easier AI creation can result in an environment with a high probability of harmful AI events—unless we implement standards, techniques, and processes to control AI usage, as well as to improve safety and trustworthiness. The AI assessment line of research will investigate the adequacy of, and make recommendations on, the adequacy of AI standards, techniques, and processes. It will also investigate how these standards inform risk mitigation and the assessment of AI safety and trustworthiness.

A variety of stakeholders are developing a broad range of AI assessment approaches, including standards, testing procedures, best practices, regulation, risk management, auditing, certification, conformity testing, and red-team exercises. These approaches provide a range of information and are often used at different stages of an AI technology's life cycle of development, deployment, and maintenance. But overall, these approaches are best thought of as tools that can help stakeholders identify AI risks, understand if an AI technology will meet their needs, and inform the conditions where AI performance will be degraded.

The evolution and adoption of these and other AI assessment processes will impact both U.S. security and economic prosperity. If they are of poor quality or unevenly adopted, they could hinder AI development and scaling across key sectors, as well as increase the number of harmful AI events, or even create an onerous burden on the economy. For the United States to lead in AI and related emerging technologies, AI assessment should be executable across the diverse range of AI products, aligned with partner governments and organizations, and not create deficiencies that may hinder national competitiveness relative to adversaries.

Research Agenda: Initial Phase

The initial phase of this research will involve foundational work that supports later research and products. It will also begin focused work on the adoption of AI assessment processes, AI harm and vulnerabilities, and software that accelerates AI use.

Using other transformative technology fields as a point of reference, it is not unreasonable to expect that the development of standards, processes, and policies for safe and trustworthy AI is likely to take a long time—potentially decades. The initial phase of this CSET research attempts not just to inform this long-term development process, but also shorten its overall timeline, and contribute to near-term AI assessment solutions needed today.

Foundational Research

At a high level, the foundational research focuses on *what is needed* and *what we have*, which then allows analyzing the gap between the two, or, in other words, *what should we get*? What are all of the characteristics of AI development and use that we need to account for when assessing AI quality, safety, and trust? What resources and capabilities do we currently have for demonstrating AI quality, safety, and trust? Answering these questions will feed into work to analyze capability and resource gaps for AI assessment processes. It will also allow for the assessment of how gaps could impact U.S. security and economic prosperity. At an international level, it will provide a basis for comparing U.S. capabilities to those of its partners and adversaries.

This research will identify which critical AI features may impact the utility and application of AI assessments. This will provide an understanding of *what is needed* to assess the safety and trustworthiness of the wide variety of AI tools, products, and services. This investigation of features is critical because the different features will require different metrics, tools, procedures, and resources for the demonstration of AI standards and testing best practices.

Since having sufficient resources is a critical part of demonstrability, it is important to understand *what we have*. During the initial phase, this work will look at what open-source and commercially available resources, processes, guidance, code, and software are available. Research into available resources will support later gap analysis that identifies where there needs to be additional infrastructure, academic research, tools, or budgetary resources to support demonstration and adoption. By understanding these capability gaps, this line of work will be able to provide U.S. policymakers with recommendations on AI safety and trustworthiness, as well as on assessment research and infrastructure.

Adoption of AI Assessment Processes

AI assessment processes that are not adopted will have little benefit and not mitigate AI risk. To better understand the varied approaches to adoption of AI assessment processes, as well as assess barriers and lessons learned, this research will examine how other fields have developed and adopted their standards. It will also explore how organizations are addressing existing AI regulations and policies.

This research effort will support recommendations on how to make AI assessment processes scalable. In this context, an AI assessment process is scalable if it is general enough that it can be adapted and applied to the wide range of AI systems created, but specific enough that information it produces can inform action to mitigate risk.

AI Risks and Vulnerabilities

Building safe, secure, and trustworthy AI systems involve knowing their risks and vulnerabilities. Discovering, understanding, and identifying risks and vulnerabilities is important to AI developers (so that they can mitigate them before deployment), end-users (so that they can be informed consumers of AI products), policymakers (so that they command and resource action), and oversight organizations (so that they can ensure that applicable policies, laws, regulations, and ethical standards are followed).

Some AI risks for harm can be deduced through research and collaboration. However, others are only discovered after the AI systems are deployed. Thus, our research will investigate and analyze AI incidents by providing a structured way of characterizing vulnerabilities and harm.

This research will continue CSET's joint effort with the Responsible AI Collaborative and their Artificial Intelligence Incident Database. CSET will help increase the number of incidents in the database and refine the definition of AI harm. The research will also revise CSET's AI harm taxonomy to distinguish between AI harm and vulnerabilities, as well as improve annotation consistency. The updated taxonomy will support research on AI harm, such as trend analysis, and enable the forecasting of future harm and vulnerabilities. It will collect data so that occurrences, trends, and emerging types of harm, risk, and vulnerabilities from AI systems can be identified and mitigated through policy, best practices, and academic research.

This research can inform the U.S. government and international organizations on how to define and structure data collection on AI incidents. Currently, there are multiple groups seeking to establish AI incident collection systems. Data collection alone has little value if actionable information cannot be extracted. The research's iterative approach to improving the taxonomy and information extraction will advance best practices for gathering and characterizing AI incident data.

AI Enablers

AI development and deployment is becoming democratized, reducing the barriers that hinder the growth of AI systems. AI enablers are the products, tools, services, data, and resources that accelerate and simplify the development, deployment, and maintenance of AI. They democratize and increase scalability of producing AI systems by facilitating customization, automating processing, or replacing hands-on coding and technical expertise with simple user interfaces. But if AI enablers do not incorporate AI standards and testing best practices, there will be a collection of AI products that will not conform to standards and will have a higher likelihood of producing unintended harm.

This research will investigate different types of AI enablers such as automated machine learning, low-code or codeless AI, or composable AI. It will explore the commercial market size and growth trends in AI enablers and how they can play an important role in AI standards adoption.

During the initial phase, the research will also investigate which U.S. government agencies are purchasing AI enablers and estimate how much is being spent. This will provide information about the impact of AI enablers on U.S. security, defense, and operations.

As the deployment and development of AI systems democratizes and scales, techniques and tools for demonstrating safe AI also need to be easily scalable, accessible, and usable by the general public. It is unclear if AI enablers have sufficient tools for assessing the safety and trustworthiness of the systems they produce. This research will investigate what assessment tools AI enablers provide for the AI systems that they can produce. U.S. government agencies purchase AI enablers. Based on the National Defense Authorization Act for Fiscal Year 2023, they are likely to acquire more. This research will help to inform what policies should be created regarding the use of these tools and the AI systems that they create.

Conclusion

The initial phase of research will provide U.S. policymakers, government agencies, and other stakeholders with foundational analysis about AI assessment, safety, and trust for the diverse range of AI products, tools, services, and resources. It will provide information on the current status and needs for AI assessment, standards, and testing, allowing stakeholders to take steps for risk mitigation and improvement. Future work will build upon the initial research phase that is U.S. focused, examining how the United States' AI assessment tools, resources, legislation, and policies compare to those of its partners and adversaries. This initial effort will lay the foundation for promoting safety, understanding how to demonstrate trust, and mitigating potential harm for the broad range of AI systems. This includes AI systems that are narrowly focused (e.g., diagnosing why a vehicle is not operating smoothly), to AI systems that directly impact many people (e.g., screening job applicants), all the way to AI systems used to assist decision makers at the highest levels of government.

Author

Heather Frase, PhD, is a senior fellow with CSET.

Acknowledgments

For feedback, assistance, and editing, the author would like to thank Mina Narayanan, Karson Elmgren, Christian Schoeberl, Shelton Fitch, Margarita Konaev, and Igor Mikolic-Torreira.

Document Identifier: doi: 10.51593/20220051