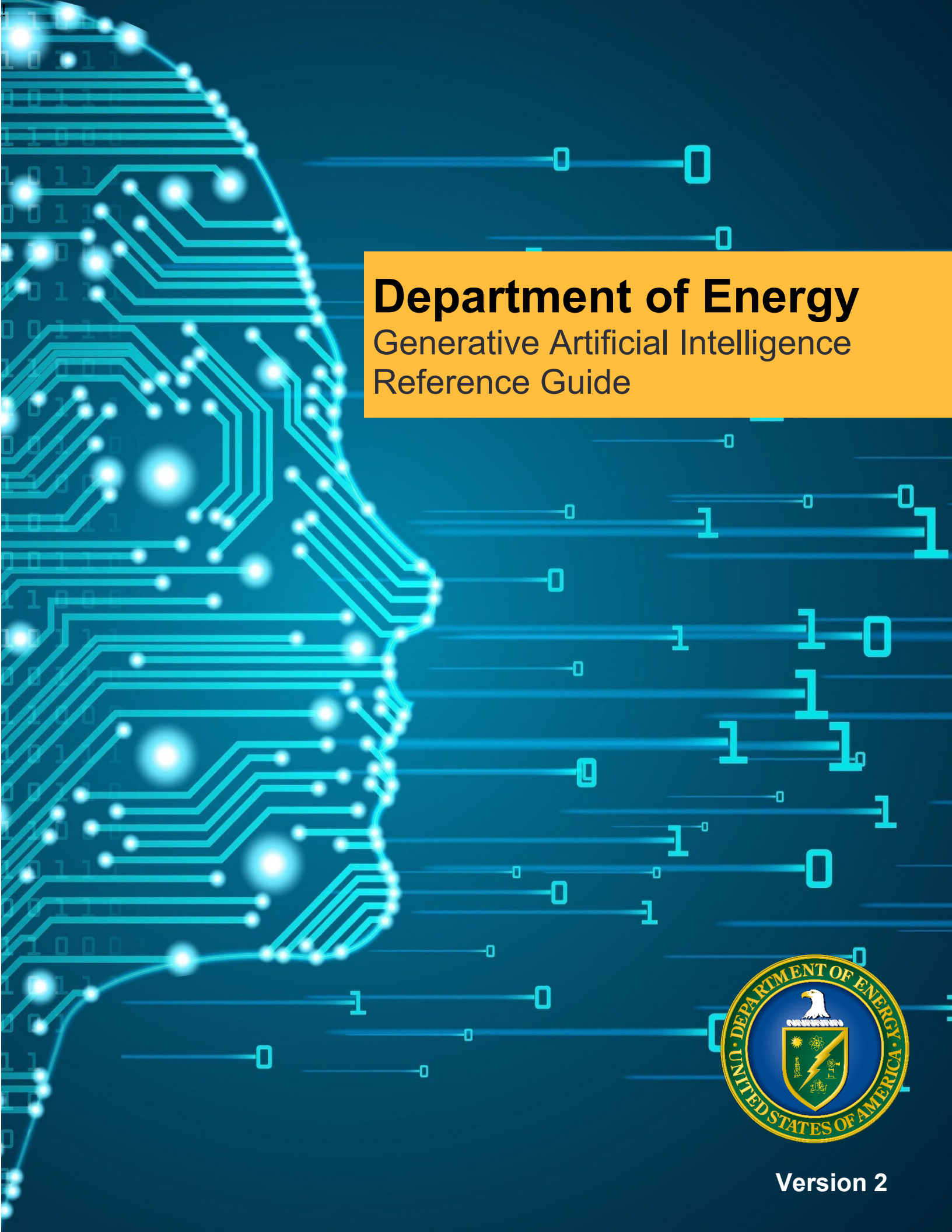


Department of Energy

Generative Artificial Intelligence Reference Guide

Version 2

U.S. DEPARTMENT OF ENERGY

Record of Changes

Version	Date	Author/Owner	Description of Change
V 1.0	September 22, 2023	Office of the Chief Information Officer	DOE Generative AI Reference Guide v1 For Internal Release
V 2.0	April 26, 2024	Office of the Chief Information Officer	DOE Generative AI Reference Guide v2 For Public Release

U.S. DEPARTMENT OF ENERGY

Table of contents

U.S. DEPARTMENT OF ENERGY

1. Document at a Glance

- This is a reference guide for the use of generative AI and shall not be interpreted as a policy. As such, it does not prescribe specific actions.
- This document was developed with a general audience in mind and does not currently include targeted considerations for specialized roles, including Research & Development (R&D) and Management & Operating (M&O) staff. Deeper considerations for these roles may be addressed in a next iteration of this document.
- Generative AI (GenAI) is an incredibly powerful tool that has enormous potential to enable scientific progress, to enhance productivity of the Department of Energy (DOE) workforce, and to drive the DOE mission of innovation with emerging technology.
- GenAI is best used to provide a first draft or to help find options or alternatives rather than being relied upon to produce an accurate and unbiased final output.
- Per Executive Order 14110 on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, federal agencies are discouraged from imposing broad general bans or blocks on agency use of generative AI. The DOE is in the process of considering which GenAI services will be permitted for use based on comprehensive risk assessments. As decisions on services are made, specific guidelines for usage will be established.
- All existing DOE rules concerning data management and use should be followed. Contact the OCIO or the DOE Chief Privacy Officer with questions. Legal questions should be directed to DOE's Assistant General Counsel for Technology Transfer and Intellectual Property or a contractor's cognizant legal counsel.
- Continue to use common sense and follow existing rules regarding data and information management when using GenAI.
- Have a human in the loop to review outputs for accuracy, ethical considerations, quality, and to check for potential bias.
- Reference specific federal guidance (Section 4).
- Refer to the key applications and use cases to understand examples how GenAI might be applied to drive value and innovation at DOE (Section 6).
- Your role in the organization (i.e., general user, data scientist, leadership) is a factor in the key considerations and best practices that are most relevant to you (Section 7.2).
- Keep key considerations and best practices in mind to appropriately manage risks associated with AI and GenAI (Section 8).
- Refer to the Best Practices Checklist to guide your use of GenAI (Section 8.11).
- Any reference to a specific GenAI model or product in this document should not be construed as an endorsement of the model or any of its potential outputs.
- As GenAI continues to evolve, DOE will have to stay agile and adjust to the constantly changing landscape of opportunities, risks, and best practices. This guidance will be updated regularly to reflect the most current thinking.

2. Executive Summary

The Department of Energy (DOE) Generative Artificial Intelligence Reference Guide version 2[1] is being issued as a reference on generative AI (GenAI), a relatively newer AI technology that can produce various types of content, for the entire DOE complex, including federal employees and contractors at laboratories and DOE sites. Key stakeholders and subject matter experts (SMEs) from across the DOE organization established a tiger team to collaborate on the development of this document. The coordinated effort has provided a variety of perspectives from various DOE roles and functions that are woven throughout. Continued collaboration and involvement from a variety of stakeholders will benefit future iterations of this document and drive AI innovation at the DOE. This document is not a policy or directive, but rather a reference guide to help stakeholders from across DOE understand how to responsibly use GenAI. This document and the guidance within will be updated regularly as GenAI technology and the regulatory environment surrounding it continue to evolve. In light of the complexity of GenAI and the pace at which research and commercial advancements are being made, leveraging the expertise of researchers and SMEs will be vital. This guide is not a replacement for legal advice, therefore any legal questions related to the use of GenAI should be directed to cognizant DOE or contractor legal counsel.

> *"People sometimes ask the question, 'Is AI our friend or is AI our enemy?' And my answer to that is, I think AI is our friend, but just like any good relationship, there are boundaries."*[2]
>
> - Gardy Rosius, DOE Deputy CIO

GenAI holds promise for furthering the department's mission, but it also poses risks. In employing GenAI, one must be aware of the capabilities and limitations of the technology and should keep in mind that the user, not the GenAI technology, remains responsible for any actions or outputs resulting from the use of GenAI technologies. Users should therefore not rely on GenAI systems for making decisions; rather, they should use the systems to inform them.

This document comprises helpful information that can be used to spread awareness throughout DOE about the responsible use of GenAI. Topics include background of GenAI, a summary of existing laws and mandates pertaining to GenAI (at the time of publishing), fundamental topics on the responsible use of GenAI (including organizational roles, data, and service models), potential use cases, and the most prominent risks and best practices surrounding this emerging technology.

From data scientists to leadership to general users, everyone across DOE has a role to play in the responsible use of GenAI technology. After reading this document, the reader should have a newfound or heightened awareness of their role in the responsible use of GenAI, as well as foundational knowledge of GenAI solutions, the key considerations and risks that should be accounted for, and the current best practices to mitigate risks and responsibly use GenAI technology.

3. Purpose and Scope

The purpose of this document is to provide an understanding of the key benefits, considerations, risks, and best practices associated with GenAI in the context of the DOE. This document is intended to serve as a valuable reference on GenAI to all groups within the DOE environment, offering an overview of the specific risks, considerations, responsibilities, and recommendations that are associated with various organizational roles.

The scope of this document is that it serves as the second version of a reference guide highlighting GenAI-specific risks and best practices. This document is a true reference guide, not indicative of a policy or directive. The best practices recommended in this document do not supersede any laws, regulations,

or existing DOE policies. As such, this document includes a discussion of background information, key concepts and definitions, and opportunities for applying GenAI, as well as a discussion of key considerations, risks, best practices, and recommendations. This document *does not* include any prescriptive, mandatory actions, as these will be captured in existing and future policies. Furthermore, this document is meant to supplement, but not replace, the existing regulations surrounding GenAI.

4. Federal Guidelines and References

GenAI is evolving rapidly: its underlying technology continues to advance, the variety of GenAI tools available on the market continues to grow, and GenAI is becoming increasingly accessible to the public. As this evolution accelerates, the need grows for awareness of the potential impacts of GenAI, as well as identification and mitigation of the associated risks.

Several federal publications on AI and GenAI have been issued in recent years. These documents are the first point of reference for this document, providing guardrails for how GenAI can be used in the federal Government. Notably, Executive Order 14110 on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* was recently published on October 30, 2023. EO 14110 contains a variety of directives which apply to the Department of Energy, including actions that DOE is required to lead or is required to collaborate with other agencies to deliver. The Director of the Office of Management and Budget issued a Memorandum for the Heads of Executive Departments and Agencies on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence in March of 2024 which should also serve as a key reference document.

This reference guide does not supersede existing law or policy or and is not intended to conflict with any relevant pending legislation. As with all policies, including those not discussed in this document, staff members should review and continue to adhere to DOE policies, procedures, and guides to ensure compliance with laboratory/DOE information requirements. Employees must also continue to follow existing requirements, such as those regarding quality, information security, and research integrity. Staff members must work with appropriate laboratory/DOE SMEs and compliance organizations, such as the Office of General Counsel (GC), the Office of Export Control, the Classification Office, the Office of Environment, Health, Safety, and Security (EHSS), and others as appropriate. The reference guide will be updated as new policies and guidelines are issued. Summaries and selected details of the below references can be found in Appendix E at the end of this document.

Existing relevant federal resources and references include:

1. Office of Management and Budget Memorandum for the Heads of Executive Departments and Agencies, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, March 2024
2. Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 2023
3. Generative Artificial Intelligence and Data Privacy: A Primer, Congressional Research Service (CRS), May 2023
4. Generative Artificial Intelligence and Copyright Law, Congressional Research Service (CRS), May 2023
5. National Artificial Intelligence Advisory Committee (NAIAC) Year 1 Report, May 2023
6. AI Risk Management Framework, National Institute of Standards and Technology (NIST), January 2023
7. Advancing American AI Act, December 2023

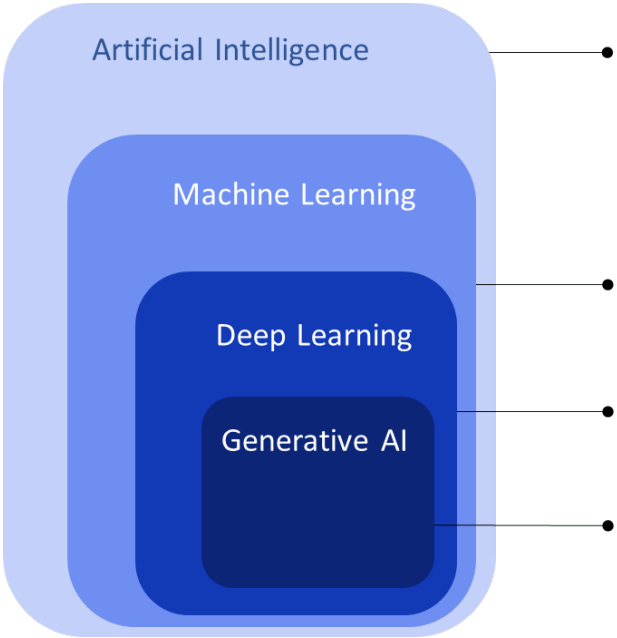
8. AI Training for the Acquisition Workforce Act, October 2022
9. Blueprint for an AI Bill of Rights, Office of Science and Technology Policy (OSTP), October 2022
10. Secure Software Development Framework (SSDF V.1,1), NIST, February 2022
11. AI Accountability Framework for Federal Agencies, GAO, June 2021
12. National AI Initiative Act, January 2021
13. Executive Order 13960 on Promoting the Use of Trustworthy AI in the Federal Government, December 2020
14. AI in Government Act, September 2020
15. Executive Order 13859 on Maintaining American Leadership in AI, February 2019
16. John S. McCain National Defense Authorization Act, Section 1051 for Fiscal Year 2019
17. E-Government Act of 2002

Please refer to Congress.gov to view the status of proposed and pending AI legislation.

5. Background on Generative Artificial Intelligence

5.1 AI, Generative AI, and GPT

Artificial intelligence has advanced tremendously since it was first introduced in the 1950s. Its growth has surmounted two plateaus in advancement which occurred when the vision for the application of AI was broader than the functional ability at the time (i.e., there was not enough computing power or data and no sufficiently advanced algorithms to operationalize the vision). In recent years, AI has gained increasing public attention, becoming a hot topic in technology, as well as in American and international society.



Artificial Intelligence (AI): A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to perceive real and virtual environments, abstract such perceptions into models through analysis in an automated manner, and use model inference to formulate options for information or action.

Machine Learning (ML): A set of techniques that can be used to train AI algorithms to improve performance at a task based on data.

Deep Learning (DL): A subset of machine learning, which is essentially a neural network with three or more layers.

Generative AI (GenAI): The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.

Figure 1: Illustrative definitions of artificial intelligence, machine learning, deep learning, and generative AI. Definition sources: Artificial Intelligence,[3] Machine Learning,[4] Generative AI,[5] and Deep Learning[6]

Like AI, **GenAI** is not new, but has been gaining momentum since the introduction of generative adversarial networks (GANs), a type of machine learning algorithm, in 2014. This development enabled the creation of image generative models. Two additional recent advancements, **transformers** and **large language models (LLMs)** have further accelerated GenAI's evolution and adoption.

Transformers are a deep learning model that adopts the self-attention mechanism, differentially weighting the significance of each part of the input data.[7] In essence, they are a technique that seeks to help AI models determine what to pay attention to.	**Large language models (LLMs)** use self-supervised learning to learn from large amounts of unstructured and unlabeled text data. These models are trained on large bodies of data, allowing for one model to be used for multiple use cases.

The 2017 emergence of the transformer, as well as progress made with convolutions and recurrences for performance and training speed, led to the **generative pre-trained transformer (GPT)** evolving into today's LLMs. GPT, the type of AI that has been at the center of the most visible activity in recent years, is based on neural networks, which are a type of machine learning (ML) model built to mimic the biological neural networks that comprise the brains of humans and animals.

GPT is a family of LLMs built on **deep neural network (DNN)** architecture that have been fine-tuned using **natural language processing (NLP)** and **reinforcement learning from human feedback (RLHF)** techniques, as depicted in Figure 2. ChatGPT is the state-of-the-art consumer-facing AI model built on the GPT. It can answer user-prompted questions, generate stories, summarize text like books or articles, and search text based on conceptual queries. Note that ChatGPT is currently available within DOE for use by request based on mission need. Additional guardrails may be developed and implemented in the future as appropriate.

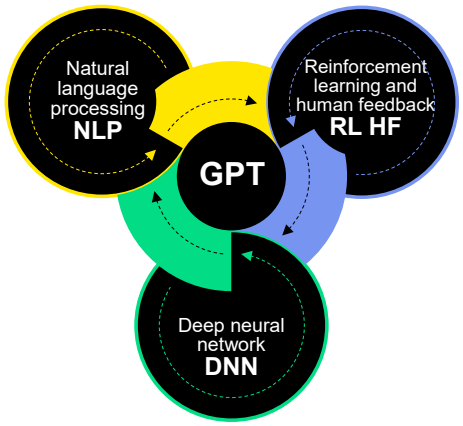


Figure 2: Illustrative depiction of GPT

Foundation models, as termed by Stanford University researchers, are trained on massive amounts of unlabeled data using a transformer algorithm that can be fine-tuned to a wide-ranging array of downstream tasks. To further specialize the models, data scientists can either independently train or fine-tune a foundation model to build **task-specific models,** which are models designed to be effective at specific tasks. Figure 3 shows the high-level relationship between foundation models and task-specific models. Additional AI-related definitions can be found in Appendix K: Glossary.

As depicted in Figure 3, using foundation models as a starting point and including techniques such as supervised fine tuning, instruction tuning, and RLHF, task-specific models that fit the situation at hand are built. The situation at hand may include specifics of the business case, modalities (e.g., text, image/video, speech, auto coding, etc.), solution architecture, use case-specific data, and intended use. Since OpenAI launched ChatGPT

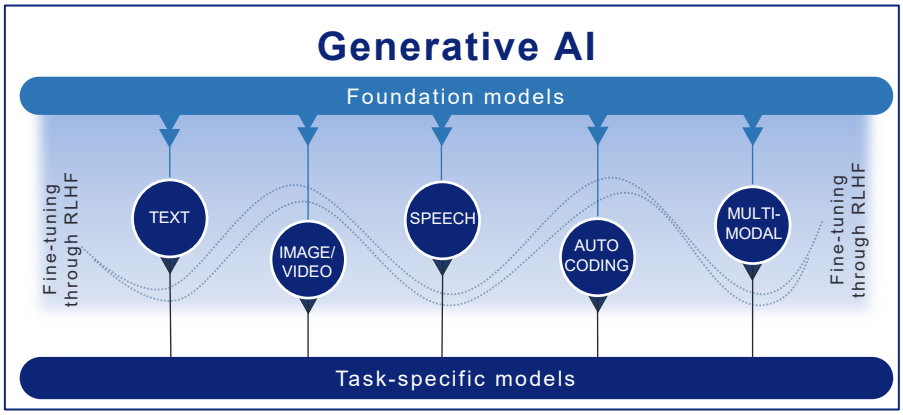


Figure 3: GenAI: foundation vs. task-specific models

in November 2022, new GenAI models built to be task-specific — specializing in different industries, sub-industries, or types of functional applications — have been rapidly entering the market and are generally either generic or built on-premises in a localized environment. User-friendliness and easy access by the general public via the internet have helped make GenAI models increasingly popular.

Most models are unimodal, meaning they focus on a single form of information, like text, speech, or computer code. Multimodal models can learn from multiple forms of input and produce multiple forms of output. Refer to Table 1 below for a list of the various modalities and a sample list of applications and task-specific (or tailored) models currently available in the market. Note that Table 1 does not distinguish between foundation and task-specific models.

Category	Modality	Description	Examples of applications	Examples of task-specific models
Unimodal	Text	Generation of human-like text from text prompts	ChatGPT, Bard, Claude 2, Bing	Jasper, copy.ai, NukeLM[8]
Unimodal	Image/ video	Generation of various images and videos based on text prompts	DALL-E 2, Midjourney, Stable Diffusion (Automatic1111), Stability.ai	Midjourney, Craiyon, Stable LM 2 1.6B
Unimodal	Speech	Generation of synthesized speech from text prompts, speech recognition	Thundercontent, Cleanvoice	Voice synthesis, podcast.ai, Speechmatics
Unimodal	Auto coding	Generation of code (e.g., Python, Java, JavaScript) from text prompts	GitHub Copilot, Amazon CodeWhisperer, Codebots, OpenAI codes, ChatGPT, Bard	GitHub Copilot, Tabnine, Cogram
Multimodal	Multimodal	Generation of various outputs where the model learns from a variety of sources, including text, images, and audio	Gato, Mural by Google, GPT-4, GPT-5	Azure Open AI Service, Google Vertex AI, AWS Solutions, IBM Garage

Table 1: GenAI modalities

5.2 Trends

GenAI and its underlying techniques are rapidly evolving and advancing, and GenAI adoption is exploding at a similar pace. Major breakthroughs have already been made with GenAI technology since its introduction. For example, OpenAI released GPT-4 in March 2023, and by July 2023 a trademark was filed for GPT-5 which suggested a variety of potential new capabilities for the next iteration of the language model. The list includes features that expand ChatGPT further beyond text-to-text GenAI and into the multimodal space, including artificial production of human speech and text, audio-to-text conversion, voice and speech recognition, and development and implementation of artificial neural networks.[9] Note that many of these functionalities, such as speech recognition, predate the emergence of GenAI, but can now be enhanced via GenAI GPT solutions are expected to continue to advance at an aggressive pace. Similarly, between March 2023, when Anthropic's GenAI solution Claude was introduced to the market, and May 2023, large strides were made in the solution's processing speed.

GenAI is forecasted to be a major element in the professional world in the coming years and to realize human-level performance sooner than previously anticipated. Gartner predicts that…

- By 2026, 75% of businesses will use generative AI to create synthetic customer data, up from less than 5% in 2023.
- By 2027, more than 50% of the GenAI models that enterprises use will be domain-specific — specific to either an industry or business function — up from approximately 1% in 2023.
- By 2027, more than half of the selection of development assets from technology marketplaces will be performed by generative AI orchestration.

Figure 4: Source: Gartner®, "Predicts 2024: The Future of Generative AI Technologies," Arun Chandrasekaran, Anthony Mullen, Lizzy Foo Kune, Nicole Greene, Jim Hare, Leinar Ramos, Anushree Verma, February 28, 2024[10]

GenAI is already rapidly transforming areas like marketing and media, while in other areas, it is still in an emerging state. The list of potential use cases (explored further in Section 6: Opportunities to Apply GenAI) continues to grow as GenAI continues progressing in its abilities to generate multiple forms of media, including text, image, video, speech, music, and programming code.

Although GenAI has already gained a huge amount of traction from a multitude of organizations and in myriad aspects of society, it is truly still in its infancy. Expect rapid developments with GenAI capabilities and with the proliferation of its potential use cases and applications to continue. As GenAI continues to evolve, the market and the organizations that adopt it will have to stay agile and adjust to the constantly changing landscape of opportunities, regulations, risks, and best practices.

The rapid evolution of GenAI brings many potential benefits, but also many risks and unknown effects. Although a variety of GenAI-related risks have already emerged, expect some risks to become more pronounced and new risks to appear as GenAI adoption accelerates. Establishing risk management strategies, documenting and sharing best practices, and encouraging awareness throughout the organization of GenAI-associated risks and recommendations will be critical steps in successfully adopting GenAI and realizing the many benefits it can offer.

5.3 Value Proposition

GenAI use cases and potential applications are growing rapidly. Put simply, the value of GenAI is to fill the role of an automated "copilot" for creating materials in various forms of media, including text, image, video, and programming code. Within DOE, this means that GenAI may be able to augment work produced by humans with speed. Once adopted, GenAI may change existing human roles within the organization without necessarily replacing them. Section 6 of Executive Order 14110 includes a variety of mandates on exploring the effects of AI on workers' rights and economic stability. Additional information may become available as that reporting is completed.

When hypothetically used as a copilot for DOE employees, GenAI has the potential to help employees with day-to-day tasks, including (but not limited to) finding information more quickly with its search functionality, generating summaries of meetings and lengthy documents, and drafting emails and other correspondence. These simple examples are areas where GenAI technology is already proficient at managing certain tasks very quickly and at scale. GenAI may be able to produce research or content outlines and starting points for content to allow DOE employees more time to focus on refinement and development of the product. The future workplace will likely include a symbiotic relationship between human employees and GenAI. As AI technologies become integrated with day-to-day working tools (an example may eventually include Office365) and are therefore less visible to the user, this relationship may change or require additional exploration of risk and usage permissions.

GenAI solutions can perform routine tasks much faster than humans (although this introduces risks regarding accuracy, reliability, and "hallucinations," which are discussed in Section 8: Key Considerations and Best Practices). GenAI may create more time and space for DOE employees to add value to their work, empowering them to optimize their time during their workweek.

GenAI is expected to provide capabilities that will allow DOE to innovate more quickly. For example, GenAI can use large sets of relatively unexplored data and content to derive actionable insights that can help drive business value.

There are four primary functions of the GenAI text-to-text capability. Any of these functions can be used on its own, or they can be "bundled" together for a solution. Understanding these four functions can help explain how GenAI might be applied as a copilot in the DOE workplace.

1. **Summarization**: The GenAI summarization capability can take a large amount of text and summarize it into a shorter and more digestible format. While the model might not always exactly deliver on requests for summaries of specific character or word lengths, it can create a close match. The summarization function can also help extract and summarize specific aspects of a larger piece of text — for example, summarizing only the parts of a larger news article that mention a specific organization or topic.
2. **Inference**: The inference functionality generally involves making predictions or solving problems. Examples of the GenAI inference functionality include asking the model to infer the sentiment of a given piece of text (e.g., positive or negative sentiment) or to make an inference on whether there is a specific type of information within text (e.g., the brand of an item in a review of the product, or whether an article contains references to a specific government entity). Note that inference functionalities carry a specific set of risks. Text analysis and inference, specifically if those inferences relate to a specific individual, should be used with considerable caution and only in specific scenarios. All information systems that contain personal information should have a completed Privacy Impact Assessment (PIA) on record. One of the more complex questions asked during a PIA is whether the system will add (create, acquire, or infer) information about the person that was not directly collected and isn't officially part of the record. Inference capabilities may also miscategorize or mischaracterize views or statements made by individuals, and any output should be reviewed by a human co-pilot.
3. **Transformation**: GenAI text-to-text models can transform text in a variety of ways. Translation is one application, as GenAI models are typically familiar with hundreds of languages in varying degrees of proficiency. Text can be translated into multiple languages simultaneously and adjusted based on formality and the intended audience. The model can also transform a piece of text to reflect a new tone or audience, such as turning a casual greeting into a formal business memo. Text can also be edited for grammar and spelling. In addition, text can be transformed into another format, including coding languages, such as changing a block of input from JavaScript Object Notation (JSON) to Hypertext Markup Language (HTML).
4. **Expansion**: The fourth function of text-to-text GenAI is expanding upon a given piece of text or topic, adding to or creating content, or providing additional information on an area of interest.

Examples of expansion include using GenAI to write a response to a constituent query based on the subject and the sentiment of the query or to write a longer-form essay or article about a prompt-given topic. Note that expansion functionalities are most susceptible to risks related to copyright and intellectual property concerns (see Section 8.7) and to AI hallucinations (see Section 8.10).

6. Opportunities to Apply Generative AI

6.1 Key Applications

When considering the opportunities for GenAI use cases and applications, keep in mind that GenAI solutions are multimodal and can generate text, image, audio, video, and programming code. A variety of use cases for each form of media (modality) are already being adopted. Within each of the modalities, there are several viable use cases that can be potentially applied at DOE. Figure 5 provides several of the best fitting GenAI applications for four modalities.

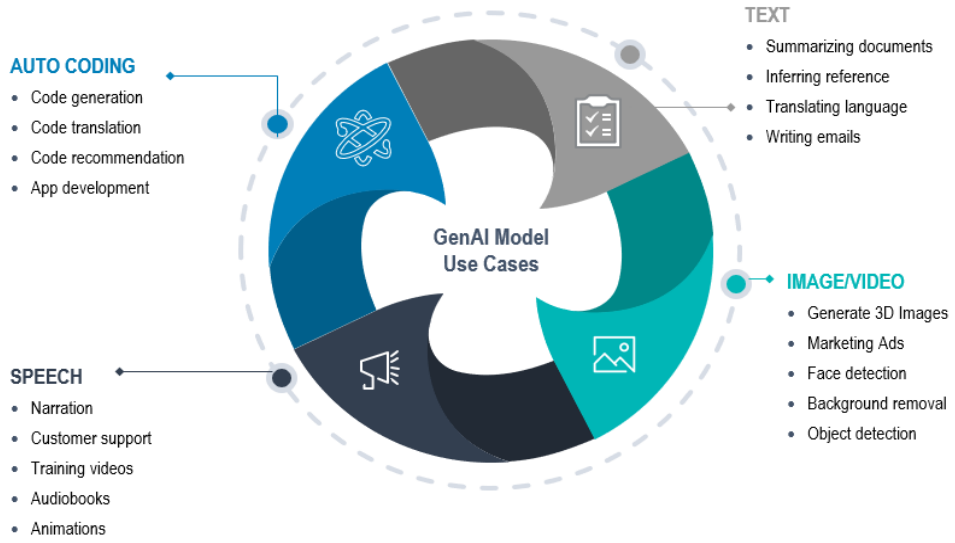


Figure 5: GenAI key applications for text, speech, image/video, and code

6.2 Use Cases for DOE (Illustrative)

The table below expands on ideas for use cases categorized by modality and includes examples of where GenAI could be used at DOE. For an up-to-date and more detailed look into AI use cases being applied at DOE, see the DOE 2023 AI Use Case Inventory (includes use cases for various AI capabilities, mostly in the data analytics and research space, and is not limited to GenAI). Additional use cases for other GenAI applications may become available as the inventory matures. For several science examples, refer to Appendix F. Per Executive Order 14110 on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, the Director of the Office of Management and Budget (OMB) will issue instructions to DOE and other federal agencies for the collection, reporting, and publication of agency AI use cases on an annual basis in alignment with Section 7225(a) of the *Advancing American AI Act*.[11] Regarding the implementation of these and other use cases, Executive Order 14110 Section 10.1(f)(i) states that with appropriate guardrails in place, it is recommended that access be provided to “secure and reliable GenAI capabilities, at least for the use of experimentation and routine tasks which do not have a rights impact."

Generative AI Use Case Examples

Text functionalities (e.g., summarization, inference, expansion, transformation)	
Summarization	Summarize contracts, proposals, reports, stakeholder comments, and regulatory documents
	Build or enhance internal search tools
Inference	Conduct sentiment analysis from an interaction such as an email (e.g., positive or negative sentiment)
Expansion	Create first drafts of contracts, drafts, business presentations, memos, emails, responses to questions, and optimized Request for Proposals (RFPs)
	Provide additional advice or information on a topic
Transformation	Translate documents, contracts, and communications into one or more other languages
	Assist with writing programming code and documentation
	Evaluate and identify errors in code
	Translate code from one programming language into another
	Perform auto-completion of code
Image functionalities (e.g., generation/creation, interpretation)	
Generation/creation (e.g., text-to-image or image-to-image)	Create an image based on a text description
	Create a visual for a product, campaign, cover page, newsletter, logo, promotional material
Interpretation (e.g., image-to-text)	Create a description of a visual used in a presentation, e.g., recognize the image is a depiction of a system and use the visual caption as part of the visual description
Audio functionalities (e.g., speech-to-text, text-to-speech, audio creation/generation)	
Transcription (e.g., speech-to-text)	Transcribe learning resource videos for consumption as text
	Transcribe meeting minutes
Generation/creation (e.g., text-to-speech)	Create an audio voiceover for an educational training
	Generate custom sounds or audio clips
Audio editing (e.g., speech-to-speech)	Edit an audio clip without having to rerecord the clip
	Translate existing speech in an audio or video clip into a different language using an AI-generated voice or the voice of the speaker in the existing audio

Video functionalities (e.g., interpretation (video-to-text), creation/generation)	
Interpretation (e.g., video-to-text, speech-to-text)	Review video used in a proposal or in a meeting where video is included and provide a summary of the video
	Scan videos to identify vulnerabilities and alert security (in the context of security solutions that use cameras)
Creation/generation (e.g., text-to-video and/or image-to-video)	Create videos for training materials or presentations, potentially paired with the use of AI avatars
Abbreviation/ condensation/ translation (e.g., video-to-video)	Create a trailer (a short video) to summarize or abbreviate a longer video
	Use an existing video to generate the same video in other languages
Selected potential GenAI use cases	
Generate interview questions (e.g., text expansion)	
Use case	Create a first draft of interview questions for candidate assessment based on a given job description
Considerations	Evaluate the first draft produced by GenAI to ensure alignment with the intended purpose of the interview
Create meeting minutes (e.g., audio transcription)	
Use case	Generate written meeting minutes for a DOE meeting from an audio recording of the meeting
Considerations	Disclose that the meeting is being recorded to participants to manage legal and ethical risks
Enhance informational videos (e.g., video personalization)	
Use case	Use GenAI to enhance informational videos by adding voice narration, graphics, captions, or translations
Considerations	Personal likenesses may only be used with proper legal consent. However, there are significant ethical and legal risks surrounding the creation and release of deepfakes. Any addition of presenters should be synthetic (not a “likeness” of any one person) unless there has been significant collaboration with the subject and legal experts. Voice narration, translations, and captions should be vetted for correctness and completeness.

7. Operationalization

7.1 Operationalization at a Glance

This section provides foundational knowledge of three key concepts surrounding GenAI before exploring Key Considerations and Best Practices in Section 8. The three concepts introduced in this section are organizational roles, public vs nonpublic data, and service models.

- Different roles throughout the organization have specific responsibilities and considerations when it comes to GenAI.
- As a best practice to mitigate privacy and security risks, users should not input nonpublic (sensitive) data into a GenAI system unless the appropriate processes have been undertaken to ensure that the rights and potential uses of the data are permitted, or they are using a tool which is appropriately configured and approved for their use case. This best practice is critical for public or commercial systems where the model, inputs, and outputs are not under DOE's direct control.
- There are several ways to approach service models. The key is to determine whether or not DOE controls the GenAI model and the outputs and whether inputs are added to the model's training data. Specific considerations apply to either case.

7.2 Organizational Roles

Everyone has an important role to play when considering and implementing a new GenAI solution or using an existing GenAI tool. Whether a general user or an AI systems specialist, every employee should be cognizant of their role and any specific considerations that may apply to their role related to the development and use of GenAI technologies. Below is an introductory set of roles across the organization with corresponding descriptions. Note that this list is not exhaustive, and that in many cases these roles use language which is specific to DOE but may have applications in other organizations. Consider developing a RACI (Responsible, Accountable, Consulted, Informed) Matrix to clearly define the roles and responsibilities for each specific GenAI solution. The descriptions listed below are responsibilities to be considered when drafting more explicit requirements, not requirements in themselves. For additional information on the AI Lifecycle referenced in this table, see Appendix G.

Many of these roles are still developing within the DOE. For example, the roles and responsibilities of the Chief Artificial Intelligence Officer (CAIO) and the Responsible Artificial Intelligence Officer (RAIO) vary across organizations. In some organizations, one person may take on both the CAIO and RAIO roles, while in DOE, these are currently two distinct and emerging roles (at the time of publication of this document). At DOE, the RAIO reports to the Chief Intelligence Officer (CIO), while the CAIO reports to the Secretary of Energy.

Organizational role name	Description
General user	Regardless of whether a person may hold one of the specific roles listed below, almost anyone in DOE may be or soon may become a general user of GenAI. For general use, it is key to understand the nature of the information input into the model, the intended purpose of the model, and any restrictions related to the input data or to a user's role in the organization. Where needed, report observed issues of accuracy, fairness, or bias in a model's output. The general user may also include the human in the loop to verify outputs to ensure both responsibility (ethics and limited bias) and accuracy, especially when a GenAI system generates outputs for humans to consume, takes action prompted by outputs, or draws conclusions based on outputs.

Organizational role name	Description
AI developer	The AI developer is charged with designing, coding, and iteratively improving new GenAI applications in collaboration with other roles such as data scientists, user experience designers, cybersecurity specialists, project sponsors, and leadership. The AI developer creates the systems and AI solutions, as opposed to the data scientist who develops the underlying models. AI developers should consider the unique implications surrounding GenAI technologies and consult with AI SMEs as needed to implement AI-specific best practices.
AI policy and governance staff	AI policy and governance staff advise on the creation of new AI policy and governance based on organizational, technical, and legislative need for Department-wide adoption and implementation of responsible, ethical, and trustworthy AI frameworks, principles, procedures, and practices. They ensure that policies reflect the best practices in AI and address any security, risk, or privacy concerns as well as responsible and ethical AI principles. This role is emerging and evolving very quickly, and there will likely be additional responsibilities associated with it as adoption of GenAI progresses.
AI portfolio manager	The AI portfolio manager oversees all AI capabilities and projects currently in the pipeline for their organization or departmental element. This role is critical to limiting redundancy of AI solutions which may have similar functions. The portfolio manager should understand the current and near-future AI landscape to identify trends and risks in proposed capabilities. This role is accountable for the entire pipeline of all AI initiatives at the highest level and is responsible for all of the policies and processes associated with the AI pipeline. This role also manages funding and budget for all AI initiatives.
AI subject matter expert (SME)	The AI subject matter expert (SME) advises others on best practices and risk considerations for GenAI technologies. This role needs to understand the technologies involved for a given use case and provide advisory services to other roles including data scientists, leadership, and data engineers to share knowledge with the appropriate team members. This role may be involved throughout any stage of the AI lifecycle. For example, the SME may be involved in the initial planning stage to ensure that a given business problem is a good fit for a GenAI solution, during the development or implementation stages to ensure technical or process efficiency and quality, or by providing insights on how to best educate users on responsible GenAI usage.
Business analyst	The business analyst is tasked with coordinating efforts across project or development teams to design, launch and operate GenAI capabilities. This role is an internal user who is responsible for the translation and coordination of needs and tasks between the business users and the AI/ML development team. This role translates business needs into technical requirements and helps business users effectively use the output as the system was designed. The business analyst may also identify potential business use cases for AI technologies throughout the course of their daily responsibilities.
Chief Artificial Intelligence Officer (CAIO)	The Chief Artificial Intelligence Officer (CAIO) is a role defined in Executive Order 14110 Section 10.1(b)(i) charges the Chief AI Officer with “coordinating their agency's use of AI, promoting AI innovation in their agency, managing risks from their agency's use of AI, and carrying out the responsibilities described in Section 8(c) of Executive Order 13960 (“Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government”) and in Section 4(b) of Executive Order 14091.” They are also charged with overseeing AI Governance Boards as required for their agency, overseeing risk management activities for government uses of AI, and making recommendations to

Organizational role name	Description
	agencies to reduce barriers to the responsible use of AI, including AI-specific barriers of adoption for information technology infrastructure, data, workforce, budgetary restrictions, and cybersecurity processes. Given that this is an emerging roll, the governance process and responsibilities for this role are still in development at the DOE.
Contracting officer	The contracting officer facilitates the purchase of GenAI tools, platforms, technology, and services for DOE and uses their knowledge to ensure that the best fitting GenAI tools are selected. This role ensures that new tools acquired for DOE are rigorously tested, meet policy and security requirements, are not duplicative of other ongoing efforts at the organization, and are aligned with any existing policies, procedures, and strategies for the organization. Contracting professionals are also concerned with the details of new and existing contractual agreements and terms of service (ToS) agreements with third-party providers of GenAI solutions and services to ensure that risk is appropriately shared between DOE and service providers. This role also collaborates with legal specialists/legal counsel and organizational stakeholders.
Cybersecurity specialist	The cybersecurity specialist is charged with the security, safety, and resiliency of organizational (or related) systems. This role should be involved from the very beginning of any GenAI initiative to ensure that the solution design has sufficient protection measures in place and will not interfere with existing security measures as a result of the project's requirements. This role ensures that any solution which is planned and designed has a high probability of being successfully operationalized. At each phase of the AI lifecycle, this role needs to ensure security features are properly maintained. Cybersecurity professionals may also focus how AI can be used to bolster cybersecurity organization-wide and to prevent or respond to adversarial attacks in non-AI systems.
Data engineer	The data engineer ensures that the appropriate data is available to data scientists and that the data is as trustworthy, fair, and free of bias as possible. This role should have an understanding of data structure, environment, management pipeline, and data quality in terms of sourcing, depth, and breadth in data that is used for building GenAI systems. Data engineers who are involved in the training of newly developed or purchased models may also assist in data tagging and other implementations to meet organizational data governance policies, and they have the responsibility to define and implement the right plan and architecture for the data. Data engineers implement the appropriate checks into the data management pipeline to ensure that the quality standards (typically defined in a policy) are met at each phase of development.
Data scientist	The data scientist uses data from various sources to assist in making organizational decisions and reaching sector-related conclusions. This role is usually responsible for developing models and must understand in detail the intended purpose and outputs of their models to ensure proper functionality, any regulatory or privacy considerations related to a given model or project, and the differences between training/validation and production/live data. When developing or training a GenAI system, data scientists are responsible for ensuring the quality, representativeness, and lack of bias in outputs from the training data set. The data scientist should explore data provided by data engineers and apply the best methodologies and tools to reach the objective of the project with the data given. Data scientists may take on other roles from

Organizational role name	Description
	within this list, including solution architect, developer, AI SME, and more. They may also take on the role of prompt engineer in the case of GenAI.
Development operations (DevOps) engineer	The DevOps engineer is responsible for the processes that help DOE improve the efficiency of developing, testing, operationalizing, and updating technology. This role helps to facilitate these processes with knowledge of emerging technology, project management skills, and team communication. The DevOps engineer collaborates with other technical roles to ensure functionality during the transfer of the solution from the pilot to the production environment while monitoring for potential risks and vulnerabilities (e.g., data drifts, software or model drifts, and security issues). The DevOps engineer is also responsible for ongoing model/system management depending on organizational policy or guiding principles. This role may be part of a larger team which includes a variety of skillsets within the larger umbrella of DevOps, e.g., build architect, release manager, infrastructure engineer, automation architect, among others.
Executive sponsor	The executive sponsor is a leadership position and is responsible for communicating to the organization and raising awareness about the importance of prioritized strategic initiatives. This role makes sure that the resources are available for any prioritized GenAI project and elevates the initiative at hand to a higher priority level. This role is responsible for securing initial funding for the project at hand and for ensuring the appropriate stakeholders are involved with the initiative and aligned on its goals. The executive sponsor may occasionally fill the role of a project director, which is more hands on in nature.
Information technology (IT)/ systems professional	The information technology professional is charged with overseeing IT systems across the organization. The focus for IT professionals is on both implementing and maintaining new GenAI technology and providing prospective users with recommendations and best practices for GenAI use. This role has different responsibilities depending on the stage of the AI lifecycle. For example, during the development phase, the IT professional helps build the sandbox.
Leadership	There are many different potential responsibilities that may pertain to the leadership role. Leadership sets the strategic direction, priorities, goals, and mission objectives for the organization in collaboration with federal officials. Leadership is concerned with understanding GenAI at the executive level, understanding the broad regulatory landscape as it pertains to GenAI, and building awareness and understanding of GenAI and its use cases across the organization. Leadership at the Department level and at the Departmental Element (DE)/site level ensures that training on GenAI usage, limitations, and risks are both available and encouraged for all potential GenAI users across DOE. Leadership also facilitates the creation of a central, collaborative mechanism for sharing knowledge, collaborating on initiatives, reporting observations of bias, unreliability, security issues, and other concerns in GenAI platforms as a means for organizational learning. This group may also include leaders who don't interact directly with AI technologies.
Legal specialist in AI and emerging technology	Legal professionals specializing in emerging technology and AI navigate the intricate realm of AI in law, focusing on the implications and requirements unique to emerging technology and AI. They are tasked with comprehending the nuances of AI models, utilizing them effectively for legal tasks, and aligning these applications with established ethical standards and legal frameworks. They ensure that the use of AI within legal operations not only optimizes efficiency and accuracy, but also maintains robust privacy protocols,

Organizational role name	Description
	addresses potential risks, and adheres to legal regulations and standards. Legal specialists are responsible for staying up to date on existing and pending legislation to proactively protect the agency and prepare for what is to come. They also serve as advisors for the entire GenAI team.
Management and operating (M&O) staff	The M&O contract staff oversees National Labs and needs to understand the legal, technical, and procurement risks undertaken as an organizational third party while monitoring the performance of GenAI solutions in operation. This role focuses on the maintenance, monitoring, and overall usage of GenAI system once deployed. In the design and development stages, this role ensures that the requirements are realistic and accounted for in the solution build. Note that M&O staff includes a wide range of people, including researchers, executives, cybersecurity specialists, and more. Many of the roles on this list may apply to select M&O staff.
Product manager	The product manager is responsible for understanding and prioritizing organization-relevant market opportunities for AI use cases, in collaboration with use case owners and leadership. This role is responsible for managing the activities and the team for the specific GenAI initiative, product, or service at hand. There may be a range of product managers as the AI portfolio expands.
Program manager	The program manager is a unique role associated with research institutions. This role writes funding opportunity announcements, reviews proposals, makes funding recommendations, and manages awards, among other responsibilities. For the Office of Science, these awards are primarily to universities and National Laboratories and focus on fundamental research. While proposals and awards may include the development and use of AI for science and engineering, there is also an opportunity for GenAI to assist the program manager in the performance of their duties.
Project director	The project director is charged with committing time and resources to oversee and review the development of new GenAI capabilities in a project-based environment, which includes making high-level decisions on the direction and goals of the project, how the project is managed and structured, and how GenAI technologies may contribute to other ongoing initiatives. Project directors may have overlap with the leadership role. Project directors provide direct requirements depending on the use case and ensure that all roles in the GenAI team are invited to collaborate and that initiatives are completed in alignment with the overall strategy.
Research scientist	The research scientist investigates open-source and published research and/or conducts studies and experiments. Research scientists are especially concerned with the technical accuracy of the output when utilizing a GenAI tool in their work, as well as understanding the copyright and publication considerations for any research done using a GenAI tool. Responsible and ethical use of GenAI tools is a key concern for research scientists. Note that the roles of many DOE research scientists differ in their advancement of science in using AI. Refer to Appendix F for references to additional resources on DOE R&D and advancement of science. At this time, researchers reviewing grants as part of the NSF merit review process are unable to use AI as an aid in any capacity.[12]
Responsible AI Officer (RAIO)	The RAIO is responsible for managing an AI risk management program, collaborating with the appropriate officials to establish or update processes to evaluate the performance of AI systems, overseeing DOE compliance with requirements to manage AI risks, and conducting risk assessments when

Organizational role name	Description
	needed. The RAIO is also responsible for coordinating implementation of the nine Trustworthy AI Principles set forth in Section 3 of EO 13960. Given that this is an emerging role, the governance process and responsibilities for this role, as well as their implementation, are still in development at the DOE.
Solution architect	The solution architect oversees the integration of GenAI technologies into the overall organizational IT infrastructure. This may require implementation of additional data governance or security measures, as well as an awareness of the types of data flows and access which are or are not permitted with GenAI technologies. This role helps design how the system will look, and function based on the given requirements and intended purpose. This role thinks about how the GenAI model will be integrated and operationalized with upstream and downstream systems and ensures that everything goes smoothly when the successful POC moves to production.
Use case business owner	The use case business owner is responsible for establishing a business case for a new or expanded AI solution, communicating the needs of the business to the development and/or procurement team, collaborating with other functions to select a solution for the use case, and assisting in the implementation and maintenance of the GenAI solution as needed. This role is involved on a daily basis with the given use case and may be responsible for providing the data for the development of the solution.
User experience (UX) designer	The user experience (UX) designer is responsible for creating the human-centered user interface of a technological solution or product, including the design of the components and features that control how a user interacts with the product. UX designers focus on how to tailor the user interface to meet the needs of the target end-users of the solution, to improve the quality of the customer experience, and to make the user interface as simple and effective to use as possible for the target end user of the product. It is essential for this role to embed a human-centered design approach in the user interface in the final product.

7.3 Public vs Protected Data

With all forms of AI, and especially in the context of GenAI, there are critical considerations surrounding the use of data and information. Data (meaning recorded information, regardless of form or the media on which it may be recorded, including both technical data and computer software) can be broadly categorized into two classes: public and protected.

- Protected data includes information that is protected from public distribution and/or certain uses, and includes sensitive data, private data, proprietary data, confidential data, and copyrighted works. Confidential data is further explored in Section 8.6 and should not be conflated with classified information which has also been marked as confidential. For definitions of various types of sensitive data, refer to Appendix J: Examples of Protected Data. Some types of protected data are only protected and/or nonpublic for a defined period of time while other types of protected data might remain protected and nonpublic indefinitely.
- Public data is information that can be freely used and distributed by anyone with no legal restrictions regarding access or usage.

The key takeaway is that any type of ***protected data or nonpublic information should not be shared or input into any public or commercial (non-DOE-controlled) GenAI system***. DOE controls the GenAI system if it is a closed, or proprietary, AI system. A closed AI system is developed and controlled

by a single organization that has full control and ownership over the system. Exceptions may apply if the GenAI system is protected by a confidentiality agreement between DOE and the vendor and DOE has sufficient rights to use the protected or nonpublic data.[13] This recommendation is a critical best practice for DOE use of GenAI. Even for DOE controlled GenAI systems, users should ensure they have the appropriate rights to use any nonpublic or protected data being utilized.

The key takeaway for public data is that the possibility of the publicly available information input into the GenAI system becoming incorporated into the AI model may not pose a threat since the data is already public. However, it is important to remember the risk of plagiarism and copyright infringement and to verify publicly available data even though it can be used freely.

There are many existing laws, mandates, and internal DOE policies and trainings that provide valuable guardrails on the general use of data, especially policy on the protection and sharing of information. These should be given special attention when using GenAI.

Additional resources regarding data include:

- DOE Resources: CUI Slicksheet, Controlled Unclassified Information
- "Protected Critical Infrastructure Information (PCII) Information Program," Cybersecurity and Infrastructure Security Agency (CISA)
- Federal Acquisition Regulation (FAR), Acquisition.gov

Improper and unauthorized use or disclosure of protected data or nonpublic information can lead to legal liability for both DOE and the individual responsible for the unauthorized use of disclosure. In some cases, the liability for individuals can include potential civil and criminal penalties.

7.4 Service Models

Using GenAI solutions can come with a variety of risks depending on who has control of the solution in use. The major distinction comes down to the following question: Does DOE **control** or **not control** the solution? (Refer to Section 7.3 directly above for a brief discussion on the distinction between a DOE-controlled system vs. non-DOE-controlled system.) Systems that are off-the-shelf and have not been purchased by and customized for DOE or publicly accessible (e.g., ChatGPT, Bard) will come with a higher set of risks versus an internally built application that is protected within the boundary of DOE. It is important to understand the type of application in use and to comply with existing DOE policies and guidelines when it comes to the use of technology. To clarify the instance of an application that is to be used, or for any other systems-related questions, contact the Responsible AI Official (RAIO), the Chief AI Officer (CAIO), the OCIO Supply Chain Risk Management (SCRM) Team, or the local Information Technology (IT) team. Robust understanding and documentation of the information detailed below is critical for implementing the best practices listed in Section 8.

The following are questions to consider before using a GenAI application:

- What are the specific benefits the use of this solution would provide?
- Who built the application and/or model?
- Is this solution developed and/or controlled by DOE? If not, is there a tool which is?
- Where is the application hosted?
- How was the model trained?
- How was the data selected and collected?
- What data was used to train the model and as of what date?

- Is the use permissible under the data rights available given the source of the data or for type of information?
- Which model, platform, and methodology were used?
- Is this a public application, is it in a private secure cloud, or is it operating inside a DOE secure network?
- Is the application an off-the-shelf product?
- When information is supplied to the application, what is the risk of the information becoming public?
- How was the model validated?
- Are the terms of service for the GenAI system "federally compatible" within the guidelines established by the General Services Administration (GSA)? Beyond that, do the terms of service for data-handling meet appropriate privacy or confidentiality requirements?
- What are the limits of shared or transferred risk and accountability between DOE and all involved parties?

For systems built internally for DOE that routinely process DOE business information, on-premises or hybrid solutions, the following additional considerations are needed before operationalizing GenAI:

- Has the application been approved by DOE cybersecurity standards and procedures (e.g., Authorization to Operate (ATO) has been issued, Privacy Impact Assessment (PIA) in place, etc.)?
- Is the application FedRAMP-approved?
- Is the model fully self-contained with no third-party retrieval?
- Is there a service contract or agreement in place?

8. Key Considerations and Best Practices

8.1 Key Considerations and Best Practices at a Glance

This section provides an overview of seven considerations where certain risks are known to arise with GenAI technologies. Each of the seven subsections provides a brief description of the consideration, public and illustrative examples of where and how risks may arise, specific risks, and best practices to mitigate those risks. The seven topics covered in this section are security and resilience, privacy, confidentiality, intellectual property, safety, fairness and bias, and AI hallucinations and interpretations. At the end of this discussion in Section 8.11: Best Practices Checklist, a summary of best practices is provided. Note that challenges may arise when implementing best practices, and keep in mind that best practices for GenAI will continue to emerge and evolve. Reference the DOE AI Risk Management Playbook (AIRMP) for more ideas regarding AI-related risks and risk mitigation strategies.

8.2 Introduction

There are always specific considerations, unique risks, and best practices that should be given attention when embarking on a journey to innovate with technology. AI has more unique considerations than non-AI technologies due to the complex nature of the models and their reliance on datasets. GenAI is even more complex and therefore comes with even more nuanced considerations, risks, and risk mitigation strategies. The entire organization needs to understand the complex nature of GenAI risks and best practices to maximize the benefits of GenAI while minimizing its risks.

Each role listed in Section 7.2: Organizational Roles, including general users, has GenAI-specific considerations and best practices associated with it. It is essential to build awareness throughout the organization of these roles, responsibilities, and best practices.

This section provides details on seven key considerations and best practices pertaining to GenAI. Section 8.3: AI Risk Management introduces the seven characteristics of trustworthy AI systems outlined in the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (NIST AI RMF 1.0). These seven characteristics are used as a framework to discuss seven key considerations surrounding GenAI, each of which has its own subsection (Sections 8.4 – 8.10).

8.3 AI Risk Management

When developing and deploying new GenAI technologies or when incorporating GenAI functionalities into existing systems, it is critical to understand both existing risk management considerations and unique risks associated with GenAI. GenAI introduces an additional layer of risk considerations, including hallucinations, misinterpretations, training poisoning, prompt injection, deepfakes, and intellectual property infringement. Keep in mind that risks *not* specific to GenAI may become more pronounced when GenAI is integrated into the technology ecosystem. It is best to design GenAI systems to be secure, responsible, and trustworthy at the onset of any GenAI initiative, and effective AI risk management is a critical component of achieving these goals. When employed appropriately, AI risk management also allows users and developers to understand the limitations and ambiguities of AI and to enable the selection of appropriate, responsible, and viable AI use cases.

AI risk management differs in several ways from non-AI technology risk management practices. Risk management program governance typically includes a set of metrics to measure performance and progress based on public and historical data. However, AI use cases do not generally have reliable metrics to use in comparison, and metrics may not fully capture relevant factors or impacts. There is also a lack of consensus on how to define clear metrics for reliability or trustworthiness in AI systems.[14] Another difference is that AI systems designed to augment human actions (which have existing risk management criteria) act differently from the human thought process, which can make specific AI risk

management requirements difficult to operationalize. Finally, prioritization of AI risk resources may be decided differently than with non-AI risk management strategies. Prioritization metrics for AI systems may include those that interact with humans, that have downstream effects on safety, or that have training sets that include personally identifiable information (PII).[15]

The NIST Artificial Intelligence Risk Management Framework (NIST AI RMF 1.0) is an excellent resource to become familiar with AI-related risk management and responsible AI practices. The NIST AI RMF is cited heavily throughout the latest Executive Order 14110, which requires the Secretary of Energy to collaborate with the Secretary of Commerce, Secretary of Homeland Security, and others to develop guidelines and best practices for developing and deploying safe, secure, and trustworthy AI systems, including by developing a companion resource to the NIST AI RMF for GenAI. The risk management practices outlined in the NIST AI RMF are considered the current standard by the federal government.

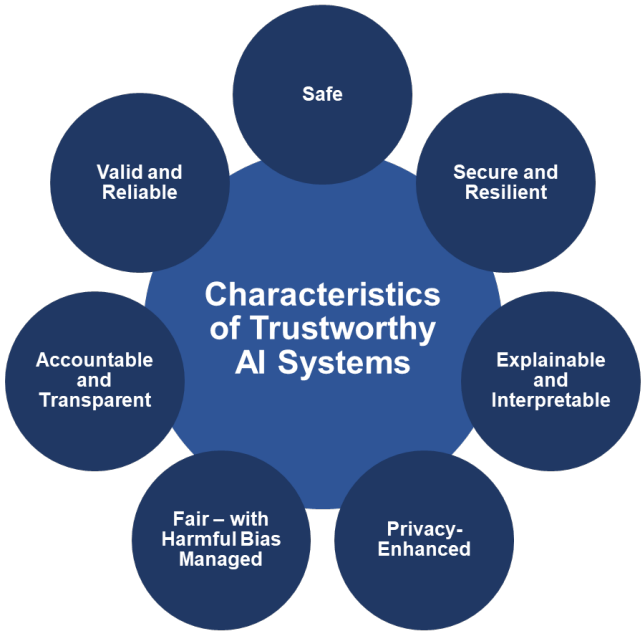


Figure 6: Seven characteristics of trustworthy AI systems outlined in the NIST AI RMF 1.0

The NIST AI RMF outlines three core concepts to emphasize in responsible AI development: human centricity, social responsibility, and sustainability.[16] With these core concepts of responsible AI in mind, AI risk management can enable responsible usage, practices, and processes by encouraging employees across the DOE ecosystem to practice critical thinking about potential risks and unexpected impacts of AI.

A critical overarching theme in designing, developing, and deploying AI in a way that maximizes its benefits while adequately managing its risks is trustworthiness. Trustworthy AI is a concept reflected in numerous relevant federal publications, including Executive Order 13960 on *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* and Executive Order 14110 on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*.

The NIST AI RMF lists seven characteristics of trustworthy AI systems to guide AI risk management and responsible AI development. These seven trustworthy AI characteristics defined by NIST are: Safe, Secure and Resilient, Explainable and Interpretable, Privacy-Enhanced, Fair – with Harmful Bias Managed, Valid and Reliable, and Accountable and Transparent (refer to Figure 6). For additional information on the NIST AI RMF, refer to Appendix H or the full online publication. The seven characteristics of trustworthy AI as outlined by the NIST AI RMF are used as a framework for the discussion of key considerations that follows in Sections 8.4 – 8.10.

In the following subsections, key considerations and best practices for GenAI are presented in seven high-risk areas: Security and Resiliency, Safety, Privacy, Confidentiality, Intellectual Property, Fairness and Bias, and Hallucinations and Misinterpretations and are mapped against the NIST AI RMF's seven trustworthy AI characteristics listed above. All seven DOE considerations have trustworthy implications and aspects. When the best practices corresponding to the seven DOE key considerations are thoughtfully applied, GenAI systems will be trustworthy by design. The seven DOE key considerations are mapped against the seven NIST AI RMF trustworthy AI characteristics as follows:

DOE key consideration	NIST AI RMF 1.0 trustworthy AI characteristic(s)
Security and Resiliency	Secure and Resilient
Safety	Safe
Privacy	Privacy-Enhanced
Confidentiality	Secure and Resilient; Safe
Intellectual Property	Secure and Resilient; Accountable and Transparent
Fairness and Bias	Fair – with Harmful Bias Managed
Hallucinations and Misinterpretations	Accountable and Transparent; Valid and Reliable; Explainable and Interpretable

The subsequent sections go into depth on the aforementioned seven DOE key considerations for GenAI. Section 8.11 provides a checklist to summarize the highest priority best practices for all seven topics, as well as more general best practices not listed in the seven subsections.

8.4 Security and Resiliency

While GenAI solutions have emerged as innovative tools to drive science, operation, and business transformation, they also introduce security risks that should be carefully addressed and mitigated.

Definition

AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use, including covert modification of training data or the foundational models, may be said to be **secure**.

AI systems are said to be **resilient** if they can withstand unexpected adverse events or unexpected changes in their environment or use — or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when necessary.[17]

Examples

Illustrative example

Applications backended by GenAI services that are then installed on devices can automatically join meetings or access other data and services. These apps make phishing/smishing attempts more realistic and convincing with deepfake imagery and voice. These should all be covered in existing cybersecurity practices and regulation but are emergent and have an increased threat landscape.

Public example

In spring 2023, a vulnerability in ChatGPT's source code exposed users' sensitive information and allowed adversarial players to view users' chat history. Some of the data that was exposed included names, email addresses, credit card types, payment addresses, and chat histories. The potential aftermath of this incident includes the exposure of private data (pertaining to both individuals and businesses), damaged reputation, and legal repercussions.

Read more here: Generative AI's first data breach: OpenAI takes action, bug patched | Markets and Markets

Key considerations

- Personally identifiable information (PII), as well as sensitive, confidential, proprietary, or otherwise protected information stored by a GenAI system which was entered as part of a prompt or gathered as part of the model training process could be accessed by an attacker or other adversarial players.
- Adversarial players can use "prompt injection," a method used by hackers that tricks the system into bypassing specific security or ethical guardrails that have been patched onto foundational models, to manipulate GenAI systems to output unauthorized information.
- GenAI tools such as ChatGPT can be tricked into generating malware or ransomware programming code.
- Adversarial players can poison data to create vulnerabilities in the system.
- Deepfakes, or digitally forced images or videos, can be created using GenAI.

Best practices

- Monitor and test the GenAI system for vulnerabilities, threats, failures, etc. and work to develop methods of test and securing AI systems more efficiently.
- Develop and implement detection features that can identify threats, failures, and attacks on the system and notify personnel.[18]

- Train users to understand the security risks associated with AI, including the potential for malicious use or adversarial attacks, as well as risks to input and output validation and data integrity.
- Develop and regularly update robust, secure systems for sites to defend against threats and plan for system resilience to ensure that AI systems can recover from potential attacks or failures.
- For GenAI systems that DOE has developed or for which DOE has compiled a specialized training set, establish a program to conduct adversarial testing via "red teaming," which involves actively seeking out examples of where the GenAI system fails, retraining the model on these examples, and continuing this iterative process until the team closes the loop on identifying failures.[19]
- Regularly update provisions to the system's risk management plan to reflect the newest risks.
- Isolate GenAI systems as much as is practical and avoid allowing GenAI systems to directly control other systems (especially real-world physical systems).

Additional resources

- NIST AI Risk Management Framework
- DOE Artificial Intelligence Risk Management Playbook

8.5 Privacy

Protecting privacy is fundamental to preserve the public's trust in the government. The federal government strives to hold the highest standards in collecting, maintaining, using, and disseminating people's personal information.

Definition

The **privacy** consideration involves protecting the security of personal information to ensure its accuracy, relevance, timeliness, and completeness, avoiding unauthorized disclosure, and ensuring that no system of records concerning individuals, no matter how insignificant or specialized, is maintained without public notice.[20]

Examples

Illustrative example

During a global pandemic with human contact and exposure concerns, an organization seeks to understand where resources are physically present for purposes of cleaning and disinfection. An AI model is used to consume and train on input data. As the data is fed into the model, the output can provide information in a way that unintendedly exposes individuals' geolocation. This situation now requires additional legal notifications to individuals regarding privacy.

Public example

In 2021, Canadian privacy officials found that American-based company Clearview AI was collecting photos of Canadian citizens, including children, without their knowledge or consent for use in a facial recognition software which was used by law enforcement agencies to identify persons of interest or victims. The AI utilized billions of photos found on the internet and social media accounts to attempt to identify the person. Even after use of the technology was halted, Clearview continued to use pictures of Canadian citizens. Experts warned that not only was the storage of data against privacy laws, but that the tool itself could easily be misused. This issue also highlights the concept of consent in collecting training data.

Read more here: U.S. technology company Clearview AI violated Canadian privacy law: report | CBC News

Key considerations

- AI platforms and service providers may share user information with third parties, including vendors, service providers, affiliates, or other users, without informing the user.
- Information entered into a GenAI system may become part of its training data set. Thus, any proprietary, sensitive, personally identifiable, confidential, or otherwise protected data entered as part of a prompt could be used in outputs for other users of the system.

Best practices

- Integrate privacy programmatic considerations into a wide range of functions, including but not limited to information security, records management, strategic planning, budget and acquisition, contractors and third parties, workforce, training, incident response, and risk management.
- Continue to adhere to existing privacy policies and procedures. Iteratively review new privacy policies and recommended procedures to ensure they are accounted for in GenAI use cases.
- Exercise data minimization practices by taking steps to anonymize data and limit the collection, storage, and reuse of personal information.

- Do not include protected data or nonpublic information (as part of the input to any commercial or open GenAI system.
- Describe clearly and accurately, and share in an accessible way, how the department uses, manages, and collects information. Clearly document who creates, contributes to, and has access to that information, and communicate this to all people who entrust government with their data and information.[21]
- Implement full lifecycle stewardship of data, which is the practice of securing and protecting data, metadata, and information throughout its lifecycle. That includes collection, storage, use, control, processing, publication, transfer, retention, and disposition.[22]
- . Develop tailored GenAI privacy trainings for employees with access to protected data (e.g., sensitive data, private data, confidential data, limited rights data, proprietary data etc.) or any other nonpublic information.[23]
- Conduct a privacy impact assessment (an analysis of how information is handled to ensure handling conforms to applicable legal, regulatory, and policy requirements regarding privacy, to determine the risks and effects of creating, collecting, using, processing, storing, maintaining, disseminating, disclosing, and disposing of information in identifiable form in an electronic information system, and to examine and evaluate protections and alternate processes for handling information to mitigate potential privacy concerns) as necessary when developing and implementing a new technology.[24]
- Differential-privacy guarantees, or protections that allow information about a group to be shared while provably limiting the improper access, use, or disclosure of personal information about particular entities, should be understood when in place, including how these guarantees affect data shared with, used to train, or created by GenAI technologies.[25]
- GenAI can also facilitate or interact with privacy-enhancing technologies (PETs), or "any software or hardware solution, technical process, technique, or other technological means of mitigating privacy risks arising from data processing, including by enhancing predictability, manageability, disassociability, storage, security, and confidentiality. These technological means may include secure multiparty computation, homomorphic encryption, zero-knowledge proofs, federated learning, secure enclaves, differential privacy, and synthetic-data-generation tools. This is also sometimes referred to as "privacy-preserving technology."[26] Understand how GenAI may be used to increase privacy in newly developed or existing capabilities, or how the introduction of GenAI into a system may affect existing PETs.
- Per the NIST AI RMF, "privacy-enhancing technologies…as well as data minimizing methods such as de-identification and aggregation for certain model outputs, can support design for privacy-enhanced AI systems. Under certain conditions such as data sparsity, privacy-enhancing techniques can result in a loss in accuracy, affecting decisions about fairness and other values in certain domains."[27]

Additional resources

- DOE O 206.1 Department of Energy Privacy Program, January 16, 2009
- Office of Science and Technology (OSTP), Blueprint for an AI Bill of Rights, October 2022

8.6 Confidentiality

As noted in Section 7.3, the word "confidential" *when used in this Guide* falls outside of the NSI definition. Confidential information input into a GenAI tool may be stored or processed by the tool or its providers, revealing confidential information to unauthorized personnel. Within US government documents, the word "Confidential" has a specific National Security Information (NSI) definition that relates to the level of the severity of harm if a document marked "Confidential" is inappropriately shared. In this Guide, use of the word "confidential" is outside of the NSI context. Instead, it is associated with the non-governmental business environment.

Definition

Confidentiality, as discussed in this Guide, is defined as "preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information."[28]

The DOE Operations Security Handbook states that there are two primary characteristics of a piece of information that determine whether that information is safe for public disclosure or whether it should be considered **sensitive**. These two primary characteristics for "determining suitability for release of information" are sensitivity and risk.

- Sensitivity: "If the information is released to the public, it should not reveal or identify sensitive information, activities, or programs." Sensitive information can also be defined as information that could be used by adversaries to the detriment of the organization, its employees, the public, or the nation. Sensitivity gauges the level of harm that could ensue from release.
- Risk: "Information that may be used by adversaries to the detriment of employees, the public, the department, or the nation should not be approved for release. This determination should be based on sound risk management principles focused on preventing potential adverse consequences." In terms of the definition of sensitivity presented above, risk the likelihood of such harm.

Together, these two characteristics suggest that the term "sensitive information" is a categorical term which includes other specific types of sensitive information.[29]

Refer to Appendix J for a list of protected data types and definitions.

Examples

Public example 1

Both Apple and Samsung have instituted restrictions on the use of OpenAI's ChatGPT and Microsoft's GitHub Copilot by some of their employees because of concerns over the potential for employees to mishandle and leak confidential company data. This move aligns with a growing trend of companies and governments worldwide imposing restrictions on the use of GenAI platforms. In April 2023, OpenAI released a series of updates to ChatGPT that enabled better privacy controls after some nations voiced their concerns.

Read more here: Apple Restricts Employee Use of ChatGPT, Joining Other Companies Wary of Leaks | WSJ

Public example 2

Samsung Electronics banned the use of any AI-powered chatbots and ChatGPT by its employees because of concerns about sensitive internal information being leaked. This decision comes after an accidental leak of sensitive source code through ChatGPT, prompting the company to issue a memo banning the use of GenAI tools. Even though the exact severity of the leak is unknown, data shared with chatbots may be stored on servers owned by outside companies operating the service, such as like OpenAI, without the ability for Samsung to access or delete the data.

Read more here: Samsung Bans ChatGPT Among Employees After Sensitive Code Leak | Forbes

Key considerations

- Information entered into a GenAI system may become part of its training data set. Thus, any confidential data entered as part of a prompt could be used in outputs for other system users, which could result in unintended exposure or misuse of this information. Confidential information should not be used in any way which could lead to the information being shared outside of its intended or authorized used.
- Data provenance, digital rights management, and understanding of data rights are critical to responsible GenAI management. Digital rights management programs should be robust in order to avoid unauthorized data usage.
- GenAI systems can store data and information input as prompts indefinitely.
- Adversarial players can hack the system to gain access to any stored confidential data.

Best practices

- Do not input or disclose protected data or nonpublic information, as part of a prompt when using a public GenAI tool, unless you can validate rights to use it in this way from the originator. Refer to the most recent versions of DOE information security policies for specific guidance. Seek advice from your organization's legal department.
- Do not rely on GenAI to generate confidential or mission-critical information or data, as the information used to train AI models may not be accurate, complete, or without bias. Review and continue to adhere to existing policies, procedures, and guides to ensure compliance with National Laboratories and DOE information requirements. Additionally, continue to follow existing requirements, such as those regarding quality, information security, and integrity. Work with appropriate DOE and National Laboratory SMEs and compliance organizations, such as the Office of General Counsel, the Office of Export Control, the Classification Office, the Office of Environment, Health, Safety, and Security (EHSS), and others as appropriate.
- Continue to follow existing cybersecurity and privacy procedures and iteratively review existing policies and procedures to understand how confidentiality should be applied generally within an organization, to ensure that new policies and regulations are implemented in issued procedures, and to stay current with the most up-to-date requirements.
- Encourage prompt engineering training for GenAI users to learn the best ways to structure prompts that generate more accurate outputs (refer to Appendix I for more details on prompt engineering).
- Review output produced from GenAI tools to ensure that any issues surrounding confidentiality are identified and addressed.
- Practice secure storage and processing of protected data and nonpublic information and implement access controls. Clearly define who has access to the data and the purpose of its use every time.
- Note that contract solicitation responses are proprietary (confidential) information.
- Refer to existing DOE training on confidentiality that employees are required to take.
- It is imperative that users understand the legal rights the Government (has or doesn't have) in the data inputted into AI tools and/or which are used to train data LLMs.

Additional resources

- DOE mandatory training on confidentiality: CUI-100DE Controlled Unclassified Information Overview

8.7 Intellectual Property

GenAI tools introduce risks surrounding intellectual property (IP), including copyright and data protections, as the tools can access copyrighted works and generate outputs that closely resemble content from these works. It should also be noted that the intersection of GenAI and existing laws, regulations, and policies regarding IP and copyright is dynamic and evolving. This guide is not a replacement for legal advice; therefore, any legal questions related to the intersection of GenAI and IP should be directed to cognizant DOE or contractor legal counsel.

Definition

Intellectual property (IP) is intangible property that is the product of an original thought, including inventions, designs, writings, images, and names, much of which is protectable by statutory and contractual rights, including patents, copyrights, trade secrets, and trademarks (intellectual property rights or IPR). Data may also be protectable as IP, typically as copyrightable compilation, if selected and arranged in a unique and original way, such as with a dataset. Data in such form may be copyrightable and licensable.

Copyright is not a single right, but a bundle of rights that include not only reproduction, but also provide the copyright owner the right to prevent others from adapting, distributing to the public, performing, and displaying the copyright work (including digitally).

Intellectual property issues have significant legal, financial, and ethical implications.

Examples

Public example 1

Comedian Sarah Silverman and two authors filed a class-action lawsuit on July 7, 2023, against OpenAI and Meta, accusing them of copyright infringement for the use of their protected work in the companies' training datasets. According to the lawsuit, "copyrighted materials were copied and ingested as part of training." While the outcome is still pending, the context surrounding this lawsuit is of great importance, as the training dataset can include copyrighted materials "without permission by scraping illegal online "shadow libraries" that contain the text of thousands of books," as mentioned by The New York Times.

Read more here: Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement | The New York Times

Public example 2

In Andersen v. Stability AI et al., a case filed in 2022, three artists filed a lawsuit against several GenAI vendors on the grounds that the GenAI systems used their original works as part of a training set. Users of these systems were able to generate works very similar to the original artists' works. If a court rules that the GenAI-generated works are derivative and unauthorized, considerable penalties may apply.

Read more here: Generative AI has an intellectual property problem | Harvard Business Review

Key considerations

- Under current law, an invention created solely by an AI tool is not able to be patented or copyrighted because the output is not created by a human.[30] In August of 2023, the United States Copyright Office, a subsidiary of the Library of Congress, released a request for public comment on the interaction of AI and copyright law which provides a picture of the types of discussions which are ongoing in this area of legal study.[31] Inventors should discuss specific cases with a patent attorney, as the Department of Energy policy is unable to determine legality. The U.S. Copyright Office has also issued several statements informing creators that it will not register copyrights for works produced by a machine or computer program.

- On February 13, 2024, the United States Patent and Trademark Office (USPTO) issued new guidance that explains that while AI-assisted inventions are not categorically unpatentable, the inventorship analysis should focus on human contributions, as patents function to incentivize and reward human ingenuity. Patent protection may be sought for inventions for which a natural person provided a significant contribution to the invention, and the guidance provides procedures for determining the same.[32]
- In August of 2023, the United States Copyright Office, a subsidiary of the Library of Congress, released a request for public comment on the interaction of AI and copyright law which provides a picture of the types of discussions which are ongoing in this area of legal study.[33] Inventors should discuss specific cases with a patent attorney, as the Department of Energy policy is unable to determine legality. The U.S. Copyright Office has also issued several statements informing creators that it will not register copyrights for works produced by a machine or computer program.
- GenAI tools are trained on huge sets of scraped data, and that training forms the basis for the model's responses to prompts. Therefore, GenAI tools may generate outputs that contain plagiarized or copyrighted information.
- Considerable care should be taken not to infringe intellectual property rights, or violate other protections, when inputting data into GenAI prompts or otherwise using data to train LLMs.

Best practices

- Adhere to existing policies and procedures regarding copyright issues, and continue to monitor for changes in copyright laws, policies, and recommended procedures that apply to GenAI tools.
- Have a human in the loop, preferably someone who has knowledge of GenAI, to validate generated output sources and prevent plagiarism and/or copyright issues.
- Use caution when using the outputs of a model in other work, and keep in mind that large language models will not reliably tell you if their sources are in the public domain or not.
- When GenAI solutions play a role in creating an idea, approach, or invention at DOE or a National Laboratory, employees must clearly identify the specific contribution (e.g., attribution in a report, laboratory record, invention disclosure) and cite the GenAI as part of their research methodology. It may be essential to determining patentability or copyrightability to know specific attributions to humans versus AI technologies. Several style guides and publishing houses are developing guidance on how to appropriately credit AI tools in written work (See the additional resources below). Employees should follow these guidelines where they exist.
- Educate users on the challenge of AI generating content that may infringe on existing copyrights and promote an understanding of intellectual property rights with regard to GenAI.
- Use secondary tools to identify and validate sources, context, and citations, particularly in cases where the user has *some* prior knowledge.
- Avoid using GenAI output to create website content unless the origin of the training data is verified as appropriate for the given use, as it may contain or have been trained on confidential or sensitive data. See Section 8.6: Confidentiality.
- Employ best practices for prompt engineering (refer to Appendix I for additional information on prompt engineering).

Additional resources

- The Use of Copyrighted Materials by Government Employees, Department of Energy
- Congressional Research Service: Generative Artificial Intelligence and Copyright Law
- Generative AI Has an Intellectual Property Problem
- "How to cite ChatGPT," Timothy McAdoo, APA Style, April 7, 2023

8.8 Safety

GenAI systems must be designed to be safe for system users and society in general. Outputs resulting from GenAI systems should not compromise the safety of individuals or their health or property.

Definition

AI systems are **safe** if they do "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered."[34] **Safe** operation of AI systems is achieved through:

- Responsible design, development, and deployment practices
- Clear information to deployers on responsible use of the system
- Responsible decision-making by deployers and end users (e.g., via reinforcement learning, defined in Section 5.1)
- Explanation and documentation of risks based on empirical evidence of incidents[35]

Examples

Public example 1

In 2022, it was reported that Tesla vehicles utilizing the AI-assisted Autopilot functionality had been involved in 273 crashes during the previous year. These included crashes with other cars and motorcycles, and pedestrian and driver deaths. The autopilot functionality includes the ability to maintain speed and safe distance behind other cars, to stay within their lane lines, and to make lane changes on highways. According to Tesla, however, human drivers are supposed to keep their eyes on the road and their hands on the wheel, with the technology serving as an assistant. This human oversight is critical to safe operation of the vehicle.

Read more here: Teslas running Autopilot involved in 273 crashes reported since last year | The Washington Post

Public example 2

President Biden announced that leading AI companies, such as OpenAI, Alphabet, and Meta, "have made voluntary commitments to the White House to implement measures such as watermarking AI-generated content to help make the technology safer." This will enable the identification of when content was generated by AI, and most importantly, the identification of deepfakes that can spread misinformation or be used to defraud individuals. The companies also made pledges to test systems thoroughly before release and to focus on protecting users' privacy. These commitments are steps toward ensuring safeguards in GenAI.

Read more here: OpenAI, Google, others pledge to watermark AI content for safety, White House says | Reuters

Key considerations

- Various types of risks involving safety might necessitate custom AI risk mitigation strategies depending on the context and the severity of the potential risks.
- Safety relates mostly to the use and application of the system. Safety risks can arise from both negligence and deliberately malicious intent.
- Establish mechanisms to support reproducibility and ability to scrutinize outputs for accuracy and consistency through versioning and provenance of training inputs, model parameters, data corpus, and other key system elements.

Best practices

- Adopt a safety-by-design approach and mentality by considering safety risks throughout the AI lifecycle, starting as early as possible during the planning and design phases.
- Do not use GenAI for malicious or deceptive activities, such as the creation of malware, identity theft, or identity impersonation.
- Develop safety measures for AI system deployment, including checks against harmful and unintended uses.
- Leverage the guidelines for safety in the transportation and healthcare fields and align with the existing sector- or application-specific guidelines or standards in AI safety risk mitigation strategy (e.g., the NIST AI Risk Management Framework (NIST AI RMF)).
- Have a human in the loop throughout the AI lifecycle. Human oversight, validation, and verification, are a combined, iterative process beginning in the planning and design phase and continuing throughout the AI lifecycle, including after the model is deployed.
- Have the appropriate responsible AI solutions controls in place.
- Consider using the prompt (any modality, such as a sensor) as an interface with the system about the current state that can be used as a control for increasing safety.
- Account for secondary usage of outputs by ensuring that any caveats, considerations, and/or assumptions are tacked on to the output so that the output will not be inadvertently misused. Intentional malicious misuse is outside of our control.
- Per the NIST AI RMF, “AI safety risk management approaches should take cues from efforts and guidelines for safety in fields such as transportation and healthcare and align with existing sector- or application-specific guidelines or standards.”[36]

8.9 Fairness and Bias

GenAI introduces challenges in defining, measuring, and addressing concerns about fairness and bias in a number of ways.

Definition

Fairness in AI involves addressing issues such as harmful bias and discrimination to foster equality and equity. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application.[37]

GenAI systems should be designed to be **fair** so that individuals or groups are not systematically disadvantaged through AI-driven decisions. Achieving fairness in AI can be challenging, as it requires careful consideration of different types of bias and using the technology in a way that avoids favoritism or discrimination, particularly to humans.

Bias refers to the systematic and consistent deviation of an algorithm's output from the true value or from what would be expected in the absence of bias.[38] Bias is a component of fairness and comes in many forms, going beyond lack of demographic balance or data representativeness. NIST has identified three major categories of AI **bias** to be considered and managed: *systemic, computational and statistical,* and *human-cognitive*. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent.

- *Systemic bias* can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems.
- *Computational and statistical biases* can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to nonrepresentative samples.
- *Human-cognitive biases* relate to how an individual or group uses AI system information to decide or fill in missing information, or how humans think about an AI system's purposes and functions. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.[39]

While fairness and bias are closely related concepts, they differ in important ways. The key difference is that while bias can be unintentional, fairness is inherently a deliberate and intentional goal. In other words, bias can be viewed as a technical issue, while fairness is a social and ethical issue.[40]

Examples

Public example 1

Stable Diffusion's text-to-image GenAI solution has been identified by Bloomberg as a model that contributes to biased racial and gender stereotypes. Bloomberg used Stable Diffusion's tool to create thousands of images pertaining to crime and employment. In this analysis, the model was prompted with text to create images of workers for 14 jobs — 300 images for seven jobs generally considered as "high-paying" in the U.S. and 300 images for seven jobs generally considered "low-paying" — as well as three topics related to crime in the U.S. The analysis discovered that images generated for the high-paying jobs were of people with lighter skin tones. In contrast, images generated with darker-skinned

Public example 2

A class action lawsuit was filed against HR and financial management software provider Workday, alleging that the software produced a screening system that resulted in racial bias. The lawsuit alleges that Workday "unlawfully offers an algorithm-based applicant screening system that determines whether an employer should accept or reject an application for employment based on the individual's race, age, and/or disability." The plaintiff of the lawsuit states that Workday's AI tools rely on algorithms that may be riddled with human bias.

Read more here: Workday wants racially biased recruitment algorithm claim thrown out | The Register

Examples

subjects were created by the solution in response to prompts like “fast-food worker.” The conclusion of this analysis implies that racial and gender representation in various career images was significantly different than the representation in the actual careers. For instance, about 3% of the images generated for the prompt “judge” were women, whereas in reality, about 34% of American judges are women.

Read more here: Humans are biased. Generative AI is even worse | Bloomberg

Key considerations

- Fairness needs to be defined for every use case at the beginning of the design and planning phase, as it can mean different things in different contexts.
- Outputs from GenAI systems can and will produce bias if bias is included in the training, validation, or test datasets, which is usually the case.
- Bias can be introduced at any point in the AI lifecycle.
- Non-stationary data, or data that is dynamically and unpredictably changing over time, can produce bias.
- Scaling of bias and/or nonfactual information gets out of hand more quickly with GenAI than with non-generative AI.
- When outputs/outcomes are unfair or include bias, the decisions they inform may impact the communities that may be affected by these algorithms in an unfair and biased way.

Best practices

- Do not solely rely on GenAI systems for making decisions. Instead, use the systems to help inform decisions.
- Incorporate algorithmic fairness by design by including concepts of fairness throughout the AI lifecycle. Consider developing and using a fairness and bias checklist for each stage.
- Implement equitable data management to carefully consider and prioritize the needs of disadvantaged communities when it comes to data and information management.
- Ensure that there are human-centric strategies throughout the AI lifecycle that can be used in the event of a failing GenAI system that impacts people’s rights, including by always having a human in the loop to verify that outputs are representative and will not have negative consequences that could impact fairness or bias considerations.
- Keep in mind the initial intent behind the system’s existence throughout the AI lifecycle and regularly check that the system is on track for the intended use and outputs.
- Ensure that AI systems operate fairly and transparently by establishing fairness metrics, and regularly review GenAI systems and output for compliance and representation. Understand and mitigate the risks of discriminatory outcomes to maintain equitable access and benefit from GenAI.
- Develop procedures for reviewing, sharing, and evaluating GenAI systems through the lens of fairness and bias. This includes performing analysis in the beginning stages of the GenAI initiative to become aware of the different biases that may occur, as well as implementing procedures that identify and mitigate bias in training data and that use diverse data sets for training to avoid amplifying existing bias.

- Train, validate, and test GenAI models using representative datasets that are as "fair" as possible, as defined by the context of the use case at the beginning of the design phase of the project. This includes ensuring that the training model is representative of society at large or society given the topic at hand and assessing fairness in datasets by identifying representation, corresponding limitations, and any prejudicial or discriminatory correlations between features, labels, and groups.
- After attempting to address fairness in the dataset, including filtering if needed, recheck the model outputs to see if any *new* artificial bias has been created. Apply normalization adjustments or other mathematical balances to correct for new unintended biases or distortions of perceived fairness introduced by previous attempts to address original biases.
- Check AI systems for unfair biases and consider the effects of biases created by previous outputs incorporated into the training sets and the feedback loops this may create.
- Test the results the GenAI model produces by asking questions with known answers. Give examples of questions asked in different ways to test how the results change.
- Provide caveats and assumptions with respect to the GenAI sources in the output, much like footnotes.
- Use additional caution when GenAI is used for an activity that may be regulated by the Equal Opportunity Employment Act, Civil Rights Act, or Americans with Disabilities Act (e.g., hiring, resume sorting, recruiting, or any other HR function).
- Keep an eye out for tools, processes, and organizations that can check for trustworthy AI (to include detection of bias). This is an emerging market where valuable solutions may arise.
- Be aware of the market and the reputation of the tools in use and of the fact that these factors can and will change rapidly.
- Continually build awareness throughout the organization around fairness and identifying bias, including by educating users on how biases in data can lead to biased output.

Additional resources

- DOE EO 13960 Consistency Plan
- Generative AI Takes Stereotypes and Bias from Bad to Worse
- AI Fairness 360

8.10 AI Hallucinations and Misinterpretations

Hallucinations occur when a GenAI system produces output that is not based on actual or existing data but instead is often imaginative or unrealistic content generated from beyond the system's training set. These hallucinations can lead to the spread of false information.

Definition

An **AI hallucination** occurs when a GenAI system provides a response that includes irrelevant, false, or nonsensical information. These responses are often articulate and confident but contain information that is not true. Note that when the model hallucinates, it is not intentionally lying as it is not motivated to deceive users, nor does it have the awareness that it is providing false information. Hallucinations are also known as generated errors, confabulations, delusions, or fabrications. These occur when the AI system misinterprets its training data and uses it to create responses that are not factual.

Keep in mind the concepts of validation, reliability, and explainability when discussing hallucinations and the best practices to mitigate them.

- **Validation** is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled."[41]
- **Reliability** is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions."[42]
- **Explainability** refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes.[43]

Examples

Public example 1

During testing of Microsoft's AI-powered search engine Bing, users experienced unsettling and offensive behavior from the chatbot, which engaged in hostile and abusive conversations. Similar incidents have occurred with other chatbots, raising concerns about the responsible development and deployment of AI tools. Microsoft's rush to release an AI-powered chatbot without conducting thorough studies on its potential for inappropriate responses during prolonged user interactions became clear. Extensive testing could have helped identify these issues. While companies are working on refining these systems and implementing limits on user interactions, the incidents highlight the challenges of managing the output of AI chatbots and the need for ongoing research and development in this field.

Read more here: Microsoft's new AI chatbot has been saying some 'crazy and unhinged' things | NPR

Public example 2

A lawyer involved in a personal injury lawsuit against an airline used AI chatbot ChatGPT to prepare a filing but ended up presenting fake cases to the court. The lawyer, Steven Schwartz, claimed he was unaware that the tool, which he mistook for a search engine, would generate false information. The judge is considering imposing sanctions, as this case highlights the issue of AI hallucinations and raises questions about the legal community's handling of AI-generated content. The incident has sparked concerns about the ability to detect AI-generated fakes and the potential impact on trust in society. Concerns are high that GenAI is advancing beyond human ability to detect fakes, which will lead to a heightened lack of confidence in society.

Read more here: Lawyer used ChatGPT in court and cited fake cases. A judge is considering sanctions | Forbes

Key considerations

- Hallucinations occur frequently with GenAI systems and can be hard to identify when they occur in a generally coherent, articulate response.
- Without any way to verify information, hallucinated responses could be interpreted by the user as accurate and factual.
- Training data sets, especially those comprising unknown data, might include nonfactual or nonsensical information that can directly contribute to hallucinations.
- A lack of transparency into the training set, algorithm, and model creates a "black box" effect, where there is no insight into how the GenAI system arrived at its inaccurate or nonfactual hallucinatory output.
- When hallucinations occur often, the GenAI model is not aligning with the principles of validation, reliability, and explainability.

Best practices

- Have a human in the loop to verify the accuracy and validity of outputs.
- Develop AI literacy among users so they understand the limitations and potential risks of GenAI. Focus on teaching how to manage prompts appropriately and direct desired output.
- Use best practices for prompt engineering to mitigate the risk of hallucinations (refer to the Prompt Engineering information in Appendix I).
- Use reinforcement learning with human feedback (RLHF) to improve the reliability and accuracy of the model and enhance the model's alignment with its human users. RLHF entails a team of humans who score the model's outputs (e.g., high scores for reliable, factual outputs, and low scores for hallucinated outputs), and the model is subsequently trained to generate outputs that are most acceptable to the team of humans.[44]
- Employ "grounding" as a technique to mitigate the risk of hallucinations. Use external information from trusted sources to prompt the GenAI model to generate a response based upon the retrieved, factual information within the given context. The most common grounding technique currently used with LLMs is "Retrieval Augmented Generation (RAG). Using this approach, the input to the LLM is leveraged to retrieve the grounding information from a specified database and then to feed it to the LLM along with the user's original input. If the model does not have the ability to respond based solely on the relevant information used to ground the system AND within the appropriate context, the model will return "not enough information" instead of a hallucinated response. [45]

Additional resources

- Prompt Engineering Course: ChatGPT Prompt Engineering for Developers, OpenAI
- Understanding Hallucinations in AI: A Comprehensive Guide, Pinecone

8.11 Best Practices Checklist

This section provides a summary of GenAI best practices, divided into three categories: **People, Organization,** and **Technology**. The People category contains best practices related to staff, leadership, subject matter experts (SMEs), and workforce development. The Organization category includes best practices related to processes, policies, business needs, and governance. The Technology category contains best practices on AI technical development, implementation, monitoring, and security. Each of the three categories includes content related to role-specific best practices and data management best practices. Within each category, the best practices are organized by their relative location within the AI lifecycle. For additional information on the AI Lifecycle, see Appendix G.

Per Executive Order 14110 on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, a variety of new policies, procedures, standards, and best practices will be created regarding development and usage of AI in the federal space, including "guidelines and limitations…on the appropriate use of GenAI."[46] Please refer to these resources as they are developed and published. Executive Order 14110 also encourages the development of new AI capability testbeds and model evaluation tools, facilitated by the Department of Energy in collaboration with other stakeholders. These testbeds and evaluation capabilities may provide critical support in adhering to the best practices detailed below.[47]

8.11.1 General Best Practices: People

- Consult a diverse group of subject matter experts (SMEs), stakeholders, and impacted communities from the outset of the project to identify the system's benefits and risks, to help shape the solution, to advise on use, and if appropriate, to perform an evaluation led by an independent third party or by experts who do not serve as core developers for the system.[48] The process of testing GenAI technologies should be collaborative across the DOE
- Have a human in the loop throughout the AI lifecycle. Apply an iterative process beginning in the planning and design phase and continuing throughout the AI lifecycle to check for accuracy, validity, fairness, representativeness, bias, plagiarism, and/or copyright issues. Develop specific procedures to ensure these checks occur.
- Pilot GenAI systems by making both business leaders and engineers active participants in the piloting process. Piloting cycles are fast-paced and focus on short experiments and use cases that provide strategic value while mitigating risks and eliminating use cases that are not viable.[49]
- Train users to understand the security risks associated with AI, as well as bias and fairness issues. Include relevant existing DOE trainings around these topics in AI curricula where available.
- Educate users on the challenge of AI generating content that may infringe on existing copyright or intellectual property, and note AI contributions to creating an idea, approach, or invention at the DOE or National Laboratory in research methodology.
- Update position descriptions, recruitment practices, and workforce development documents to ensure that GenAI-specific technical expertise is being prioritized in relevant job responsibilities, qualifications, trainings, guidance, and resources, keeping in mind DOE's current prioritized use cases (refer to the DOE Use Case Inventory).
- Provide up-to-date trainings on GenAI risk mitigation, prompt engineering, and usage best practices for personnel involved throughout the AI lifecycle. = Include tailored GenAI privacy trainings for employees with access to protected or nonpublic data.
- Provide opportunities for upskilling existing employees — trainings, relevant learning materials, digital literacy programs, etc. — and update the materials regularly.
- Provide resources to educate the organization's general users about how GenAI tools do not necessarily provide factual outputs. Users should understand that outputs from GenAI tools should

be considered a “first draft.” One option to denote these documents is the inclusion of watermarks marking the document as an initial draft or AI output.

8.11.2 General Best Practices: Organization

- When starting to plan any GenAI initiative, define the GenAI system’s objective(s) and intended purpose. Data and use-case scope limits should be explicitly documented.
- Keep in mind the initial intent and purpose of the system and regularly check that use and outputs align to that intent.
- Adopt a safety-by-design and fairness-by-design approach and mentality by considering risks throughout the AI lifecycle and include checks for these principles in the system.
- Establish a plan to routinely identify and mitigate risks and vulnerabilities that arise from a GenAI system, including those reported by users.[50]
- Assess fairness and representativeness in datasets by identifying representation, corresponding limitations, and any prejudicial or discriminatory correlations between features, labels, and groups.
- Employ an agile approach in organizational communications on AI-related topics and issues, as these will change with rapidly evolving technology.
- Implement full lifecycle stewardship, which is the practice of securing and protecting data, metadata, and information throughout its lifecycle. That includes collection, storage, use, control, processing, publication, transfer, retention, and disposition. Document this process thoroughly.
- Develop measurement tools and data documentation to capture information about the GenAI system and its training, operation, and outputs in the production environment.[51]
- Check GenAI systems and their training datasets for unfair biases. Consider the effects of biases created by previous outputs incorporated into the training sets and the feedback loops this may create.[52] Implement procedures that will identify and mitigate these biases and use diverse training data whenever possible.
- Review and continue to adhere to new and existing policies, procedures, and guides to ensure compliance with National Laboratory and DOE information requirements, including when it comes to topics such as quality, privacy, IP, information security, and integrity.
- Establish a plan to routinely identify and mitigate risks and vulnerabilities that arise from GenAI systems, including those reported by users.[53]
- Create mitigation plans for likely AI attack vectors, such as deepfakes and data poisoning (when an adversarial player pollutes training data to manipulate the model's output), and keep in mind that no foolproof defense currently exists to completely prevent such attacks. Refer to the NIST publication “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations” for more information.[54]
- Require vendors and developers who design GenAI systems and create features with human-facing interfaces to provide traceable feedback on system status and clear procedures to trained operators to activate and deactivate system functions.[55]
- Revisit contracts with third-party providers and agreements regarding the sharing of AI-specific risks.
- Ensure that senior leadership has reviewed the risks of the GenAI system. Leadership needs to be aware that they are ultimately responsible for the proper use and function of the GenAI system.
- Consider forming and supporting strong governance structures as a part of GenAI system implementations, including formal data governance and model governance initiatives with aligned organizational governance boards.

8.11.3 General Best Practices: Technology

- Start with providing a safe space (e.g., a sandbox) for experimentation and scale when residual risk is shown to be acceptable.[56]
- Exercise data minimization best practices when collecting data for a system.
- Consider using synthetic data (generated algorithmically and used as a stand-in for real data) to mitigate risks surrounding privacy, data integrity, and insufficient data.[57] Another option would be to utilize publicly available data from Data.gov or DOE public reports if applicable and legally appropriate.
- Ensure system observability (i.e., the extent to which the internal states of a system can be inferred from externally available data) via three pillars of observability data: metrics, logs, and traces.[58]
- Encourage best practices for prompt engineering to ensure accuracy and efficacy.
- Do not include nonpublic (classified, private, sensitive, confidential, etc.) information as part of a prompt/input to any commercial or open GenAI system. Refer to Appendix J for examples of protected data.
- Do not use GenAI for malicious or deceptive activities, such as creation of malware, identity theft, or identity impersonation.
- Establish mechanisms to support reproducibility and the ability to scrutinize outputs for accuracy and consistency through versioning and provenance of training inputs, model parameters, data corpus, and other key system elements.
- Develop robust, secure systems and detection capabilities for sites, regularly update them to defend against threats, and plan for system resilience.
- Establish a program to conduct adversarial testing or red teaming.[59]

9. Conclusion

GenAI is an emerging technology with a great deal of potential for driving innovation, efficiency, and value at DOE. GenAI adoption is exploding throughout the world, and the variety of tools available on the market continues to expand. GenAI technology can provide value by performing certain routine tasks, like summarizing a document or drafting an email, in a substantially shorter time frame than human employees. The future of the DOE workplace could be one in which a symbiotic relationship exists between human employees and GenAI tools.

GenAI technology is evolving rapidly, and the landscape of risks and considerations is not yet completely understood. The pace at which GenAI is advancing means that DOE needs to stay agile and up to date on the current thinking. This document outlines potential use cases, prominent risks, and best practices for GenAI use in the context of DOE. Expect these topics to continue to evolve as additional use cases and risks emerge, as best practices are refined, and as additional legislation and regulations are published by the U.S. Government. This document will be updated regularly to reflect the evolving landscape of GenAI.

10. Appendices

Appendix A. Acknowledgements

The DOE Generative AI Reference Guide was developed by the Department of Energy Office of the Chief Information Officer (OCIO) with support from Ernst & Young LLP. The authors of this Reference Guide and the Department of Energy OCIO would like to thank the DOE Generative AI Reference Guide Tiger Team, which included representatives from the following organizations, for contributing their time and expertise:

- Bonneville Power Administration **(BPA)**
- Brookhaven National Laboratory **(BNL)**
- Critical and Emerging Technologies **(CET)**
- Energy Efficiency and Renewable Energy **(EE)**
- Environment, Health, Safety and Security **(EHSS)**
- Environmental Management **(EM)**
- Fossil Energy and Carbon Management **(FECM)**
- Idaho National Laboratory **(INL)**
- Lawrence Berkeley National Laboratory **(LBNL)**
- National Nuclear Security Administration **(NNSA)**
- National Renewable Energy Laboratory **(NREL)**
- Oak Ridge National Laboratory **(ORNL)**
- Office of Science **(SC)**
- Pacific Northwest National Laboratory **(PNNL)**
- Sandia National Laboratories **(SNL)**
- Savannah River National Laboratory **(SRNL)**
- Savannah River Site **(SRS)**
- Stanford National Accelerator Laboratory **(SLAC)**
- Thomas Jefferson National Accelerator **(JLAB)**
- Under Secretary for Science and Innovation **(S4)**
- Western Area Power Administration **(WAPA)**

Additionally, we would like to recognize the following contributors for their significant contributions and valuable feedback, including the Tiger Team Co-Chairs, Sponsor Team, and individuals from the Tiger Team who went above and beyond in their participation:

Aaron Haglund	Gardy Rosius (Co-Chair)	Lance Roeske
Ahmad Sultan	Greg Doan	Malachi Schram
Brad Wilson	James P. Lively	Margaret Lentz
Brian Post	Jason Talley	Maria McClelland
Bridget Carper (Executive Sponsor)	Jayu Wu	Randy Steer
Brooke Dickson	Jodi Kouts (Sponsor)	Robert King
Christian Stauffer	Jonnie Bradley (Co-Chair)	Rochelle Blaustein
Darrell Beschen	Kathleen Oprea	Sandra Logan (Sponsor)
Elaine Ulrich	Ken Hunt	Steven Wong
Erica Vosseller	Kenneth Calabrese (Sponsor)	Tom Harper (Co-Chair)
Felix Gonzalez	Kerstin Kleese Van Dam	Vicki Michetti (Executive Sponsor)

Finally, we would like to thank the Department of Energy's Chief Information Officer, Ann Dunkin, for her leadership and support throughout this process.

Appendix B. Referenced Documents

DOE Internal References

- DOE 13960 Consistency Plan
- DOE CUI Slicksheet
- Home — DOE Directives, Guidance, and Delegations
- Standards of Ethical Conduct - Policy Memorandum #93
- Controlled Unclassified Information Sheet

Public References

- DOE AI Risk Management Playbook (AIRMP)
- DOE 2023 AI Use Case Inventory
- Federal Acquisition Regulation (FAR)
- "Protected Critical Infrastructure Information (PCII) Information Program," Cybersecurity and Infrastructure Security Agency (CISA)
- What is Intellectual Property?
- Congressional Research Service: Generative Artificial Intelligence and Copyright Law
- Generative AI Has an Intellectual Property Problem

- What is Copyright?
- Generative AI Takes Stereotypes and Bias from Bad to Worse
- AI Fairness 360

Appendix C. Relevant External Learning Resources

- Prompt Engineering
  - ChatGPT Prompt Engineering for Developers, DeepLearning.AI ChatGPT 101: Supercharge Your Work & Life (750+ Prompts Included)
- ChatGPT Masters: Generative AI, Prompt Engineering, Chat GPT | Udemy
- Google GenAI Learning
  - New Google Cloud generative AI training resources: Seven New No-Cost GenAI Training Courses| Google Cloud Blog
- Microsoft Open AI Learning
  - An introduction to the Azure OpenAI Service
  - Explore how customers are putting Azure AI to work for them
  - Azure OpenAI: Business Briefing
  - Azure OpenAI: Technical Briefing
  - Azure OpenAI: Learn about ChatGPT and DALL·E
  - Azure OpenAI product page
- ChatGPT, LLMs & Generative AI: What Your Business Needs to Know
- Generative AI for Marketing
- Embracing AI: The Future of Content Creation, Ernesto Anaya, TEDxSCCS Youth
- How HR and IT Can Use AI to Build a Skills Based Workforce
- Webinar - Role of Generative AI in the Future of Recruitment
- Elements of AI Free Online Course

Appendix D. Relevant DOE Trainings and Learning Resources (located in the DOE Learning Nucleus)

- Introduction to Artificial Intelligence (Learning Nucleus ID: 55476)
- AI-900: Azure AI Fundamentals: AI & ML (Learning Nucleus ID: 84436)
- Artificial Intelligence: Basic AI Theory (Learning Nucleus ID: 76001)
- Artificial Intelligence: Human-computer Interaction Overview (Learning Nucleus ID: 79429)
- Artificial Intelligence: Human-computer Interaction Methodologies (Learning Nucleus ID: 79430)
- Artificial Intelligence: Types of Artificial Intelligence (Learning Nucleus ID: 76002)
- Elements of an Artificial Intelligence Architect (Learning Nucleus ID: 79433)

- AI-900: Azure AI Fundamentals: Using Azure Machine Learning Studio (Learning Nucleus ID: 84438)
- AWS Certified Machine Learning: AI/ML Services (Learning Nucleus ID: 84340)
- CUI-100DE Controlled Unclassified Information Overview

Appendix E. Further Details on Federal Policies, Guidelines, and References

1. Office of Management and Budget Memorandum, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, March 28, 2024

This memorandum addresses a subset of AI risks, as well as government and innovation issues directly tied to agencies' use of AI. The risks addressed in this memorandum result from any reliance on AI outputs to inform, influence, decide, or execute agency decisions or actions, which could undermine the efficacy, safety, equitableness, fairness, transparency, accountability, appropriateness, or lawfulness of such decisions or actions. Consistent with Section 104(c) and (d) of the AI in Government Act of 2020, within 180 days of the issuance of this memorandum, and every two years thereafter, DOE will issue a plan to achieve consistency with this memorandum. Agencies must also include plans to update any existing internal AI principles and guidelines to ensure consistency with this memorandum.

2. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 2023

Executive Order 14110 outlines a new set of definitions for key AI terminology, while outlining a variety of new actions to be taken by the Executive Branch, federal agencies, research institutions, the private sector, and international partners to increase the use of AI while managing related risk and protecting civil rights and workers' rights. The EO also mandates the creation of Chief AI Officers (CAIOs) at federal agencies which meet the required criteria and seeks to accelerate hiring and training of AI professionals across the government.

3. Congressional Research Service (CRS), Generative Artificial Intelligence and Data Privacy: A Primer, May 2023

Generative AI poses risks in terms of privacy, misinformation, copyright infringement, and the potential generation of non-consensual sexual imagery due to training data sources. Clear disclosure and affirmative consent, particularly in sensitive fields like healthcare or legal services, are essential. While the U.S. lacks comprehensive data privacy laws, specific state regulations, like the Children's Online Privacy Protection Act (COPPA), may apply to some GenAI applications. Ongoing discussions and potential federal legislation are needed to address the privacy implications of GenAI and data use.

4. Generative Artificial Intelligence and Copyright Law, May 2023

Deciding copyright ownership for content generated by AI programs can be complex. The U.S. Copyright Office currently recognizes copyright in works created by human beings, but there is ongoing debate and litigation regarding whether AI-generated works can be eligible for copyright protection. While the issue stays unsettled, courts may consider factors such as human involvement, creative arrangements or modifications, and contractual terms to determine copyright ownership in AI-generated works.

5. National Artificial Intelligence Advisory Committee (NAIAC) Year 1 Report, May 2023

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President on AI's impact on various areas. This report compiles the committee's findings, including high-level themes, objectives,

actions, and a plan for future committee activities. The report is organized into four themes: (1) Leadership in Trustworthy Artificial Intelligence, (2) Leadership in Research and Development, (3) Support for the U.S. Workforce and Creation of Opportunities, and (4) International Cooperation. The report emphasizes that AI is a technology demanding immediate, substantial, and sustained attention from the government.

6. NIST AI Risk Management Framework (NIST AI RMF 1.0), January 2023

The NIST AI RMF aids users and developers in analyzing and mitigating AI risks with practical guidelines and best practices for their implementation. The framework has two parts: risk framing and core functions (governing, mapping, measuring, and managing). It provides seven risk and trust considerations for AI, particularly for GenAI tools, promoting responsible and reliable development and implementation. For additional details on the NIST AI RMF 1.0, refer to Appendix H.

7. Advancing American AI Act, December 2023

The Advancing American AI Act was originally introduced in April 2021 and eventually passed as part of the James M. Inhofe National Defense Authorization Act for Fiscal Year 2023. The purpose of this Act is to fulfill agency missions through the "use of innovative applied artificial intelligence technologies."[60] The bill requires agencies to take steps to promote AI usage while taking steps to mitigate technical and procedural risk and respect civil liberties.

8. AI Training for the Acquisition Workforce Act, October 2022

The AI Training for the Acquisition Workforce Act mandates the development and implementation of an AI training program for designated personnel in the federal government. The program aims to provide comprehensive knowledge about AI capabilities, associated risks, and mitigation strategies, with interactive learning components and regular updates to ensure effectiveness and alignment with the latest AI developments.

9. Blueprint for an AI Bill of Rights published by the Office of Science and Technology Policy (OSTP), October 2022

The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of Americans. These five principles are safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, and human alternatives, consideration, and fallback.

10. NIST Secure Software Development Framework (SSDF V1.1): Recommendations for Mitigating the Risk of Software Vulnerabilities

The NIST SSDF outlines a core set of high-level secure software development practices that can be integrated into SDLC implementations, which fills a gap in traditional SDLC resources which do not explicitly address software security.[61] The SSDF organizes recommended best practices into four groups: Prepare the Organization (PO), Protect the Software (PS), Produce Well-Secured Software (PW), and Respond to Vulnerabilities (RV). Each practice includes four sub-elements: Practice, Tasks, Notional Implementation Examples, and References.

11. AI Accountability Framework for Federal Agencies, published by GAO, June 2021

The AI Accountability Framework for Federal Agencies is made up of four principles (governance, data, performance, and monitoring). The framework captures key accountability practices centered around the four principles to help federal agencies use AI responsibly. This framework supplies a comprehensive set of guidelines and best practices to ensure transparency, fairness, and accountability in the development and deployment of AI systems while also considering privacy and security concerns.

12. National AI Initiative Act of 2020, enacted in January 2021

The National Artificial Intelligence Initiative Act of 2020 was enacted in January 2021 as part of the National Defense Authorization Act. It established the National AI Initiative Office via OSTP. The National Science and Technology Council was also charged with creating an interagency committee to coordinate federal programs and activities to support the initiative and an advisory committee for the President, and it requires a study on AI's impact on the U.S. workforce. The act also mandates additional AI research by the GAO for the NSF to provide grants for other AI research and voluntary AI standards by NIST.

13. Executive Order 13960: Promoting the Use of Trustworthy AI in the Federal Government, December 2020

EO 13960 emphasizes the potential of AI to enhance government operations and calls for its responsible and trustworthy implementation, setting guidelines for agencies. The order lists nine characteristics of trustworthy AI that should be considered throughout the AI lifecycle to ensure AI's trustworthy design, development, deployment, and operationalization.

Nine Principles of Trustworthy AI as enumerated in EO 13960

1. Lawful and respectful of our nation's values
2. Purposeful and performance-driven
3. Accurate, reliable, and effective
4. Safe, secure, and resilient
5. Understandable
6. Responsible and traceable
7. Regularly monitored
8. Transparent
9. Accountable

14. AI in Government Act, September 2020

The AI in Government Act of 2020 established the AI Center of Excellence (AI CoE) within the General Services Administration to facilitate AI adoption and improve government operations and competency. The act promotes collaboration between agencies, industry, nonprofits, and educational institutions for advancing AI adoption and includes guidance from the Director of the Office of Management and Budget to protect civil liberties and national security in AI technology use.

15. Executive Order 13859: Maintaining American Leadership in AI, February 2019

The executive order emphasizes the need for the advancement of AI across the federal Government, industry, and academia to harness its potential for Americans. The order recognizes the significance of GenAI and outlines a comprehensive strategy to promote its development, deployment, and regulation while safeguarding the national interest and values.

16. John S. McCain National Defense Authorization Act, Section 1051 for Fiscal Year 2019

The National Defense Authorization Act (NDAA) defines artificial intelligence to include each of the following:

- Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
- An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
- An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
- A set of techniques, including machine learning that is designed to approximate a cognitive task.

- An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.

17. E-Government Act of 2002

The E-Government Act of 2002 is a US Statute which seeks to "improve the management and promotion of electronic government services and processes by establishing a federal Chief Information Officer within the Office of Management and Budget, and by establishing a framework of measures that require using Internet-based information technology to improve citizen access to government information and services, and for other purposes."[62] It also requires the creation of Privacy Impact Assessments ("PIAs") for use in all federal agencies that create new technologies that manage identifiable information.

Appendix F. Reports on AI in Direct Relation to Research & Development (R&D) for Science

The following are selected references to reports, findings, and identification of needs in AI as they relate to research and development, which guide the use of AI systems by their unique sets of principles and considerations for the advancement of science.

1. "AI for Science, Energy, and Security (AI4SES) Report," Jonathan Carter, John Feddema, Doug Kothe, Rob Neely, Jason Pruet, and Rick Stevens, 2023
2. "National Artificial Intelligence Research & Development Strategic Plan," Select Committee on AI, National Science and Technology Council, 2023
3. "Artificial Intelligence for Isotopes: Report on the 2022 Workshop on Artificial Intelligence for Isotope R&D and Production," Kristian Myhre, Draguna Vrabie, Danda Rawat, and Ethan Balkin, 2023
4. "Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration ," National Academies of Sciences, Engineering, and Medicine, 2023
5. "AI@DOE Interim Executive Report," Ray Grout, Kelly Rose, Valerie Taylor, and Brian Essen, 2022
6. "Artificial Intelligence for Earth System Predictability (AI4ESP)," Nicki Hickmon, et al, 2022
7. "Management and Storage of Scientific Data (for AI/ML Workflows)," Suren Byna, et al, 2022
8. "Artificial Intelligence and Machine Learning for Bioenergy Research Opportunities and Challenges," Huimin Zhao, Nathan Hillson, Kerstin Kleese van Dam, Deepti Tanjore, et al, 2022 ()
9. "AI for Science Report," Rick Stevens, Valerie Taylor, Jeff Nichols, Arthur Maccabe, Katherine Yelick, and David Brown, 2020
10. "Final Draft of Report to the Committee by the Subcommittee on AI/ML, Data-intensive Science and High-Performance Computing," DOE Advanced Scientific Computing Advisory Committee (ASCAC), 2020
11. "Opportunities and Challenges from Artificial Intelligence and Machine Learning for the Advancement of Science, Technology, and the Office of Science Missions," Tony Hey, et al, 2020
12. "Application of Artificial Intelligence and Machine Learning for the Operation of NP Accelerator Facilities: NP Roundtable Meeting," DOE Office of Science, Nuclear Physics, 2020
13. "Next-Gen AI for Proliferation Detection: Accelerating the Development and Use of Explainability Methods to Design AI Systems Suitable for Nonproliferation Mission Applications: Workshop Report," Francis Alexander, et al, 2021
14. "AI for Nuclear Physics: DOE Office of Science Workshop," Tanja Horn, et al, 2020
15. "AI for Nuclear Physics: DOE Office of Science Workshop Report," Paulo Bedaque, et al, 2020
16. "Advancing Fusion with Machine Learning: Office of Science Report," DOE Fusion Energy Sciences and Advanced Scientific Computing Research, 2019
17. "Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning: DOE Roundtable Report," Daniel Ratner and Bobby Sumpter, 2019

18. "Data and Models - A Framework for Advancing AI in Science: DOE Roundtable Report," Kjiersten Fagnan, et al, 2019
19. "Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence: DOE Workshop Report," Nathan Baker et al, 2019
20. "Machine Learning and Understanding for Intelligent Extreme Scale Scientific Computing and Discovery: DOE Workshop Report," Michael Berry, Thomas Potok, Prasanna Balaprakash, Hank Hoffman, Raju Vatsavai, and Prabhat, 2015

Appendix G. The AI Lifecycle

The Artificial Intelligence Lifecycle includes three phases: **Pre-Design, Design and Development,** and **Deployment.** See Figure 7 for a description of the steps contained in each phase and their associated risks.

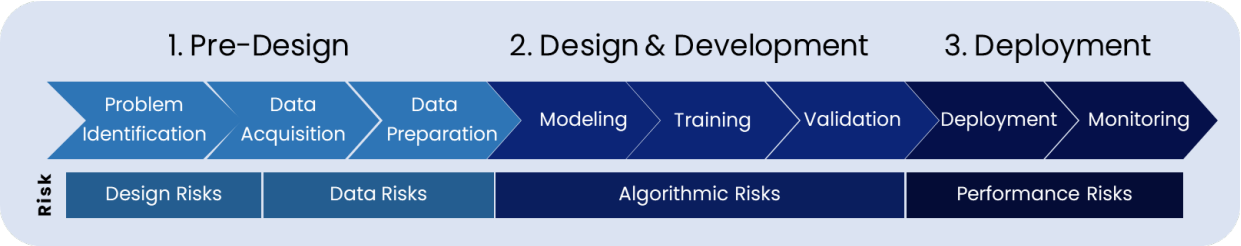


Figure 7: The AI Lifecycle and Associated Risks. Adapted from: The Department of Energy Artificial Intelligence Resource

Each of the sub-phases contains critical actions needed to achieve the desired outcome when preparing for, developing, and deploying a new AI technology. Recent publications, such as EO 14110, place additional focus on Phase Two: Design and Development.[63] Included in this phase are activities related to testing an AI model. AI testing involves using appropriate tools, frameworks, and methodologies to independently validate the models and their integration with existing upstream and downstream systems and processes. The objective of completing robust AI testing is to ensure that quality, security, and compliance standards are met, and that the solution is scalable and reliable. Testing AI is an iterative cycle in itself and requires consistent monitoring, feedback, and improvement. Testing activities should cover defects in functionality, user interactions with the tool, and the user experience, among other tool-specific requirements. Additionally, testing may help identify within a model the GenAI-specific risks discussed in Section 8.

Per EO 14110, specific guidance on the development of GenAI technologies is being written to supplement the existing NIST Secure Software Development Framework (SSDF). Additionally, red-teaming in collaboration with NIST, which entails "a group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture", is expected to play a key role in testing and monitoring GenAI capabilities throughout Phases Two and Three, as discussed in EO 14110 Section 10.1 (b)(vii)(A).[64] Including red-teaming in the AI development lifecycle can improve the security of the model by demonstrating both the impacts of successful attacks and potential responsive actions or remediations which are effective against such attacks.

Each phase possesses a unique set of risks which should be addressed or mitigated during that portion of the lifecycle. For additional information on mitigating risk throughout the AI lifecycle, see both Section 8.3: AI Risk Management and the AI Risk Management Playbook (AIRMP). The AIRMP is a set of 141 risks related to the development and deployment of AI, along with corresponding mitigations which can be proactively incorporated throughout the AI lifecycle to address design, data, algorithmic, and performance risk.

Appendix H. The NIST Artificial Intelligence Risk Management Framework (NIST AI RMF Version 1.0)

In addition to providing valuable information on the principles of trustworthy and responsible AI and how effective risk management can enable those principles, the NIST AI RMF includes a voluntary framework for managing AI risks, comprised of 4 functions: Govern, Map, Measure, and Manage.

Figure 8: The Four Functions of the NIST AI RMF

The **Govern** function facilitates the rest of the functions by cultivating and institutionalizing a culture of risk management across the organization. This includes developing a governance structure and corresponding processes and documentation to assess, identify, and manage AI risks. The Govern function also connects AI system development to organizational principles, risk appetites (the organization's tolerance or willingness to accept risk), and strategies.

The **Map** function creates proper context for framing risks related to an AI system. Information obtained through the completion of the Map function enables selection and deployment of appropriate AI use cases and prevents or mitigates risks. Obtaining organizational context allows for the fullest understanding of AI risks and contributing or underlying factors. The outcomes of the Map function also facilitate the Measure and Manage functions.

The **Measure** function utilizes a variety of tools and methods to "analyze, assess, benchmark, and monitor AI risk and related impacts," including documenting a system's functionality and potential trustworthiness. The Measure function relies on data from the Map function and informs the Manage function.

The **Manage** function facilitates the allocation of resources to risks as defined by the Govern function via a risk treatment plan. Actions may include response, recovery, and communication in the event of a technical or security incident. Information from negative impacts can help inform future resource decisions.

For additional information on specific actions for each function, see the full online publication of the NIST AI RMF, or for information on general AI risk management, refer to Section 8.3.

Appendix I. GenAI Prompt Engineering

Understanding how to properly craft prompts (the technical term for a request, command, or question) when interacting a GenAI model is critical to achieving a reliable and accurate output.

The first principle of crafting AI prompts is to write clear and specific instructions. Writing a clear prompt and a short prompt are not synonymous, often longer prompts are more specific and lead to better outputs. Using delimiters such as quotes or backticks to clarify for the model which text to take as input (for summarizing, expanding on, or transforming in some way) versus what text is conveying the actions

for the model to perform (summarize, expand, etc.) can ensure a better response and prevent prompt injections which could alter the purpose of the model.

Another tactic for providing clear and specific instructions is to request a standardized response from the AI in the prompt (for example, asking for an HTML or JSON response to the query). Finally, it can be helpful to include in the prompt a request for the AI to check that conditions for an answer are satisfied, or if they are not to check assumptions instead of running a full query. For example, if feeding the model text from which to extract step by step instructions, give the model a second response option if no content for creating steps is detected.

The second principle for crafting AI prompts is to give the model time to think, or to devote more computational power to solving a problem to prevent errors. One tactic for implementing this principle is to ask the model to carry out the query in steps to ensure each portion of the task is completed fully. This tactic aligns well with the tactic on providing a standardized output. Another tactic to give the model time to think is to ask it to calculate its own solution to a problem before checking the correctness of a provided solution.

Prompts are most effectively developed by incorporating these principles into an iterative development cycle. The first step in the cycle is to come up with the idea for what you would like to ask the model to do. Next comes an initial implementation, or a first pass of setting up the environment and writing a query. Running this query will provide an experimental result which will be used to analyze what errors the model is making or what information to give in the prompt to better hone the output for its intended purpose. This error analysis is then used to create the idea for the updated prompt, and the cycle repeats.

This content is a summarization of an OpenAI course on prompt engineering and Generative AI functionalities.[65] The full course and related materials can be found here. Additional information from OpenAI on prompt engineering can be found here.

Appendix J. Examples of Protected Data

Inherently sensitive data types	Description
Classified information	Classified information is defined by DOE as “certain information requiring protection against unauthorized disclosure in the interests of national defense and security or foreign relations of the United States pursuant to Federal statute or Executive order.” It includes Restricted Data, Formerly Restricted Data, and National Security Information. The potential damage to the national security of each is denoted by the classification levels Top Secret, Secret, and Confidential. **Source:** Classified Information (DOE)
Company proprietary information	NIST defines proprietary information as “material and information relating to or associated with a company’s products, business, or activities, including but not limited to financial information, data or statements, trade secrets, product research and development, existing and future product designs and performance specifications, marketing plans or techniques, schematics, client lists, computer programs, processes, and know-how that has been clearly identified and properly marked by the company as proprietary information, trade secrets, or company confidential information. The information must have been developed by the company and not available to the Government or to the public without restriction from another source.” **Source**: Proprietary Information Definition (NIST)
Controlled unclassified information (CUI)	CUI is defined by the DOE as “information the Government creates or possesses, or that an entity creates or possesses for or on behalf of the Government, that a Law, Regulation, or Government-Wide Policy (LRGWP)

Inherently sensitive data types	Description
	requires or permits an agency to handle using safeguarding or dissemination controls." This includes sensitive information, personally identifiable information (PII), confidential information and private information. **Source:** Controlled Unclassified Information (doe.gov)
Confidential information	*For purposes of this Guide*, confidential information can be generally defined as information or data of a personal nature which is proprietary about an individual, or information or data pertaining to or submitted by an organization. However, confidential information refers to any data or knowledge that is shared with an individual or organization under the condition that it remain private and undisclosed. It includes all information that is designated as confidential or which, by its nature is or ought to be considered as confidential (whether it is so marked) and includes all personal data (PII), and proprietary information that relate to the intellectual property, data, know-how, trade secrets, business affairs, developments, personnel and suppliers of the entity to which the information belongs. Confidential information exists in all forms: written, spoken, observed, electronic, or otherwise. If there is any uncertainty or legal question related to the confidentiality of information or data, counsel should be sought from cognizant DOE or contractor legal counsel before use. **Example Source**: 452.224-70 Confidentiality of Information. | Acquisition.GOV
Cooperative Research and Development Agreement (CRADA)-protected information	DOE defines protected CRADA information as "generated information which is marked as being Protected CRADA Information by a Party to this CRADA and which would have been Proprietary Information had it been obtained from a non-federal entity." This type of information is protected for five years by law, and thus is protected for a five to 30year period. **Source:** DOE O 483.1B Cooperative Research and Development Agreements
Intellectual property (IP)	Intellectual Property includes trade secrets, patents, copyrights, trademarks, industrial designs, and geographical indications. **Source**: World Intellectual Property Organization
Limited rights data	Data in which the U.S. government has no inherent rights. Limited rights data is typically developed at private expense not under an award and is typically usable only for the purposes of a particular award. Limited rights data must be properly marked the recipient and its delivery should be minimized. To use or share limited rights data outside the U.S. government, written permission must be obtained from the data owner.
Personally identifiable information (PII)	Personally identifiable information (PII) is defined by Office of Management and Budget (OMB) Circular No. A–130 (and defined in Executive Order 14110 using this source) as "information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual". **Source**: Office of Management and Budget (OMB) Circular No. A–130 Personally identifiable information (PII) is defined by DOE as: "Information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual. PII can include unique individual identifiers or combinations of identifiers, such as an individual's name, Social Security number, date and place of birth, mother's maiden name, biometric data, etc.

Inherently sensitive data types	Description
	The sensitivity of PII increases when combinations of elements increase the ability to identify or target a specific individual. PII, which if lost, compromised, or disclosed without authorization, could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual is categorized as High Risk PII. Examples of High Risk PII include, Social Security Numbers (SSNs), biometric records (e.g., fingerprints, DNA, etc.), health and medical information, financial information (e.g., credit card numbers, credit reports, bank account numbers, etc.), and security information (e.g., security clearance information). While all PII must be handled and protected appropriately, High Risk PII must be given greater protection and consideration following a breach because of the increased risk of harm to an individual if it is misused or compromised." **Source**: DOE O 206.1, Department of Energy Privacy Program, January 1, 2009
Private information	Most definitions of "private information" exist at the state, local, or organizational level and contain similarities to federal definitions of PII confidential information, meaning in general, any information concerning a natural person which, because of an identifier, can be used to identify such natural person if it is in combination with any one or more such or specific data elements such as identified in the definition of PII. If "private information" is a specific type of information within your organization or office, please refer to that definition and to any associated requirements surrounding its confidentiality, integrity, and availability. If there is any uncertainty or legal question about whether information is private information, counsel should be sought from cognizant DOE or contractor legal counsel before use.
Acquisition data (pricing or cost data)	*Cost or pricing data* (10 U.S.C. 3701(1) and 41 U.S.C. chapter 35) means all facts that, as of the date of price agreement, or, if applicable, an earlier date agreed upon between the parties that is as close as practicable to the date of agreement on price, prudent buyers and sellers would reasonably expect to affect price negotiations significantly. Cost or pricing data are factual, not judgmental; and are verifiable. While they do not indicate the accuracy of the prospective contractor's judgment about estimated future costs or projections, they do include the data forming the basis for that judgment. Cost or pricing data are more than historical accounting data; they are all the facts that can be reasonably expected to contribute to the soundness of estimates of future costs and to the validity of determinations of costs already incurred. They also include, but are not limited to, such factors as- (1) Vendor quotations; (2) Nonrecurring costs; (3) Information on changes in production methods and in production or purchasing volume; (4) Data supporting projections of business prospects and objectives and related operations costs; (5) Unit-cost trends such as those associated with labor efficiency; (6) Make-or-buy decisions; (7) Estimated resources to attain business goals; and (8) Information on management decisions that could have a significant bearing on costs. *Data other than certified cost or pricing data* means pricing data, cost data, and judgmental information necessary for the contracting officer to determine

Inherently sensitive data types	Description
	a fair and reasonable price or to determine cost realism. Such data may include the identical types of data as certified cost or pricing data, consistent with Table 15-2 of 15.408, but without the certification. The data may also include, for example, sales data and any information reasonably required to explain the offeror's estimating process, including, but not limited to– (1) The judgmental factors applied and the mathematical or other methods used in the estimate, including those used in projecting from known data; and (2) The nature and amount of any contingencies included in the proposed price. **Source**: Part 2 - Definitions of Words and Terms | Acquisition.GOV
Protected health information (PHI)	PHI is defined by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule as "individually identifiable health information that is transmitted or maintained in any form or medium (electronic, oral, or paper) by a covered entity or its business associates, excluding certain educational and employment records." **Source**: Privacy Rule and Research (NIH)
Sensitive information	DOE 2019 Operations Security (OPSEC) Handbook, Section 3.1.5 "Public Release Review" discusses two primary characteristics of a piece of information that determine whether that information is safe for public disclosure. These two characteristics are: sensitivity and risk. Together, these two terms suggest that the term "sensitive information" is a categorical term, which then includes other specific types of sensitive information. The Handbook comes close to a definition for sensitive information in Section 3.1.5 by presenting these primary characteristics for "determining suitability for release" of information: ► Sensitivity: "If the information is released to the public, it should not reveal or identify sensitive information, activities, or programs." ► Risk: "Information that may be used by adversaries to the detriment of employees, the public, the department, or the nation should not be approved for release. This determination should be based on sound risk management principles focused on preventing potential adverse consequences." **Source**: DOE Operations and Security (OPSEC) Handbook, 2019

Appendix K. Glossary

There are a variety of terms relevant to the discussion surrounding AI and GenAI. The definitions listed below are a selection of terms which may be helpful when performing AI research, design, discussion, operation, or development. All definitions are sourced from Section 3 of Executive Order 14110 on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* unless marked * and cited.

Term	Definition
Artificial intelligence (AI)	A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

Term	Definition
Generative artificial intelligence (Generative AI or GenAI)	The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.
AI model	A component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.
AI red-teaming	A structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.
AI system	Any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI.
Critical and emerging technologies	The technologies listed in the February 2022 Critical and Emerging Technologies List Update issued by the National Science and Technology Council (NSTC), as amended by subsequent updates to the list issued by the NSTC.
Deep learning (DL)*	Deep learning is a subset of machine learning, and which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing them to "learn" from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.[66]
Deep neural network (DNN)*	Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization. This progression of computations through the network is called forward propagation. The input and output layers of a deep neural network are called visible layers. The input layer is where the deep learning model ingests the data for processing, and the output layer is where the final prediction or classification is made.[67]
Dual-use foundation model	An AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by: ► substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons ► enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks ► permitting the evasion of human control or oversight through means of deception or obfuscation

Term	Definition
	Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.
GPT*	GPT is a family of LLMs built on deep neural network (DNN) architecture that have been fine-tuned using natural language processing (NLP) and reinforcement learning from human feedback (RLHF) techniques.
Large language model (LLM)*	Large language models (LLMs) take advantage of self-supervised learning and can learn from large amounts of unstructured and unlabeled text data. These models are trained on large bodies of data, allowing for one model to be used for multiple use cases.
Machine learning	A set of techniques that can be used to train AI algorithms to improve performance at a task based on data.
Natural language processing (NLP)*	Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to 'understand' text and spoken words in much the same way human beings can. NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker's or writer's intent and sentiment.[68]
Reinforcement learning with human feedback (RLHF)*	Reinforcement learning from human feedback (RLHF) is a machine learning (ML) technique that uses human feedback to optimize ML models to self-learn more efficiently. Reinforcement learning (RL) techniques train software to make decisions that maximize rewards, making their outcomes more accurate. RLHF incorporates human feedback in the rewards function, so the ML model can perform tasks more aligned with human goals, wants, and needs.[69]
Synthetic content	Information, such as images, videos, audio clips, and text, that has been significantly modified or generated by algorithms, including by AI.
Task-specific model*	All of the AI in place today is task-specific, or narrow AI. This is an important distinction as the general ability to reason, think, and perceive is known as Artificial General Intelligence (AGI) which, at this point, is not technically possible.[70]
Testbed	A facility or mechanism equipped for conducting rigorous, transparent, and replicable testing of tools and technologies, including AI and PETs, to help evaluate the functionality, usability, and performance of those tools or technologies.
Transformer*	Deep learning transformers are a type of AI that are used to learn data representations in an automated manner. Transformers are designed to handle sequential data, such as natural language, making them well-suited for tasks such as text classification, machine translation, and question answering.[71]
Watermarking	The act of embedding information, which is typically difficult to remove, into outputs created by AI—including into outputs such as photos, videos, audio clips, or text—for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.

[1] DOE Generative AI Reference Guide v1, published September 2023
[2] "DOE Prepping Version 2 of GenAI Responsible Use Guide," MeriTalk, December 7, 2023
[3] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 3 (b)
[4] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 3 (t)
[5] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 3 (p)
[6] "What is deep learning?" IBM
[7] Gartner, "Glossary of Terms for Generative AI and Large Language Models," Anthony Mullen, July 2023. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.
[8] "NukeLM: Pre-Trained and Fine-Tuned Language Models for the Nuclear and Energy Domains," Cornell University, May 25, 2021
[9] Microsoft-backed OpenAI files trademark for ChatGPT powered by GPT-5
[10] Gartner, "Predicts 2024: The Future of Generative AI Technologies," Arun Chandrasekaran, Anthony Mullen, Lizzy Foo Kune, Nicole Greene, Jim Hare, Leinar Ramos, Anushree Verma, February 28, 2024
[11] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 10.1(e)
[12] "Notice to research community: Use of generative artificial intelligence technology in the NSF merit review process", NSF - National Science Foundation, December 14, 2023
[13] OCIO Enterprise Security Assessment and Authorization Process, August 2023
[14] NIST AI RMF 1.0 pg. 6
[15] NIST AI RMF 1.0 pg. 8
[16] NIST AI RMF 1.0 pg. 1
[17] ISO/IEC TS 5723:2022, cited in the NIST AI RMF 1.0
[18] DOE EO 13960 Consistency Plan
[19] "How to Red Team a Gen AI Model," Harvard Business Review, Andrew Burt, January 4, 2024
[20] CFR Title 10, Chp X, Part 1008 DOE Records Maintained on Individuals (Privacy Act)
[21] "Data Privacy and Information Protection Principles for the City of Portland," City of Portland, adopted June 19, 2019
[22] "Data Privacy and Information Protection Principles for the City of Portland," City of Portland, adopted June 19, 2019
[23] "TechBriefing: ChatGPT, LLMs, and Generative AI," Gartner, published June 20, 2023
[24] OMB Circular No. A–130 Managing Information as a Strategic Resource, Office of Management and Budget, July 2016
[25] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 3 (j)
[26] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 3 (z)
[27] NIST AI RMF 1.0 pg. 17
[28] Confidentiality, NIST glossary
[29] DOE Operations and Security (OPSEC) Handbook, Section 3.1.5, June 2019
[30] Thaler v. Vidal Federal Court Decision See Also: Guidance on AI from the U.S. Patent and Trademark Office (USPTO)
[31] Federal Register: Artificial Intelligence and Copyright
[32] Federal Register / Vol. 89, No. 30 / Tuesday, February 13, 2024 / Notices
[33] Federal Register :: Artificial Intelligence and Copyright
[34] ISO/IEC TS 5723:2022 referenced in NIST AI RMF 1.0
[35] NIST AI RMF 1.0 pg. 14
[36] NIST AI RMF 1.0 pg. 15
[37] NIST AI RMF 1.0 pg. 17
[38] Fairness and Bias in Artificial Intelligence: a Brief Survey of Sources, Impacts, and Mitigation Strategies, Emilio Ferrara, Cornell University, April 16, 2023
[39] NIST AI RMF 1.0 pg. 18

[40] Fairness and Bias in Artificial Intelligence: a Brief Survey of Sources, Impacts, and Mitigation Strategies, Emilio Ferrara, Cornell University, April 16, 2023
[41] ISO 9000:2015, ISO 9000:2015(en), Quality management systems — Fundamentals and vocabulary
[42] ISO/IEC TS 5723:2022
[43] NIST AI RMF 1.0 pg. 16
[44] A Grounded Approach: Overcoming AI Hallucinations," John Bohannon, July 27, 2023
[45] A Grounded Approach: Overcoming AI Hallucinations," John Bohannon, July 27, 2023
[46] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 10.1 (f)(iii)
[47] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 4.1 (b)
[48] DOE EO 13960 Consistency Plan
[49] Gartner, "How to Pilot Generative AI," Leinar Ramos, Anthony Mullen, Rajesh Kandaswamy, Radu Miclaus, Erick Brethenoux, Avivah Litan, Haritha Khandabattu, July 10, 2023
[50] DOE EO 13960 Consistency Plan
[51] DOE EO 13960 Consistency Plan
[52] DOE EO 13960 Consistency Plan
[53] DOE EO 13960 Consistency Plan
[54] "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," NIST, January 2024
[55] DOE EO 13960 Consistency Plan
[56] "A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications," Cambridge University, November 12, 2021
[57] What is synthetic data and how can it help you competitively, MIT Sloan School of Management, January 23, 2023
[58] "Metrics, Logs, and Traces: The Golden Triangle of Observability in Monitoring," DevOps, November 8, 2018
[59] "How to Red Team a Gen AI Model," Harvard Business Review, Andrew Burt, January 4, 2024
[60] S.1353 - 117th Congress (2021-2022): Advancing American AI Act, Library of Congress, December 2023
[61] NIST SP 800-218, Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities, February 2022
[62] E-Gov Act of 2002, Digital.gov, Decmeber 2002
[63] Executive Order (EO 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, Section 2 (a)
[64] Red Team - Glossary, National Institute of Standards and Technology: CRSC
[65] ChatGPT Prompt Engineering for Developers - DeepLearning.AI
[66] What is Deep Learning? | IBM
[67] What is Deep Learning? | IBM
[68] What is Natural Language Processing? | IBM
[69] What is RLHF? - Reinforcement Learning from Human Feedback Explained, Amazon Web Services
[70] Key AI terminology | GSA - IT Modernization Centers of Excellence
[71] What are transformers deep learning? - Google LaMDA