

Generative AI and WMD Nonproliferation: A Practical Primer for Policymakers and Diplomats

Natasha E. Bajema, PhD

CNS

OCCASIONAL PAPER

#63 · DECEMBER 2024



Middlebury Institute of
International Studies at Monterey

James Martin Center for Nonproliferation Studies

Generative AI and WMD Nonproliferation: A Practical Primer for Policymakers and Diplomats

Natasha E. Bajema, PhD

**James Martin Center of Nonproliferation Studies
Middlebury Institute of International Studies at Monterey**

460 Pierce Street, Monterey, CA 93940, USA

Phone: +1 (831) 647-4154

Fax: +1 (831) 647-3519

www.nonproliferation.org

www.middlebury.edu/institute

Acknowledgments: The author would like to thank the Carnegie Corporation of New York for their generous support. The author is also grateful to Dr. Ferenc Dalnoki-Veress for his helpful feedback and Dr. Stephen Herzog for his significant contribution in providing edits and constructive advice.

The views, judgments, and conclusions in this report are the sole representations of the authors and do not necessarily represent either the official position or policy or bear the endorsement CNS or the Middlebury Institute of International Studies at Monterey.

Cover image: Deposit Photos

©2024, The President and Trustees of Middlebury College

Contents

EXECUTIVE SUMMARY	1
INTRODUCTION	3
CHAPTER 1: ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND GENERATIVE AI	7
ARTIFICIAL INTELLIGENCE	7
MACHINE LEARNING	8
DEEP NEURAL NETWORKS	11
PREDICTIVE VERSUS GENERATIVE	13
Predictive AI	13
Generative AI	16
Narrow Artificial Intelligence Versus Artificial General Intelligence (AGI)	20
CHAPTER 2: THE GENERATIVE AI LANDSCAPE	23
TRAINING TECHNIQUES	24
Supervised Learning	24
Unsupervised Learning	24
Reinforcement Learning	25
Reinforcement Learning from Human Feedback (RLHF)	25
TYPES OF GENERATIVE AI MODELS	26
Generative Adversarial Networks (GANs)	26
Large Language Models (LLMs)	27
Diffusion Models	28
Multi-Modal Models	28
World Models	30
CLOSED VERSUS OPEN-SOURCE MODELS	31
Performance Metrics	32
Benchmarks and Evaluations	33

Contents continued...

CHAPTER 3: GENERATIVE AI APPLICATIONS	35
IMPROVING PRODUCTIVITY WORKFLOWS THROUGH AI ASSISTANTS (CHATBOTS)	36
UPGRADING SEARCH ENGINES WITH GENERATIVE AI	37
CUSTOMIZING GENERATIVE AI MODELS FOR SPECIFIC DOMAINS	39
Prompt Engineering	39
Retrieval Augmented Generation (RAG)	40
Fine-Tuning	41
IMPROVING AND AUTOMATING WORKFLOWS WITH AI AGENTS (AGENTIC AI)	42
Plugins and Agent Actions	43
Agent Frameworks and Multi-Agent Collaboration	44
CHAPTER 4: FLAWS, RISKS, AND LIMITATIONS ON THE GROWTH OF AI	47
FUNDAMENTAL DESIGN FLAWS OF GENERATIVE AI	47
Hallucinations	47
Data Bias	48
Copyright and Intellectual Property	49
Complexity	49
Explainability	50
Cyber Vulnerabilities	51
OTHER RISKS AND VULNERABILITIES	52
Data Privacy Issues	52
Disinformation	53
Misuse for Malicious Purposes	54
Lack of Human Alignment	54
LIMITATIONS ON THE GROWTH OF GENERATIVE AI	55
Data Shortages	55

Contents continued...

- Energy Resources 56
- Economic Return 57

- CHAPTER 5: REGULATORY FRAMEWORK AND MITIGATION MEASURES 59**
- U.S. REGULATORY FRAMEWORK FOR AI 60**
- AI Red Teaming 62
- AI Benchmarks and Safety Evaluations 63
- EUROPEAN UNION’S REGULATORY FRAMEWORK FOR AI 64**
- GLOBAL GOVERNANCE OF AI 64**

- APPENDIX A: The Fundamentals of Using Generative Ai Models, Prompt Engineering,
and Productivity Use-Cases 67**
- THE BASICS 67**
- Poe AI 67
- ChatGPT 68
- Claude 68
- Gemini 68
- Flux 68
- Midjourney 68
- PROMPT ENGINEERING 68**
- Zero-Shot Prompt 69
- One-Shot and Few-Shot Prompts 70
- Many-Shot Prompts 71
- Chain-of-Thought / Tree-of-Thought 71
- Prompt Frameworks 72

Contents continued...

PRODUCTIVITY USE-CASES	74
Uploading Documents	74
Customized GPTs	74
NotebookLM	74
Watsonx	75
WMD NONPROLIFERATION – AI TOOLS USE-CASE WORKSHEET	77
END NOTES	81

Executive Summary

This primer provides a comprehensive overview of generative artificial intelligence (AI) and its implications for weapons of mass destruction (WMD) nonproliferation. It addresses five key areas, beginning with fundamental AI concepts that explain the evolution from traditional AI to current generative AI systems. The primer distinguishes between predictive and generative AI models, emphasizing how the newest AI models, particularly large language models (LLMs), differ from previous technologies in their ability to generate novel content rather than simply making predictions.

The document provides a detailed analysis of various generative AI architectures, including LLMs, diffusion models, and emerging world models. It outlines different training techniques (supervised, unsupervised, reinforcement learning) and explains how these systems are developed and improved through methods like fine-tuning and retrieval augmented generation (RAG). The primer then explores current applications of generative AI, from basic chatbot interactions to sophisticated agentic AI systems capable of autonomous action. It details how organizations can customize AI models for specific domains and discusses the emergence of AI-enhanced search engines and workflow automation.

Several critical challenges are identified, including design flaws (hallucinations, data biases, and copyright issues), implementation risks (data privacy, disinformation, and malicious use potential), and growth limitations (data shortages, energy constraints, and uncertain economic returns). The primer pays particular attention to specific WMD-related concerns, such as the potential for misuse in weapons development and proliferation. The document also outlines current and emerging regulatory approaches, including U.S. initiatives like Executive Order 14110, the European Union's AI Act, global governance efforts through international organizations, and specific measures for addressing WMD-related risks.

The primer concludes with practical guidance for policymakers and diplomats, including detailed instructions for using AI tools and frameworks for evaluating their potential benefits and risks in the nonproliferation domain.

Keywords:

Artificial Intelligence (AI), Weapons of Mass Destruction (WMD), Large Language Models (LLMs), Generative AI, Nonproliferation, AI Safety, Machine Learning, Deep Neural Networks, AI Regulation, Prompt Engineering, AI Agents, Red Teaming, Data Privacy, Cyber Vulnerabilities, AI Governance, Risk Assessment, AI Benchmarking, AI Ethics, National Security, Technological Innovation, AI Policy, International Security, Dual-use Technology, AI Alignment, Emerging Technologies

Introduction

The emergence of artificial intelligence (AI) presents significant opportunities and challenges for WMD nonproliferation. On the one hand, AI introduces new risks, as state and non-state actors could employ these new technologies to enable weapons development and use.¹ AI could also introduce other potentially existential risks, some of which we may not have imagined previously.² On the other hand, AI has the potential to be a powerful tool for detecting and analyzing proliferation risks, supporting arms control verification and treaty negotiation, and gaining new insights into the decision calculus of proliferators.³ Yet, even the benefits of AI come with major risks when applied to weapons of mass destruction (WMD) nonproliferation. To harness these benefits, policymakers must also contend with new risks, including cyber vulnerabilities that are relatively new to WMD nonproliferation but inherent to AI systems that rely upon software, hardware, and penetrable networks.⁴

The breathtaking pace of progress in the development of AI demands urgent action from the WMD nonproliferation community to get ahead of the curve, and for policymakers and diplomats to develop AI literacy. As AI capabilities rapidly mature in the coming years, potentially enabling new existential threats, we confront a closing window to steer major outcomes toward the positive effects of these technologies. We must, therefore, work together to help develop and leverage AI to serve nonproliferation goals while anticipating and formulating responses to potential misuse.

Policymakers and diplomats worldwide have already started to make the critical connection between the complex, abstract, fast-paced, and dynamic landscape of AI and the threats presented by WMD. However, there is often a significant disconnect between AI's perceived risks and benefits, alongside confusion about what AI is, and what it can and cannot do. This knowledge gap obscures the specific risks and opportunities relevant to the WMD domain and prevents the development of practical solutions to mitigate new risks.

The James Martin Center for Nonproliferation Studies (CNS) has initiated an ambitious research and training agenda on AI explicitly tailored for the field of WMD nonproliferation. CNS is examining the nexus of AI and WMD nonproliferation from several perspectives. This research is informed by a core group of staff with deep expertise on relevant issues. As a major part of its mission, CNS seeks to educate and support next-generation experts, diplomats, and policymakers engaged with complex topics, including the implications of generative AI.

Ever since the release of the OpenAI ChatGPT platform in 2022, much buzz has focused on the risks and benefits of generative AI models. Such models are a specific category of AI tools built on deep neural networks that allow for human-like interactions. As part of our education and training efforts, CNS has produced this primer on generative AI for policymakers and diplomats to provide meaningful context and frame the potential risks and benefits of these developments. We aim to provide readers with a deeper understanding of these revolutionary tools.

This primer will help cut through the noise and provide necessary clarity about generative AI, its specific relevance within the broader AI context, and its implications for the future. It also

will provide policymakers and diplomats with the foundational knowledge to understand the potential risks and benefits of generative AI for WMD nonproliferation. Finally, it serves as a practical guide to begin using generative AI within your organization’s daily work. Whether you wish to better understand how the tools work, or hope to enhance your productivity using AI tools, this primer should be of use to you and your organization.

This guide is organized into five chapters and also contains an appendix providing suggestions about how to apply your new knowledge of generative AI:

- Chapter 1 places the development of generative AI into the broader context of advancements in machine learning and explains how these new tools differ from other approaches to AI.
- Chapter 2 provides a brief overview of generative AI and describes the different types of tools that fall within this category.
- Chapter 3 discusses the enormous variety of applications built around generative AI to extend the current capabilities of the foundation models. It also examines the near-term advancement of the agentic AI revolution.
- Chapter 4 examines the flaws, risks, and limitations of generative AI and explores the prospects for future advancement.
- Chapter 5 provides a brief overview of current efforts to develop mitigation measures for reducing the risks of AI systems.
- Appendix A provides a practical guide for using AI models to improve productivity and offers tips for prompt engineering

The guide uses the following symbols to highlight important issues at the nexus of AI and WMD.



Key insight for the WMD domain⁵



Use-case for the WMD domain⁶



Benefit or opportunity for the WMD domain⁷



Risk for the WMD domain or other downside of AI tools⁸

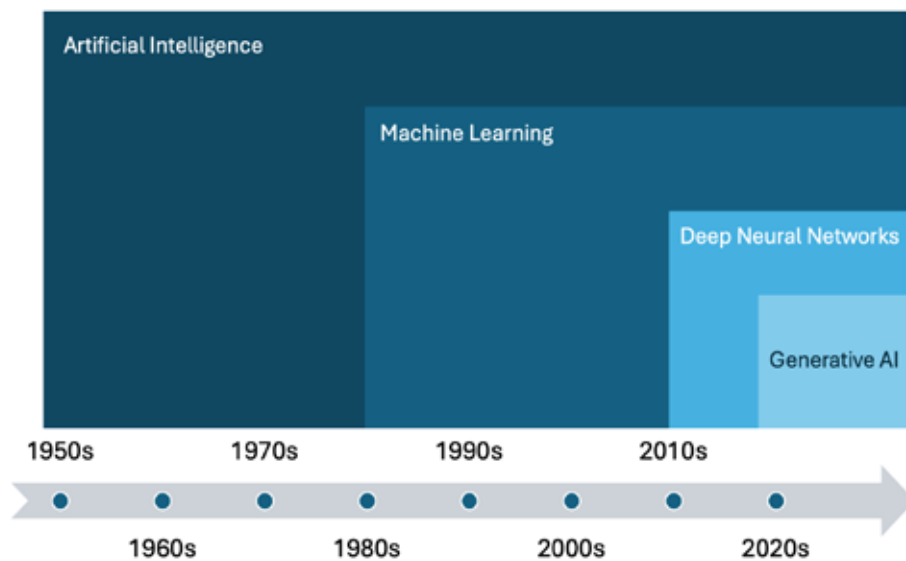


Key governance issue for the WMD domain⁹

Chapter 1: Artificial Intelligence, Machine Learning, and Generative AI

Neither artificial intelligence nor machine learning are new fields of scientific and technological inquiry within the broader discipline of computer science.¹⁰ Both emerged in the 1950s alongside the development of computers.¹¹ In the decades since then, AI has gone through several cycles of hype and disappointment known as “AI winters.”¹² With the recent rise of generative AI, it remains unclear whether AI will face another period of setbacks or if the current trendline of rapid advancement will continue onward without abatement.

Figure 1: Development of Artificial Intelligence



A significant challenge in understanding and assessing the nexus between AI and WMD relates to the complexity of the field of AI itself. This field is currently dominated by machine learning approaches and now consists of many different types of tools. These tools are built on varying architectures and designed to achieve diverse outcomes. Within the AI subfield of machine learning, several branches of tools are currently being developed,¹³ divergent schools of thought on algorithms and architectures exist,¹⁴ and numerous training techniques are used.¹⁵

This chapter will place generative AI within a broader context, allowing you to distinguish between it and other types of AI tools. In doing so, this primer endeavors to deepen your understanding of the nexus of AI and WMD nonproliferation.

ARTIFICIAL INTELLIGENCE

Artificial intelligence, first coined by John McCarthy at Dartmouth College in 1956, refers to “the science and engineering of making intelligent machines and software, especially intelligent computer programs.”¹⁶ For many decades, computers have been programmed to perform complex tasks, previously done by humans, using AI. Over time, computers have accomplished

a growing number of tasks both faster and more accurately than humans, using various forms of AI embedded in operating systems and software.¹⁷



AI is Not New

Until recently, a common misperception persisted in the WMD nonproliferation field about the risks of AI. Many argued that since it was a digital technology, it would have minimal intersection with the physical aspects of a WMD development program—particularly in the nuclear domain. However, computer programmers have been making computers more intelligent since the 1950s with the latest AI techniques. In other words, AI and WMD programs have practically grown up together. To properly understand the impact of machine learning on WMD nonproliferation and evaluate its potential intersections, policymakers need to first examine where and when computers have already been used to aid in their development, existing areas of automation across the WMD development cycle, and the presence of relevant data at different stages of development that machine learning models could leverage.

Since the release of ChatGPT in 2022, the AI risk perception has largely reversed. However, too much emphasis has been placed on the risks of foundation models for the WMD domain. Generative AI is accessible and cheap and may lower barriers to useful WMD-related knowledge. Policymakers need to assess whether generative AI models lower the barriers to WMD development and use and monitor the advancement of those capabilities over time. Once the nonproliferation community has established initial benchmarks for AI's capabilities to assist nefarious actors in developing WMD (e.g., to lower tacit knowledge barriers), policymakers need to monitor how future advances might change the context.

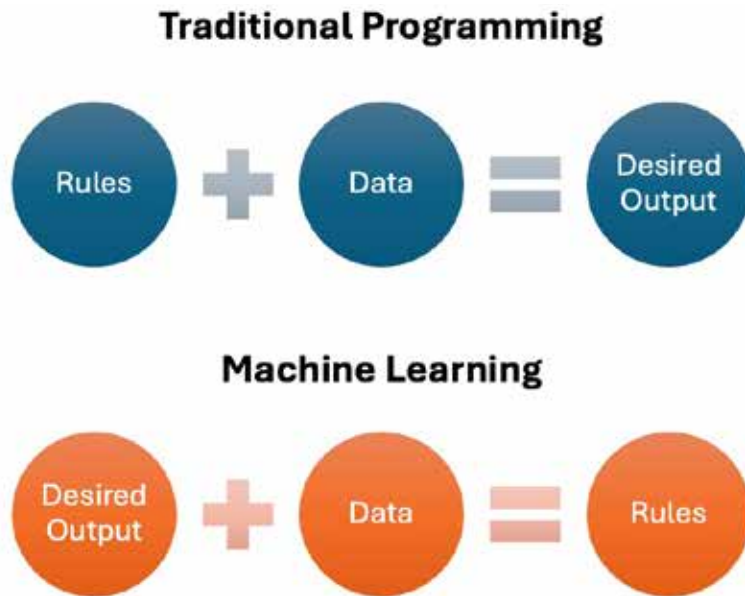
This primer defines AI as the science and engineering of making intelligent machines, but we acknowledge that there is no consensus definition.¹⁸ By integrating AI into computers and machines, they can perform complex tasks better than humans in areas such as planning, understanding language, recognizing objects and sounds, learning, reasoning, and problem solving. AI-enabled systems can achieve some tasks at levels that match and even exceed human capability. They can also perform tasks that humans are not at all capable of completing.¹⁹

MACHINE LEARNING

Although artificial intelligence has been a field of computer science for nearly seven decades, recent breakthroughs in machine learning have propelled AI back into the spotlight. For example, consider the growing public interest in the development of large language models (LLMs). First coined by Arthur Samuel in 1959, the term machine learning refers to “the ability to learn without being explicitly programmed.”²⁰

At its essence, machine learning represents a significant shift in how we program computers to solve complex problems. In traditional programming, developers specify a set of logical rules and data structures in order for the computer to produce desired outputs. To do this, programmers must know a specific domain's rules. That means humans must fully understand the problem before they can program computers to solve it.

Figure 2: Traditional Programming Versus Machine Learning



With machine learning, the developer instead specifies desired outputs up front and provides data, but humans do not give the computer explicit instructions. The machine learning algorithm then analyzes the data and automatically identifies statistical relationships and rules that allow it to generate the requested outputs from the new inputs. This new approach is enormously powerful because humans do not need to understand every aspect of a complex problem in advance to program a computer to solve it.



Solving Complex WMD Problems with AI

Today's AI can handle more complex problems using machine learning tools. The sophistication of current AI tools generates many possibilities for nefarious actors and policymakers engaged in WMD nonproliferation. With the ability to analyze massive volumes of unrelated data, AI may also allow policymakers to gain a better understanding of longstanding puzzles within the field of WMD nonproliferation—e.g., why certain countries develop nuclear weapons and others give them up. With the aid of AI tools to solve complex problems, policymakers may be able to develop countermeasures that deny the potential effects of using WMD. Doing so might reduce incentives to proliferate or enhance open-source analysis, which can help to hold nefarious actors accountable for their actions and aid in verifying existing nonproliferation agreements.

Machine and human learning involve different methodologies; both approaches produce meaningful albeit different outcomes. In the use case of visual object identification, a machine learning tool can be trained to determine if an image contains a dog. The tool is trained using a massive data set of labeled images (tagged as containing either a dog or no dog) and a classifier algorithm.²¹ Once data scientists tweak the algorithms to produce the most accurate outcomes possible, the machine learning tool can be unleashed to tag dog pictures on its own. At this point, the tool has “learned” what characteristics represented in the data indicate that an image contains a dog. However, this tool still does not “know” what dogs are; it can only determine if a picture contains one.

The above example illustrates the significant difference between machine learning and human learning. Humans can simultaneously identify dog pictures, understand how dogs behave, and determine their role within the broader context of the world. AI tools do not learn as humans do by collecting knowledge about a topic and deepening their understanding of the world. Instead, they analyze patterns in a dataset and predict outcomes with a probabilistic likelihood of being correct on a specific query.

In most cases, the quality and volume of training data are more important for predicting accurate outcomes than the quality of the algorithm itself.²² An average algorithm with high-quality data can outperform a superior model lacking the same quality or quantity of data.²³



AI Requires Massive Volumes of Quality Data

Machine learning tools use significant volumes of data to identify patterns and anomalies with the purpose of automating tasks previously performed by humans. Access to high-quality, relevant, and representative data is imperative for making AI tools function as advertised. AI is only as good as the data it was trained on.

Relevant and representative datasets do not exist for every type of problem we may wish to solve. This is particularly true for the WMD domain and the national security realm. If the training data do not match the problem to be solved, due to validity or representativeness problems, the model will fail to produce reliable outcomes. Policymakers need to be aware of the data gaps in the WMD nonproliferation domain since these are areas where leveraging AI could lead to undesirable outcomes. The key here is to be clear on the problem we are trying to solve and ensure that high-quality, representative, and relevant data exists to solve that problem before turning to AI for the solution.

DEEP NEURAL NETWORK

Most of the current breakthroughs in AI can be attributed to deep neural network architectures. The design of these architectures is based on the vast network of “neurons” in the human brain—the lowest level of computation.²⁴

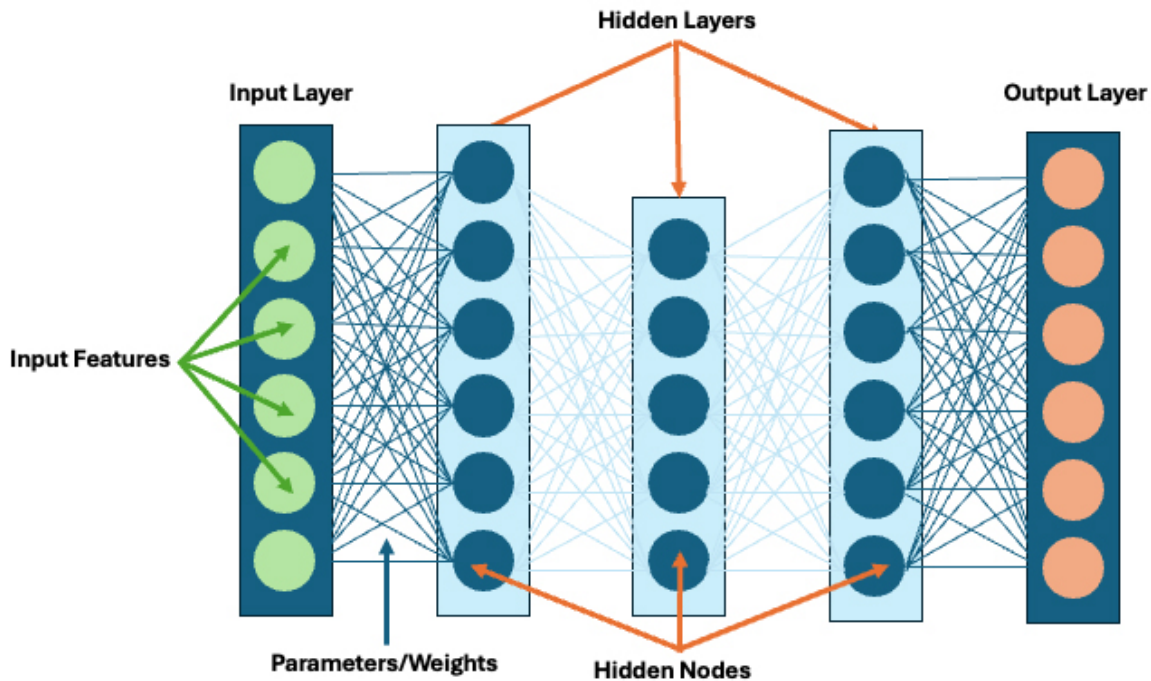
Deep neural networks contain many layers, each having nodes (also called variables or units). The units in one layer are connected to the nodes in another, and the strength of those connections is represented by weights (also called parameters). The input data (e.g., an image) passes through the different layers of nodes, with each node calculating an output from an input until it reaches the output layer and generates an outcome.

Unlike earlier neural networks, which tended to be limited to just a few layers, today’s deep neural networks contain many more layers of nodes. Today, billions of computations now occur simultaneously. This impressive computational architecture allows today’s neural networks to model highly complex relationships and feature intricate hierarchies within large datasets. However, the many hidden layers make it difficult to predict and explain why an input produces a specific output within an AI system. This lack of explainability is called the “black box.”

AI models are trained using a large dataset to determine the appropriate parameters/weights. These are the set of rules defining relevant relationships. Computer programmers can tweak models to achieve greater accuracy and make better predictions by adjusting the weights in a backpropagation process. In basic terms, this involves moving input data through the network and comparing the actual outputs to desired outputs (the difference is called the error). Then, the error margin is passed backward through the model to adjust the assigned weights (making corrections) and reduce errors. This process is repeated until the error margin is deemed

acceptable. The acceptable error margin will depend on the task, with applications in areas involving requests demanding the lowest error margins.

Figure 3: Deep Neural Network Architecture



Starting around 2012, costs for computing power and data storage dropped significantly and the production and availability of big data surged on the Internet. Consequently, major tech companies began deploying highly accurate and reliable deep neural networks to support their consumer products:

- Amazon: product recommendation/advertising system, Alexa, and autocomplete
- Facebook: content recommendation/advertising system and content moderation
- Netflix: content recommendation system
- Google: Google browser search, sponsored ads, and autocomplete

If you've used computers in the past ten years, you're already familiar with some of the world's most sophisticated AI models. Today, machine learning is used in practically every up-to-date software application, making it extremely difficult to know what capabilities are new and what has been in place for several years. Interestingly, we tend to refer to the most recent developments as AI and then take the rest for granted. This habit also leads us to interpret every new development as revolutionary until outsized expectations level off, and we adjust to a new normal.

PREDICTIVE VERSUS GENERATIVE AI

Before proceeding, it is helpful to think about machine learning tools as being categorized conceptually into two broad types—predictive AI (or task-based AI) and generative AI—in order to assess potential use-cases for the WMD domain.²⁵ A use-case is a specific scenario that describes how an AI tool can be used to achieve a particular goal.

Dividing AI tools into these two categories is not to say that only one type of AI tool is predictive in the technical sense; all AI tools are based on probabilistic predictions from their training data. However, making this distinction helps us to understand the primary characteristics and design of different types of tools and to place generative AI into its broader context.

Figure 4: Predictive Versus Generative AI: Comparing the Features

	Predictive AI	Generative AI
Function	Analyzes patterns within data to predict outcomes	Generates novel content based on patterns within existing data
Accessibility	Limited access and can be cost-prohibitive	Available online, with free versions and low monthly subscription fees
Accuracy	Generates outcomes with a specified level of accuracy	Accuracy is not guaranteed by design and varies widely
Outcomes	Fixed; models are designed to answer specific questions	Dynamic; models are designed to produce novel content every time
Methodology Example	A visual object identification tool make predictions based on labeled data	A large language model (e.g., ChatGPT) uses a transformer architecture to statistically predict the next words in a sentence based on existing data
Complementary Potential	Enhance the performance of generative AI models by classifying outputs based on pre-defined characteristics	Enhance the performance of predictive AI models by generating synthetic data for improved training

Predictive AI

Most of the robust AI systems deployed in the past decade are considered “narrow AI” and predictive tools, which are now often called traditional AI. These systems are specialized tools that exceed human capabilities on specific, well-defined tasks.²⁶ The models are trained on massive datasets, and often but not always, each observation is labeled in advance. They produce evidence-based outcomes by analyzing patterns found within the data with a probabilistic level of certainty.

To provide an example, a predictive AI model may be trained to predict with a ninety-percent likelihood that a given image contains a cat. It does so based on a large dataset of labeled images indicating the presence of a cat or the absence of a cat. If this margin of error is too


significant for the use-case, a computer programmer will tweak the model by changing weights in order to achieve higher accuracy.

Predictive models are designed to analyze the relationships among data and generate valuable, evidence-based answers to solve complex questions using statistical methods and pattern recognition. This is similar to applying statistics to large datasets to determine correlative and causal relationships using the scientific method. Predictive models rely upon massive volumes of high-quality, representative, and relevant training data to function correctly.

It is also essential to understand that predictive models “will do precisely what they are programmed to do.”²⁷ They are designed to perform well for the exact use-case for which they have been trained and cannot be transferred to solve another problem set. The questions to be answered by the tool must be well-defined in advance, and the data collected needs to answer those questions. If, for example, the relevant data changes or there is a change in the questions posed, the data scientist must start over by developing a new dataset and tool from scratch.

Thus, predictive AI models still lack humans’ multifaceted, flexible, general intelligence and our innate ability to transfer learning from one area to another. However, this finding should not be interpreted to undercut predictive AI models’ utility in achieving outcomes. Predictive AI tools have broad applicability to many sectors and tasks. They perform specific tasks, analyze data to make predictions, categorize inputs, and generate insights. Common types of tools and potential WMD domain use-cases are described in Table 1.

Table 1: Predictive AI Tools and Potential WMD Domain Use-Cases

Type of AI Tool	General Applications	 Potential WMD Domain Use-Cases
Classifiers	Decide between two options (e.g., spam detection) Choose among multiple options (e.g., identifying the type of animal in a photo) Assign several labels to each item (e.g., tagging a news article with multiple keywords)	Analyze shipping manifests and sensor data to detect suspicious materials that could indicate the smuggling of nuclear, chemical, or biological substances. Interpret satellite images for signs of clandestine WMD, such as unusual construction or activity patterns.
Recommender systems	Suggest items based on what similar users liked (e.g., recommending movies on a streaming platform) Recommend items like ones you've liked previously (e.g., suggesting books based on your past reads)	Recommend articles and reports of interest on WMD topics to policymakers Recommend tools or workflows to enhance productivity and streamline tasks based on work habits and needs.
Regression models	Predict a specific value based on a straight-line relationship (e.g., estimating house prices from size and location) Predict values when the relationship is not a straight line (e.g., predicting sales growth, which might accelerate or decelerate)	Estimate the likelihood of proliferation activities based on geopolitical, economic, and social factors. Analyze trends in compliance data to predict potential violations or lapses in security protocols.
Anomaly detection systems	Find unusual patterns without needing labeled examples (e.g., detecting unusual bank account and credit card transactions) Learn from past examples to spot abnormalities (e.g., identifying manufacturing defects in products)	Detect unusual patterns in imports and exports that could indicate illicit movement of WMD-related materials. Identify anomalies in the supply chain that may suggest unauthorized access or diversion of sensitive components.
Forecasting models	Predict future values based on past data trends (e.g., forecasting the stock market direction) Estimate future demand for products or services (e.g., predicting how many units of a product will sell)	Anticipate advancements in technology that could impact proliferation risks and require regulatory updates. Forecast emerging proliferation threats by analyzing geopolitical trends and intelligence data.
Dimensionality reduction tools	Simplify data by reducing the number of variables but keeping the essential information (e.g., condensing survey data to key factors/crosstabs) Uncover simpler underlying structures in complex data (e.g., finding the core trends in customer feedback)	Reduce complex datasets, such as trade and communication records, to identify key variables and patterns indicative of proliferation activities. Identify critical factors that contribute to proliferation risks, guiding targeted resource allocation and intervention strategies.
Sequence prediction models	Predict future actions based on past behavior (e.g., anticipating a user's next click on a website)	Anticipate potential routes and methods for smuggling WMD-related materials based on historical data and patterns. Forecast suspicious behavior sequences in personnel access or actions within sensitive facilities.

Sentiment analysis tools	Examine text to determine the mood or opinions it expresses (e.g., analyzing customer reviews to gauge satisfaction)	Assess sentiment in social media and online forums to identify potential threats or support for proliferation activities. Evaluate sentiment in international communications to gauge the stance of different countries on WMD issues.
Reinforcement learning models	Develop strategies to make a series of decisions that lead to a goal, learning from past outcomes (e.g., a robot learning to navigate a maze)	Model and predict adversary behavior in proliferation scenarios, enhancing strategic planning and decision-making. Develop adaptive response strategies that adjust to changing dynamics in real-time, helping to de-escalate tensions.
Image analysis tools	Identify what is depicted in a photograph or image (e.g., recognizing whether a photo contains a cat or a dog) Locate and identify objects within an image (e.g., spotting and labeling cars in a street scene) Divide an image into parts relevant for deeper analysis (e.g., isolating individual cells in medical imagery for detailed examination)	Detect construction of suspicious facilities or changes in known sites that could indicate WMD development. Analyze X-ray and other imagery of cargo to identify hidden or illicit materials related to WMD. Assist in verifying compliance with international treaties by analyzing imagery from inspections and audits.

Generative AI

In contrast, generative AI tools are trained “to produce new data that is similar to a given dataset.”²⁸ These models are primarily trained using unsupervised learning techniques—and often fine-tuned and improved using a few other techniques—on a massive dataset. Such a dataset usually consists of many examples of the data to be generated. The models identify patterns and trends within the training dataset and generate a set of rules about the relationships among the data. Then, they mimic those patterns while extending them by introducing novel features absent from the original inputs. In other words, put simply, generative AI tools produce novel outputs.

Generative AI models are “probabilistic rather than deterministic” in that they produce an unlimited variety of outputs “rather than get the same output every time.”²⁹ This characteristic is critical to understanding the power of generative AI models and their limitations for providing accurate, fact-based answers. Even with the same prompts, every output of a generative AI tool is novel. But sometimes, the outputs can be entirely fake or inaccurate. Since novel outcomes are determined by probabilistic distributions within the training data, the outputs of generative AI models contain an element of randomness. In other words, accurate outputs are not guaranteed by design, and these models often provide false outputs called hallucinations.




AI Models Exhibit Flaws, Risks, and Limitations

A critical obstacle to leveraging generative AI for WMD nonproliferation is the phenomenon of hallucinations, wherein the AI system generates inaccurate or misleading information that appears plausible. For example, imagine using a text generation tool to design scenario narratives for training exercises and strategic planning, helping stakeholders understand potential outcomes. Amid the scenario exercise, you discover the model got several technical details incorrect, raising questions about the viability of the exercise. This problem underscores the importance of keeping human experts in the loop to ensure the reliability and accuracy of AI outputs. Human oversight is crucial for verifying AI-generated insights, contextualizing findings within the broader strategic landscape, and making informed decisions that could have profound implications for global security. By combining the strengths of AI with the expertise and judgment of human analysts, we can better navigate the complexities of WMD nonproliferation while mitigating the risks associated with AI.

Policymakers also need to consider the unintended side effects of the AI revolution, including its impact on climate change and nuclear proliferation. The more complex the AI models become, the more energy they burn. Several tech developers have expressed interest in nuclear energy, especially small modular reactors, to provide sufficient clean power for AI.

Generative AI includes tools like large language models (LLMs) such as ChatGPT, diffusion models such as text-to-image generators, and music and video generators. As such, these models can synthesize novel content, including images, text, computer code, musical notes, video, and audio. Common types of tools and potential WMD domain use-cases are described in Table 2.

Table 2: Generative AI Tools and Potential WMD Domain Use-Cases

Type of AI Tool	General Applications	 Potential WMD Domain Use-Cases
Text generation	<p>Create text such as articles, stories, or reports based on given inputs or prompts (e.g., generate news articles from data inputs like sports scores or financial data, or automate routine reporting tasks in journalism)</p> <p>Generate conversational responses in real-time to interact with users, simulating human-like discussions (e.g., provide customer support chatbots that can manage inquiries and solve basic consumer problems)</p>	<p>Automatically generate detailed reports on WMD-related activities or compliance assessments for policymakers and international bodies.</p> <p>Create realistic scenario narratives for training exercises and strategic planning, helping stakeholders understand potential outcomes.</p> <p>Assist in drafting diplomatic communications or policy documents related to WMD treaties and negotiations.</p>
Image generation	<p>Generate new artworks or images based on various styles or prompts (e.g., artists can use AI to explore new creative styles or to generate art pieces based on specific themes or historical art movements)</p> <p>Improve or alter photos by enhancing resolution, adjusting colors, or adding elements that weren't originally there (e.g., real estate companies can enhance property photos automatically to show homes in different lighting conditions or to beautify surroundings, making listings more appealing)</p>	<p>Create realistic images for use in training scenarios, helping personnel recognize WMD-related materials or equipment.</p> <p>Generate visuals to support educational materials that raise awareness about the dangers and prevention of WMD proliferation.</p> <p>Enhance datasets for machine learning models by generating diverse images, improving model accuracy in detecting WMD-related activities.</p>
Music and sound generation	<p>Compose new pieces of music in various styles or continue a given musical piece (e.g., a composer working on a film score can input a theme and have AI develop variations to fit different scenes, saving time, and sparking creative ideas)</p> <p>Create sound effects for use in games, movies, and other media, often from scratch or by modifying existing sounds (e.g., a game developer can use AI to produce a library of unique sound effects, enhancing the immersive quality of the game world)</p>	<p>Create immersive audio environments for training exercises, enhancing realism and engagement for personnel.</p> <p>Develop soundtracks or audio content for educational videos and campaigns to raise awareness about nonproliferation.</p> <p>Create audio for conferences or workshops focused on nonproliferation, setting the tone and enhancing the experience.</p>

Video generation	<p>Create new video content or alter existing videos, such as changing day scenes to night or adding objects that weren't originally present (e.g., filmmakers can alter scenes in post-production, such as changing weather conditions or adding crowd scenes, without costly reshoots)</p> <p>Create realistic training videos (e.g., for emergency response teams, by simulating various crises that are difficult, impossible, or unethical to film in real life)</p>	<p>Develop realistic training videos that simulate scenarios involving WMD threats, enhancing preparedness for security personnel.</p> <p>Produce instructional videos for inspectors, demonstrating best practices and what to look for during facility inspections.</p>
3D model generation	<p>Design virtual environments (e.g., for training simulations in sectors like aviation or military, where real-world training can be hazardous or expensive)</p> <p>Companies can quickly prototype new product designs, visualizing and iterating on 3D models before committing to physical prototypes (e.g., automotive companies create and iterate on 3D models of new car designs virtually, speeding up the prototyping process and reducing manufacturing costs)</p>	<p>Create realistic 3D models of WMDs and related equipment for use in training simulations, helping personnel recognize and respond to threats.</p> <p>Model facilities to evaluate security vulnerabilities and optimize layouts for compliance with nonproliferation standards.</p> <p>Use 3D models to train inspectors on what to look for during site visits, improving accuracy and efficiency in compliance verification.</p>
Data augmentation	<p>Generate synthetic data or enhance existing datasets to improve the training of machine learning models (e.g., for example, anonymized healthcare data to train predictive models without compromising patient privacy)</p> <p>Enhance limited datasets in machine learning projects (e.g., adding synthesized weather conditions to improve models predicting energy usage in smart grids)</p>	<p>Enhance datasets with synthetic examples to improve the accuracy of models used for detecting proliferation activities.</p> <p>Generate variations of satellite images or surveillance footage to train models to identify suspicious activities or facilities.</p>
Code generation	<p>Generate code snippets or entire programs based on specific requirements, aiding in software development (e.g., allowing developers to focus on more complex and innovative aspects of projects)</p> <p>Help researchers or developers by suggesting new algorithmic approaches (e.g., solving complex problems in fields like cryptography or network security)</p>	<p>Create code for processing and analyzing large datasets related to trade, finance, and communications to identify proliferation risks.</p> <p>Develop custom simulation tools to model potential proliferation scenarios and assess the effectiveness of prevention strategies.</p> <p>Build interactive educational platforms that simulate WMD threat scenarios for training purposes.</p>

Narrow Artificial Intelligence Versus Artificial General Intelligence (AGI)

Until recently, most deployed AI models belonged to the category of narrow AI. This refers to models designed to perform a single, narrowly defined task. However, narrow should not be confused with less useful or simple. For example, AlphaGo is an algorithm developed by DeepMind based on a deep neural network; it plays the board game *Go* and became capable of beating world champions in 2016. Yet, as good as the model is, it cannot play chess or even checkers without modifying the data and algorithm; this makes it narrow AI.

The recent emergence of generative AI is extraordinary because of these models' general capabilities. Whenever the model improves, it improves across all its capabilities at the same time.

Many researchers are deeply concerned about the development of computerized general intelligence. Nick Bostrom, claimed in the 2014 book *Superintelligence* that "if somebody were to succeed in creating an AI that could understand natural language as well as a human adult, they would in all likelihood also either already have succeeded in creating an AI that could do everything else that human intelligence can do, or they would be but a very short step from such a general capability."³⁰ A host of other experts have warned that superintelligence will emerge shortly after the achievement of artificial general intelligence (AGI). This fear refers to the idea that machines will eventually learn and perform equally or superior to humans across unlimited tasks.

Are we on the cusp of creating artificial general intelligence?

Experts remain divided on this issue. Today's AI relies heavily on mathematical and statistical techniques like machine learning. However, many tasks that humans excel in cannot be neatly reduced to numerical computation. For example, open-ended creative work, complex social interactions, and tasks requiring contextual reasoning defy straightforward mathematical formalization. While architectures like deep neural networks take inspiration from the brain, they are still very far from emulating human cognition and behavior.



Predictive AI Tools Are Already Being Deployed

Despite the attention around generative AI, the widespread deployment of AI tools in the national security arena is already underway. Most of the AI tools deployed in the near-term will belong to the predictive, narrow, task-oriented category. As just one example, Project Maven started in 2017 and used machine learning to analyze video footage captured from U.S. uncrewed aerial systems overseas to identify potential targets. Project Maven is now a major contributor to the Pentagon's Combined Joint All Domain Command and Control concept—which is AI-enabled decision support and situational awareness system for conventional and nuclear operations.

Within the WMD domain, the most immediate impact with predictive AI models will occur in data/automation-heavy stages of the development pathways, including research and development, production, and delivery, but their impact will vary greatly across the development of nuclear, biological, and chemical weapons. As they advance, generative AI models may someday produce effects for all WMD across their development pathways.

Policymakers need to examine how predictive and generative AI capabilities have changed the game for nefarious actors when it comes to WMD development and use. To measure future impact and avoid hyperbole, experts must establish baselines for today's AI capabilities related to WMD and track them as they evolve (i.e., benchmarks and evaluations). A future area of concern will entail building tools that leverage the synergies between predictive and generative AI such as DeepMind's AlphaFold 3.

In Chapter 2, we will examine the workings of generative AI to understand how the technology works. Several features of the technical architecture warrant caution when drawing conclusions about AGI and superintelligence.

Chapter 2: The Generative AI Landscape

Generative AI is an expanding class of powerful tools built on machine learning and deep neural network architectures. These include large language models like ChatGPT and Llama, diffusion models like Midjourney and DALL-E, and multi-modal models like Gemini. Since their original releases, each model has been updated several times, and AI companies continue to work around-the-clock to advance their respective technologies. Despite their release for public use, as of the writing of this primer, these models often act like beta versions of new software and should be considered experimental. They can produce unexpected or undesired outputs, and users should be prepared for these outcomes.

As discussed in Chapter 1, generative AI tools are designed to produce novel data based on patterns and trends found in their training datasets. They generate novel outputs rather than produce specific, evidence-based, or fact-based answers with a certain level of accuracy. To accomplish their intended purpose, generative AI models learn the “probability distribution” of the training data and develop a corresponding set of rules that enables them to produce new content. Such models tend to produce content like the most common examples within a dataset.

This technical consideration is essential for understanding how generative AI models work and being aware of some of their most important flaws (for further discussion, see Chapter 4).



AI Models Exhibit Flaws, Risks, and Limitations

Generative AI models are useful for producing creative content, brainstorming solutions, and other conceptual tasks, but human expertise is still needed to evaluate accuracy and quality. Generative AI models produce novel content, which can sometimes be factual, but accuracy is not guaranteed. Although these tools can help to improve productivity and automate specific workflows, a human needs to remain in the loop to prevent embarrassing or damaging mishaps. Imagine a policymaker using an AI tool to draft a plan to address the cyber vulnerabilities of AI tools for WMD nonproliferation because they do not have a cyber expert on staff. The AI generates a comprehensive cyber defense plan that appears logical on paper. However, without verification by a human cybersecurity expert and other stakeholders, several critical issues are missed:

- Oversimplified explanation of attack vectors
- Outdated references to vulnerabilities and adversarial techniques
- Incompatibility with legacy systems
- Unrealistic timelines for security upgrades
- Resource requirements that exceed agency capabilities

This example illustrates why human expertise, knowledge, and stakeholder input remain essential in policy decisions, even when using AI tools.

A probability distribution is a statistical function of all possible values and likelihoods for a continuous random variable within a given range. We tend to be most familiar with a normal distribution (also called a Gaussian distribution, or the infamous “bell curve”). A normal distribution shows the distribution of a continuous random variable in relation to the mean, or the average, of the data. For all normal distributions, 68.2-percent of the observations in a dataset will appear within one standard deviation of the mean (plus or minus)—i.e., within the middle part of the bell curve. This explains why the models produce outcomes that seem accurate and expected but vary slightly for each prompt.

Generative AI models can use advanced statistics to produce novel text, images, videos, and sounds. Whereas LLMs can generate human-like text in almost any language (including computer code), diffusion models can generate high-quality images, video, and other visual data from text descriptions.

TRAINING TECHNIQUES

Knowing how generative AI models are trained to produce their outcomes is helpful for understanding how they work. Several different techniques exist, and they determine, in part, how algorithms use data (observations) to generate outcomes (inferences).

Supervised Learning

Supervised learning techniques require labeled datasets consisting of sample inputs (e.g., images) matched with the corresponding outputs (e.g., an image tagged to contain a dog or no dog). During the training process, an algorithm develops a set of mathematical rules to map the relationships between these inputs and outputs. The more complex the model, the more complex the behavior that can be learned from the training data, and the more training data that will be needed to generate reliable outputs. Once trained, data scientists can tweak the algorithm to minimize error rates and get more accurate outputs using back propagation. They can then (re)deploy the tool to achieve its intended purpose or use-case.

Unsupervised Learning

Unsupervised learning techniques use unlabeled datasets (i.e., raw data). This approach works well “when there is not a clear outcome of interest about which to make a prediction or assessment.”³¹ Data scientists tend to use this technique when they are looking to discover something new about the data, such as the hidden structure, patterns, distribution, or correlations within the dataset. Unsupervised learning systems do not receive external feedback on their predictions or inferences; the outcomes are solely driven by patterns and trends in the training data. In their primary training phase, LLMs are typically developed using this technique. These models identify patterns in the training data, which is often scraped from the Internet. This allows them to predict the next words in a sentence, for instance.

Reinforcement Learning

Reinforcement learning techniques generate “synthetic” data to train AI models. This process occurs as a machine learning algorithm learns how to play a game or operate in a specific environment with predetermined rules and boundaries. The algorithms are not taught the rules of the game or given any data to analyze in advance. Instead, they are expected to reach a series of decisions given a stated goal of achieving optimal outcomes—e.g., in the case of a game, the optimal outcome is to win. At the end of the gameplay or series of decisions, the algorithm receives feedback in the form of a reward or a penalty. Over time, as the algorithm plays the game repeatedly, it learns the most optimal sequence of moves to win the game and receive the greatest rewards. This technique was most famously used to train DeepMind’s AlphaGo, which beat world champion Lee Sedol and made a particularly expected move for its 37th move. At first, many experts thought it might be a mistake, but then the move changed the course of the game, leading to another victory by AlphaGo.³² The machine learning algorithm had discovered a new way to play the ancient game.

Reinforcement Learning From Human Feedback (RLHF)

Reinforcement learning from human feedback (RLHF) is similar to reinforcement learning as a general concept, but it involves human feedback. The technique leverages human feedback in the rewards function to ensure the AI models perform tasks in a way that is aligned with human goals, wants, and needs (this is called alignment). RLHF is also the main technique tech developers employ to prevent their models from being misused or from engaging in undesirable behaviors. It also helps to ensure that outputs are truthful, harmless, and helpful and is associated with the notion of “guardrails” in the field of AI safety.³³



AI Safety and Developing Effective Guard Rails

To implement guardrails in AI safety, reinforcement learning from human feedback (RLHF) is often used. Guardrails refer to measures and protocols designed to ensure safe, ethical, and reliable operation. They include guidelines, constraints, and safety checks that prevent models from producing harmful or biased outcomes—including limitations on what types of information the models share related to WMD. The RLHF technique involves training AI models through receiving feedback from human evaluators. With positive and negative feedback, the model learns to align its outputs with desired behaviors and safety. By setting these boundaries, developers can minimize risks and ensure AI systems act within acceptable norms and standards. The type of guard rails varies by AI model, and the differences are especially profound between closed/proprietary models such as Open AI’s ChatGPT and open-source models such as Meta’s Llama. With decent prompt engineering (i.e., the crafting of prompts), however, determined actors can get past these guard rails; such efforts are called jailbreaking, which refers to getting the models to provide outputs the developers tried to prevent them from offering.

TYPES OF GENERATIVE AI MODELS

This section examines different types of generative AI models and explains them in very basic terms: what they are and how they work. At least six families of generative AI models exist, but we will focus on the most relevant types for the WMD nonproliferation domain.

Generative Adversarial Networks (GANs)

Over the past decade, generative adversarial networks (GANs) have received much attention in the national security space given their incredible capacity to produce “deep fakes.” This term refers to images, videos, and other types of data capable of deceiving humans into thinking they are real, thus exacerbating disinformation challenges on the Internet.

Ian Goodfellow introduced the notion of a GAN in 2014 as a powerful way to generate realistic data from an existing dataset (e.g., text, images, audio, and video). His ideas have shaped the field of generative modeling ever since.³⁴ Unlike the newer modeling approaches we are more familiar with today, a GAN consists of two deep neural networks trained together using an adversarial process wherein they compete to achieve opposing objectives.³⁵

The two networks are trained using an unsupervised learning technique. The “generator” is trained to create fake data indistinguishable from its training data, and the “discriminator” is trained to detect fake data produced by the generator. At a basic level, the generator creates fake data by sampling the original dataset and then converting random noise into an image or video that matches the selected sampling. The discriminator compares the fake data to the original dataset and predicts whether the new observation is authentic or fake with a certain level of accuracy. At the start of this process, the generator is not good at achieving its task and produces noisy images that are easy to detect. Similarly, the discriminator exhibits a large margin of error in predicting the validity of the data.

However, as the networks are trained to achieve better outcomes through a competitive process, they innovate and improve over time. Although this process sounds simple, GANs are extremely difficult and expensive to train.³⁶



The Growing Challenge of Scalable Disinformation

The challenge posed by deep fakes for the WMD nonproliferation domain will accelerate in the near-term. National security experts have discussed the risk of deep fakes for several years already. Open-source tools for producing disinformation (GANs) have been around for some time. Since 2022, however, generative AI models (LLMs), have made such tools available to anyone with a computer and Internet connection. Although generative AI models are not guaranteed to produce factual outcomes, their ability to get close enough and trick human minds, eyes, and ears, is sufficient to make them powerful disinformation tools capable of creating content at incredible scales. The challenge of scalable disinformation consists of tools for automated content generation and rapid dissemination, and the ability to target a specific audience. This combination leads to rapid amplification of messaging and immediate impact.

Large Language Models (LLMs)

Large language models, often called generative pre-trained transformers (GPTs) or chatbots, represent a special class of models that generate text in natural language. LLMs like OpenAI's ChatGPT can engage in conversational exchanges, answer questions, and generate human-like text "on demand" by modeling statistical regularities in enormous natural language datasets. As explained above, these models use advanced statistics to predict the most likely text response based on the specific prompt given by a user.

LLMs use a transformer architecture to develop a working model of natural language—a set of rules about the relationships between words and sentences. A transformer architecture is designed to process and generate data in a sequence, called a token. When given an input, the models then predict the next token or words in a sentence. Currently, such transformers "are pre-trained," meaning they do not have direct access to information on the Internet beyond their last training date. This feature affects their accuracy in providing responses to current events; LLMs either make up a false answer or respond with their information cut-off date.

LLMs are typically trained using a hybrid approach (i.e., more than one learning technique) involving three or more steps. In the first step, the models are trained using unsupervised learning on massive datasets containing pre-processed text scraped/collected from the Internet or other sources, which serves as their raw data. These models "learn" what rules to follow—grammar, context, sentiment, and knowledge domain patterns—from the training datasets. During the training process, the LLM is presented with a sequence of words and asked to predict the next words in the sequence.

In the second step, the models are fine-tuned using a supervised learning technique, often involving labeled data, to improve their performance (sometimes for a specific domain). Finally, many models are put through a third training phase that leverages RLHF. In this

step, human evaluators read the responses from LLMs and provide positive feedback to reward good responses and negative feedback to indicate poor responses.³⁷

This intensive process results in what we call a “foundation model” or “frontier model.” These models can be further fine-tuned for carrying out specific tasks and leveraged to produce applications (for more on this topic, see Chapter 3).

This description explains why these models can produce text that sounds right but is not guaranteed to be accurate. Foundation/frontier models do not know things like humans do; they are also not concerned with the truth.³⁸ While their outputs appear intelligent, the systems have no real understanding of the content being generated. They lack a coherent internal mental world driven by experiences, self-awareness, and goals. They are simply predicting the next words in a sentence based on the context provided by the prompt.

Understanding the intensive training process also offers insights into how these models produce novel content. LLMs can generate poems, novels, screenplays, computer code, and human-like text responses to user prompts. The outputs of LLMs captivate us because the statistical repetition of patterns evokes the superficial appearance of intelligence—both expected and surprising at the same time.³⁹ Despite the basic skill of predicting the next words in a sequence, the models can do many valuable and powerful things that some developers never predicted.

Generative AI models also do not create content the same way humans do. They follow statistical rules that are learned from patterns in the data and reproduce those patterns in novel content. In contrast, human creativity involves imagination and branching out in novel directions based on unique experiences in the real world.

Diffusion Models

In the visual domain, diffusion models have become more popular than their GAN precursors due to their superior performance and greater ease of training. However, they share many of the same ideas and concepts as GANs.⁴⁰ Interestingly, the name diffusion model comes from the concept of thermodynamic diffusion, a physical property of atoms that has been leveraged to enrich uranium for producing nuclear energy and nuclear weapons.

In basic terms, diffusion models are trained to add Gaussian noise (this looks like static on an old-school television) to training data (e.g., images). Then, they are trained to remove the noise by predicting how it was added. Once the model is fully trained in adding and attenuating or reversing noise, it creates realistic images based on text prompts by users (i.e., text-to-image conversion). This process can also work for other types of visual data, including videos.

Several diffusion models have been made available to the public (e.g., DALL-E, Midjourney, Stable Diffusion). They generate high-quality imagery from text prompts (i.e., text-to-image).

Multi-Modal Models

Multi-modal models can convert between two or more modalities of data. For example, this might mean text-to-image, text-to-video, text-to-audio, and the reverse of each. These models integrate recent developments in computer vision, speech recognition, and LLMs.⁴¹ Multi-modal

models are far more complex than text-to-text models because they “must also learn how to cross the bridge between multiple domains and learn a shared representation.”⁴² Basically, they must be able to convert text to an image without any loss of information and produce novel outputs that may have never existed in an image.

The diffusion models described above are multi-modal. That is, they can create image and video data from text prompts, which involves different modalities of data (i.e., text-to-image).

To combine different data inputs and outputs into a single model, each piece of data must be encoded as an embedding (i.e., a vector of numbers) using a software tool called a variational autoencoder. This AI tool compresses each data type into a much simpler form, allowing the model to compare different categories of embeddings. The various data types, now represented in a similar form (embeddings), can be combined and decoded to produce the correct output.

Once trained, a multi-modal model can receive different data modalities as inputs and provide the expected response modality. For example, a model that can handle text and images can respond to text-to-image and image-to-text prompts. In this case, a user can input a photo of a flower and prompt the model with a text question like: “What type of flower is this?” Both pieces of data (image and prompt text) will be encoded as embeddings, combined, and then decoded to produce the text response: “a sunflower.”

Another user may want to generate an image of a sunflower. In this case, the user will create a text prompt to generate their desired outcome: “a field of sunflowers on a sunny day.” The diffusion model will encode the text prompt as an embedding, create a noisy image, and then produce an image matching the text prompt by removing the noise.

Google’s Gemini model represents one of the more powerful multi-modal models available to the public. It can work with text, code, images, and video. The model can therefore recognize images, interpret complex visuals and handwritten notes, and even translate across different languages.⁴³



The Value of Multi-modal Models for WMD Nonproliferation

Multimodal models are invaluable for organizations with vast collections of information in diverse formats, such as text, audio, video, and images including those in the WMD nonproliferation field. These models can simultaneously process and analyze multiple types of data, providing comprehensive insights that single-modal models might otherwise miss. By integrating and understanding correlations across different data forms, multimodal models enhance decision-making and uncover patterns more effectively. The technical process involves using AI tools that can handle various input types, combining them into a unified representation. This differs from the manual and time-intensive process of cataloging information and adding metadata such as keywords. Multimodal models analyze the content directly, allowing for deeper, context-rich interpretations beyond basic descriptive tags. This capability enables organizations to leverage their entire data ecosystem.

World Models

World models represent the next frontier in generative AI models.⁴⁴ Some experts claim that Open AI's Sora (image-to-video) model may come close to crossing this threshold.⁴⁵ The broad concept of a world model describes the process through which humans learn things about the world around them via information received through their senses. People then create an internal mental map of how the world works. As multi-modal generative AI models become better at producing outcomes across different data types, they may eventually possess an artificial understanding of the world.

In this case, rather than defining rules about the relationships between words and sentences, such models would establish complex rules about an environment or the entire world. For example, they could create rules that provide a detailed understanding of the world, including the “whys” and “hows,” the laws of science, and the nature of time. These models would be “capable of understanding, interpreting, and interacting with the world in a generalized way.”⁴⁶ Such models integrate multiple data sources and can produce outputs across numerous domains. They can also understand both contexts and causal relationships. They can learn and adapt their rules to changes in their environment. World models therefore aim to extend their complex understanding of the world to new situations and unknown data while making accurate predictions.

Given the complexity of world models, they are generally trained with multiple techniques. These may include unsupervised, supervised, and semi-supervised learning approaches, standard and model-based reinforcement learning, and more. Unsupervised learning is a valuable technique for discovering patterns in data without pre-existing labels, which is useful in complex environments where not everything can be neatly categorized or predicted based on past data alone. Supervised learning is used when historical data with known outcomes are available to train the models that can predict the next state of an environment based on current inputs, which is a crucial component of many world models. Model-based reinforcement

learning is used when the model needs to understand or simulate the environment, not just to determine the best actions within a specific environment. For example, reinforcement learning is used frequently in dynamic environments where the model must make decisions that affect future states, such as in robotics, gaming, and autonomous vehicles.

Currently, world models are primarily deployed in research and development settings as test beds. They are particularly useful, at present, in robotics and autonomous vehicle navigation, where an understanding of complex and dynamic environments is crucial for performance. These models are also becoming increasingly popular for virtual testing environments for AI because they reduce costs and increase safety. They do so by allowing models to learn and fail in a consequence-free setting before applying their learned behaviors to the real world.



Leveraging World Models for WMD Nonproliferation

World models, which simulate complex environments and predict potential outcomes, can significantly assist policymakers in preventing WMD proliferation. By creating dynamic simulations of geopolitical scenarios, these models allow policymakers to explore the consequences of various actions and strategies. They help in identifying potential proliferation activities by modeling interactions between nations and predicting the impact of policy decisions. World models enable scenario planning, risk assessment, and the testing of nonproliferation strategies, providing a virtual testing ground for policies before implementation. This proactive approach aids in developing informed, strategic decisions, ultimately enhancing efforts to prevent the spread of WMD.

CLOSED VERSUS OPEN-SOURCE MODELS

Closed and open-source models represent two distinct approaches to developing and distributing AI technology. Each comes with its own set of characteristics and implications for the WMD nonproliferation domain.

Closed models are developed, owned, and controlled by specific organizations like OpenAI or Anthropic. These companies often require a license to use the models, and the source code is closed to the public. This exclusivity can lead to higher quality assurance, more consistent updates, and dedicated support, but it also often comes with higher overall costs and less flexibility for customization.

In contrast, open-source models, such as Meta's Llama or Mistral, are available for anyone to use, modify, and distribute. The source code for these models is openly shared (including the weights), fostering a collaborative environment where developers worldwide can contribute to improvements and innovations.

Open-source models lower the barriers to entry for AI development, promoting wider accessibility and more rapid proliferation of the technology. However, in some cases, these models lack the dedicated support and update process that closed models offer, potentially leading to issues with maintenance, security, and consistency in performance across many different model versions.

Closed models come with licensing fees, whereas open-source models incur costs primarily related to the computational resources needed to run them. Each model type serves different needs and scenarios. Closed models are often favored in commercial applications that require robust support and liability protection. Open-source models are preferred in academic settings and among developers seeking flexibility and community collaboration.



The Open-Source Dilemma of AI Regulation

The differences between closed versus open-source models are critical for policymakers when considering their risks for WMD nonproliferation. Thus far, most policy solutions tend to focus on ensuring the safety and security of closed models such as Open AI's ChatGPT. Open-source models such as Meta's Llama tend to have fewer guard rails, since companies make the code and internal model weights available to other developers (and countries) who can then modify how the model functions. This division between closed and open-source models will pose a pivotal challenge to regulating AI and protecting society from harmful effects.

Performance Metrics

As implied above, generative AI models vary significantly across performance metrics such as speed, capability, ease of use, and cost. Each metric is influenced by the model's design and intended application. The models also vary in their scores for achieving certain widely recognized benchmarks for measuring AI's different capabilities.

- **Speed** - The speed of AI models varies widely. Simpler models can generate outputs quickly and efficiently. Complex models—particularly those involved in generating high-resolution images or videos or performing complex reasoning tasks—require more computational power and time.
- **Capability** varies primarily in terms of the quality and diversity of the outputs they can generate across different domains. Some models might excel in creating highly detailed and complex outputs while others might focus on delivering simpler, more consistent results.
- **Ease of use** is another critical factor. Models with user-friendly platforms and online interfaces are more accessible to non-specialists, whereas custom-built solutions might require significant machine learning and coding expertise.

- Cost is influenced by the resources needed for training and running these models. Larger, more complex models require substantial computational resources, thus increasing operational costs. Some of these costs are passed on to the users.

Benchmarks and Evaluations

Benchmarks are standardized methods for measuring and comparing the capabilities of different AI models. They offer a useful way to compare the performance of models at any given time.⁴⁷ For example, in natural language processing, various benchmarks might measure a model's ability to understand context, generate coherent text, perform mathematical calculations, or answer questions accurately. In image generation, benchmarks could assess the realism and resolution of the generated images. Benchmarks not only facilitate a direct comparison across the models, but they also help identify the strengths and weaknesses of different models.

Evaluations are crucial for assessing the capabilities of AI models, ensuring they perform as intended and meet desired standards. They are closely tied to benchmarks. By using benchmarks, researchers can assess how well AI models perform on specific tasks and datasets, ensuring consistency and fairness in evaluations.



New Tools For Preventing WMD Proliferation

The potential for generative AI tools to lead to a decline in tacit knowledge and specialized know-how about WMD development and delivery is a key issue for policymakers to monitor. Currently, most studies find that the models do not lead to an effective transfer of tacit knowledge but still may provide easier access to WMD-related information than the Internet. However, even then, given the tendency of generative AI models to produce fake results, human experts would still need to evaluate the validity of outputs.

Benchmarks and evaluations can help policymakers safeguard AI models by ensuring they are robust, secure, and ethically aligned—and do not provide nefarious actors with assistance in developing WMD. By setting clear performance and safety standards, benchmarks help assess whether AI systems can resist manipulation or misuse. Evaluations identify vulnerabilities and biases that could be exploited by nefarious actors. Regular testing against these standards ensures that AI models are not only effective but also resilient to threats. This proactive approach helps policymakers enforce stringent guidelines and implement necessary safeguards, preventing the misuse of AI in developing WMD.

Chapter 3: Generative AI Applications

Generative AI models are revolutionary, in large part, because they are broadly accessible to individuals, businesses, and developers alike. Until recently, small businesses and individuals could not easily gain hands-on experience using sophisticated AI models due to their costly and proprietary nature. However, as of late 2022, with the release of generative AI models by various companies, this landscape changed.

For most generative AI models, users do not need to be able to write any computer code to learn about them, use them, generate useful results, and integrate them into their daily workflows. As many of these models operate as online chatbots, they are accessible to anyone with Internet access, and many companies offer free versions to get users started. These interfaces/chatbots generate responses to natural language instructions (called prompts) entered in a textbox (called a context window).

Much of the buzz around generative AI has less to do with their function as chatbots than with their potential as engines for work automation. According to technology experts, these experimental AI models are “just the first step in the evolution of accessible AI for software development.”⁴⁸ In this chapter, we will review the basic functions of generative AI models and then discuss the current trends toward work automation through AI agents (also called agentic AI).⁴⁹



Improving Productivity at WMD Nonproliferation Organizations

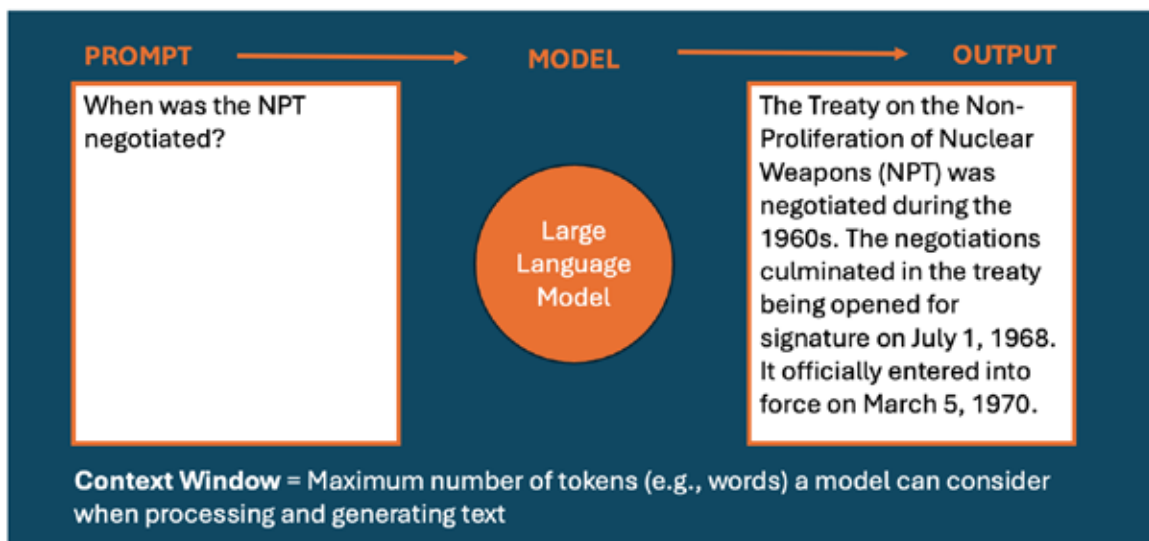
Organizations dedicated to WMD nonproliferation can leverage generative AI models to enhance productivity and streamline workflows. For example, by automating routine tasks, such as writing emails, summaries, and even draft reports, AI models can allow experts to focus instead on providing strategic analysis and insights. Large language models can create realistic simulations for training and scenario planning. Additionally, they can assist in drafting policy documents and communication materials, ensuring consistency and efficiency. Overall, generative AI can help to optimize processes and enhance the effectiveness of operations. However, current generative AI models are still experimental; many flaws and limitations remain unresolved. For this reason, caution is warranted when using generative AI models, even at a basic level, to help simplify work tasks. Given their tendency toward hallucination, professionals should not rely upon results without fact-checking and conducting a comprehensive review.

IMPROVING PRODUCTIVITY WORKFLOWS THROUGH AI ASSISTANTS (CHATBOTS)

As of the writing of this primer, most individuals and small businesses use generative AI models by entering natural language prompts into the context windows to generate results to improve work productivity. They use the models to achieve tasks including writing text—blogs, essays, poems, etc., creating images, summarizing uploaded documents, translating uploaded text, writing computer code, or merely engaging in a straightforward question-and-answer session (called inferences).

The art of writing prompts and getting higher-quality results from generative AI models is called prompt engineering. The sophistication of a user’s prompt is only limited by the context window. This limit is posed by the maximum number of words (or tokens) a model can consider when processing and generating outcomes. As illustrated below, a basic inference is a simple prompt without providing further instructions or examples.

Figure 5. Visual Depiction of a Basic Inference



The example above is called a zero-shot prompt; it is the most common way people get started with using the models. Consequently, they often walk away disappointed because the models are not designed to perform well in response to simple factual questions and also frequently provide inaccurate answers. Users should ideally engage in prompt engineering techniques to gain more useful and sophisticated results. Appendix A provides a practical guide for using different types of prompts to improve your productivity with generative AI models.

Moving beyond this basic use-case for AI models, this chapter will examine more advanced generative AI applications that help explain the current excitement around these models. Most of the activity in the field of AI and the commercial sector today revolves around leveraging generative AI models to upgrade web searches, enhance AI models for use in specific domains, and automate workflows using agents (also called agentic AI).



Investments in Basic Training on Prompt Engineering

Investing in prompt engineering training can significantly enhance the value AI models provide to organizations focused on WMD nonproliferation. Prompt engineering techniques can yield more sophisticated results than basic prompting by refining and guiding AI model responses. Techniques such as using contextual keywords, specifying format, chain-of-thought prompting, or including examples help the model understand the task better and produce more accurate and relevant outputs. By iteratively refining prompts based on feedback, users can optimize responses for complex tasks, ensuring the AI aligns more closely with specific needs and objectives. By equipping staff with the skills to craft precise and effective prompts, organizations can extract deeper value from AI models.

UPGRADING SEARCH ENGINES WITH GENERATIVE AI

A tense competition to upgrade search engines using generative AI is underway among major tech companies.⁵⁰ In addition to the emergence of new startups such as Perplexity AI, Google and other companies owning search engines have been exploring the use of AI to enhance search results for several years now. Most recently, they have started to integrate generative AI into their search engines or search features into their AI models. OpenAI launched a prototype AI-driven search engine called SearchGPT in July 2024, which was renamed ChatGPT Search and now allows ChatGPT to search the Internet for up-to-date results.⁵¹

Google operates the most used search engine in the world, receiving 8.5 billion searches per day. That's roughly 99,000 queries every second (as of April 2024). To maintain its competitive edge, Google has recently introduced AI overviews with generative AI to "enhance" its browser search results.⁵² Google's data scientists have tried to get around the problem of the "models making up stuff" by having the algorithm source "quality" responses from the Internet and provide links to cited sources. However, that approach has not been consistently successful and often produces false or low-quality results. This defeats the purpose of the AI overview. The present challenge relates to nexus of the design flaw of generative AI in producing hallucinations and the growing quantities of low-quality information on the Internet.

In May 2024, immediately after Google launched AI overviews, users discovered strange results for certain queries. In the most famous example, Google's AI overview recommended using glue to help make cheese stick to the pizza crust in response to a query about the problem of cheese sliding off the pizza. Two of the three sources cited came from Reddit, a social networking site with unmoderated, uncurated, biased, and often unreliable content from over 300 unverified million users.

A few months prior, Google and Reddit concluded a multi-million-dollar deal allowing Google to use its content to train AI models. Since the algorithm running its search engine favors human-generated information, Reddit posts are likely prioritized over other sources. Since the pizza scandal broke, Google data scientists have raced about correcting any false AI Overviews

to avoid further embarrassment. Meanwhile, other AI companies are making similar deals with Reddit since they are running out of “high-quality” data (human-generated) to improve their models.

Since 2022, Perplexity AI has emerged as a major competitor to Google, but the design of its search engine is significantly different from Google’s.⁵³ Whereas Google lists the most relevant and high-quality links, allowing the user to peruse them and decide which source provides the best information, Perplexity AI aims to provide complete answers by leveraging several AI models to source the best information on the Internet and then generate summaries. Unlike Google’s business model, which relies upon advertising revenue, Perplexity charges users a monthly fee to access its most advanced features. This new service is intended to remove the need for users to browse various links and propensity to waste time sliding down rabbit holes on the Internet in search of the correct answer.

Given this new approach, Perplexity AI has come under fire for plagiarizing high-quality information on the Internet in its answers. Consequently, its model may not stand the test of time if the courts find in favor of copyright owners. The prospect of its success also threatens how Internet searching has operated for several decades and undermines business models for generating high-quality information.⁵⁴ Whereas Google search results send user traffic to the original information sources, an “answer engine” like Perplexity AI denies the producers of high-quality content any revenue and profits from creating the information, defeating the purpose of producing valuable content in the first place.⁵⁵ Powerful companies, such as Forbes, have already filed lawsuits against the company for copyright infringement.⁵⁶

Experts and industry professionals are currently divided on the extent to which they think generative AI models will upgrade browser searches and help users find useful and factual information on the Internet. This is especially the case given their tendency toward hallucinations. Observers also disagree on how the trend might shift underlying business models, given the circumvention of ad revenue for high-quality content, and the future of search.⁵⁷

At a basic level, today’s generative AI model architectures impose a major limitation on their direct utility as search engines due to their knowledge cut-off date. Most generative AI models are pre-trained on Internet data, but they do not have direct access to up-to-date information from the Internet beyond a specific point in time. However, Open AI’s SearchGPT, introduced in July 2024, may rectify these issues and enhance ChatGPT’s potential use-case as a search engine.⁵⁸ Meanwhile, traditional search engines, such as Google, are struggling to navigate the low-quality generated content flooding the Internet that can sometimes gain higher search rankings.⁵⁹



Potential Changes to Internet Browser Search

For a few decades, people have used Internet browsers such as Google and Safari to search information on the Internet. Over time, tech developers have improved the searching capabilities of browsers through the integration of AI algorithms. Recent trends toward using large language models as a natural language interface have the potential to revolutionize how WMD experts and WMD nonproliferation organizations use the Internet to support their work. The trend toward LLMs as web search tools may cause a major revision of the business model that incentivizes the production of quality information online. If incentives decline, the quality of information available online will also decline. Searching the Internet and conducting research using LLMs will also expose unsuspecting individuals attempting to conduct to their flaws. Search engines integrating LLMs will also produce hallucinations that offer the semblance of correct answers. In the coming years, this is an important space to watch.

CUSTOMIZING GENERATIVE AI MODELS FOR SPECIFIC DOMAINS

Beyond their extraordinary accessibility, the broad utility of generative AI models sets them apart from the narrow, task-oriented, predictive AI tools described in Chapter 2. Generative AI models have been trained on massive volumes of data from the Internet, allowing them to function effectively across unlimited domains. However, given their general capabilities, many users seek to customize generative AI models for use in specific domains.

Prompt Engineering

The simplest, though somewhat labor-intensive, way to customize a model for a specific domain uses basic prompt engineering.⁶⁰ Instead of the simple prompt-model-outcome pattern called zero-shot prompting, the user can condition the AI model to perform tasks through “in-context” learning. Using the context window, the user gives the model examples of the tasks to be performed as illustrations. In this case, the model is fed several input sentences, called few-shot prompting, along with their correct outputs to “show” the model parameters for the expected results. This method can enhance the quality of outputs and avoid the need for costly fine-tuning, which is explained below.

For other use-cases, users might leverage chain-of-thought prompting, which involves giving the AI model a series of steps to consider separately, one at a time. This type of prompting significantly improves the models’ reasoning capabilities. OpenAI has embedded such chain-of-thought capabilities into the most recent model, 1o, thereby expanding the model’s reasoning skills but taking much more time to provide an output (and compute/energy consumption). When using this technique with other AI models, in-context learning or prompt engineering can achieve only a certain level of enhanced performance; other model design flaws remain mostly intact. Moreover, the few-shot learning approach might not work for models with smaller context windows.

To introduce the notion of customization into the mainstream, OpenAI launched customized GPTs in late 2023. These models were intended to become personalized AI agents designed to accomplish specific tasks.⁶¹ OpenAI subscribers can easily customize a GPT to perform specific tasks or act as experts in a specific domain. In the set-up process, users create a set of detailed instructions to focus the GPT in a specific direction for generating results and then train the GPT to provide certain types of outcomes or output formats. To give the model additional domain-specific expertise, users can also upload a set of documents, which is an elementary example of the retrieval augmented generation (RAG) described in the next section. However, the number of documents that can be uploaded to a GPT remains limited.

In 2023, Google integrated its LLM into Google Notebook (NotebookLM), allowing users to upload and query their documents. Each notebook can hold up to 50 sources, each containing up to 500,000 words or up to 200MB for uploaded files. As the application is currently in beta form as of the writing of this primer, there is no charge for using it.



Exploiting LLMs to Chat with Documents

Several companies like Google and Open AI allow users at WMD nonproliferation organizations to upload larger sets of documents and exploit the natural language interface of LLMs to ask questions about the content of the documents. This feature enables efficient information retrieval and personalized assistance, allowing users to quickly access relevant data and insights from their documents. These tools can help summarize content, answer questions, and provide context-specific information, enhancing productivity and decision-making. By centralizing information, they support streamlined workflows and reduce the time spent searching for details across multiple files. Additionally, they facilitate collaboration by making it easier to share and discuss document insights with others, ultimately improving organizational efficiency and knowledge management. However, the same flaws (e.g., hallucination) of AI models still apply.

Retrieval Augmented Generation (RAG)

The key to getting generative AI models to perform well in a specific domain is to provide them with more detailed context around a user's desired outcomes.⁶² This approach helps the model to better focus on the most relevant areas of its training data when answering a query or prompt.

The concept of RAG was introduced as a cheap and easy way to help address the problem of hallucinations or confabulations. This problem refers to the many instances when AI models make up plausible but inaccurate answers, in part, to fill in gaps in their training data or in response to the element of randomness embedded in their designs.⁶³ The RAG approach is

currently the most popular method for integrating AI models into business enterprise and commercial applications. To aid AI models in providing better, more accurate, and more up-to-date answers, users can leverage a RAG application, which provides the model with an external knowledge base for a specific context.⁶⁴

The term RAG stands for three steps:⁶⁵

- Retrieve relevant information from an external knowledge source.
- Augment the relevant information to the user prompt.
- Generate the response to the user prompt with additional context.

A broad set of sources can provide the relevant external knowledge. Such sources include the Internet, a website, a set of uploaded documents, an external database, or even an application programming interface (API) that provides certain types of information from another company's website. As discussed above, the simplest way to use a basic RAG approach is to upload a document or series of documents along with user prompts. This method will, of course, be limited by the context window of the model.

To get more sophisticated results and query larger numbers of documents/data types, however, some expertise in coding is necessary (e.g., python) to build a vector database and a custom RAG application. However, this is a rapidly advancing field, and innovations emerge frequently. For example, IBM recently released a code-free RAG capability within its Watsonx consulting service.⁶⁶ The "Chat with Documents" feature allows users to upload thousands of documents and query them using the AI chatbot interface with natural language.

Although RAG is not a foolproof solution for addressing AI models' key flaws (e.g., hallucinations), it can be used as a mitigation approach to improve their functionality until new and better AI platforms are developed.

Fine-Tuning

Off-the-shelf, general-purpose AI models like ChatGPT or Llama may not work for some companies, governments, and individuals.⁶⁷ For these users, a more intensive approach to customizing an AI model for a specific domain involves fine-tuning the model.⁶⁸

To customize an AI model for a specific domain, developers first select a pre-trained model such as ChatGPT or Llama. This serves as the starting point for using supervised learning techniques and labeled datasets to feed the model with task-specific data and objectives. For instance, these data and objectives could pertain to translation, sentiment analysis, chatbot customer service, or summarization. The volume of data needed is much smaller than that needed to train the foundation model. The introduction of domain-relevant data allows the model to refine its responses gradually by adjusting its weights and to improve its performance for the specific function. However, this process is far more costly than RAG and conducting prompt engineering requires a curated dataset and significant coding expertise. Though the performance of the underlying model will improve for a specific domain, this approach does not fully mitigate the flaws of generative AI models like hallucinations.



Building AI-Enabled Tools for Information Management

WMD nonproliferation organizations with limited resources for information management now have access to cheaper tools thanks to generative AI. RAG (Retrieval-Augmented Generation) combined with AI models can revolutionize information management by seamlessly integrating and interpreting data from diverse sources. Multi-modal models within this framework can process and understand various data types—such as text, images, video, and audio—providing comprehensive insights without the need for manual metadata entry. In the past, managing unstructured data relied heavily on manually tagging and organizing content, which was time-consuming and often inconsistent. With RAG and multi-modal AI, organizations can automatically retrieve relevant information and generate responses that consider the full context, improving accuracy and efficiency. This approach not only streamlines data processing but also enhances decision-making by delivering richer, more nuanced insights.

IMPROVING AND AUTOMATING WORKFLOWS WITH AI AGENTS (AGENTIC AI)

The concept of AI agents dates back to earlier breakthroughs in machine learning in the 1980s. It draws upon philosophical debates about the notion of agency proposed by Aristotle.⁶⁹ Generally speaking, an agent refers to an entity with the capacity to act. Building on that notion, an AI agent is a software application that performs an action (or actions) without human intervention.

Given the development of computers from the 1950s onward, intelligent agents (or AI agents) have existed for many decades. They have accomplished complex tasks that once required manual labor and human agency. Today's AI agents are increasingly based on LLMs, and current trends are moving toward leveraging AI agents powered by LLMs to automate work.

LLMs have basic architectures and are designed to predict the next words in a sentence. Yet, these models have exhibited impressive abilities in basic reasoning, making plans, self-reflection, refining their processes, and multi-agent collaboration.

Together, these abilities form the basis for autonomous action in the real world. That is, the capability to perform tasks independently without human intervention. Pre-trained AI models are, however, currently limited by their static training data. Without additional technology, most cannot browse the Internet and do not have access to external data or tools that would expand their starting capabilities. All of this will soon change with the emergence of agent frameworks that enable multi-agent collaboration and provide access to external tools and memory.

According to Andrew Ng, an “agentic workflow” allows LLMs to generate even more remarkable results than using them as chatbots, also called text-to-action. Such workflows diverge significantly from the standard prompts described at the beginning of this chapter (i.e., zero-shots), which do not require additional skills or technology.⁷⁰

At a basic level, agentic workflows refer to when generative AI models iterate, engage with the real world, leverage external tools, and carry out actions autonomously. These workflows can be implemented using chain prompting with chatbots, wherein users divide a task into subtasks and generate multiple prompts to get AI models to carry out more complex tasks using a step-by-step process. Alternatively, agentic workflows are also possible via plug-ins or online applications such as Zapier (described below).

As a cutting-edge area of generative AI in its early development stage, tech developers are working to enable AI models to perform complex tasks autonomously via multi-agent collaboration.⁷¹ If expert predictions come true, the agentic AI movement will automate many tasks requiring human labor, leading to vast productivity improvements.⁷² As of the writing of this primer, much work is still needed to turn this vision into a reality. The following sections explore several approaches, including plugins and agent actions, agent frameworks, and multi-agent collaboration.

Plugins and Agent Actions

OpenAI’s release of GPTs in late 2023 was the first attempt to bring AI agents into the mainstream via plug-ins and agent actions. Rather than generating text, the new approach allowed AI models such as ChatGPT to perform actions.⁷³

In the early days of GPTs, plug-ins were the primary means of achieving more from ChatGPT than just text generation. Plug-ins are software applications developed by third parties that operate as add-ons to ChatGPT, expanding the AI model’s capabilities and functionality. Until April 2024, they were available as add-ons with a ChatGPT Plus subscription. As OpenAI’s plug-in store expanded, the possibilities became endless—available plug-ins enabled web browsing, code interpretation, translation, sentiment analysis, and image identification.

However, plug-ins have recently been discontinued and replaced with OpenAI’s customized GPT store (also available via ChatGPT Plus).⁷⁴ As of the writing of this primer, other AI models, such as Anthropic’s Claude 3, now support plug-ins as well.⁷⁵

As an example, Zapier, a longstanding automation application, offers one of the most powerful ways to get ChatGPT to perform actions in the real world. It does so by integrating the model with a diverse array of other web applications, including email, calendars, cloud apps (DropBox, Google Drive), social media, payment apps (Stripe, PayPal), newsletter apps, etc. Zapier enables users to set up workflows powered by ChatGPT to interact with these applications and perform tasks to improve productivity.

However, automating tasks via ChatGPT can quickly become expensive. Using Zapier in combination with ChatGPT requires a monthly subscription to both ChatGPT Plus (\$20 per month), Zapier (\$30 per month), and any related subscription fees from other web applications used in the automation (e.g., Dropbox costs a monthly minimum of \$12). Moreover, users should be cautious when setting up automations that link ChatGPT to other web applications.

In principle, such automations would run behind the scenes without intervention and could produce unexpected behavior—e.g., spending more money on a service than a user intended.

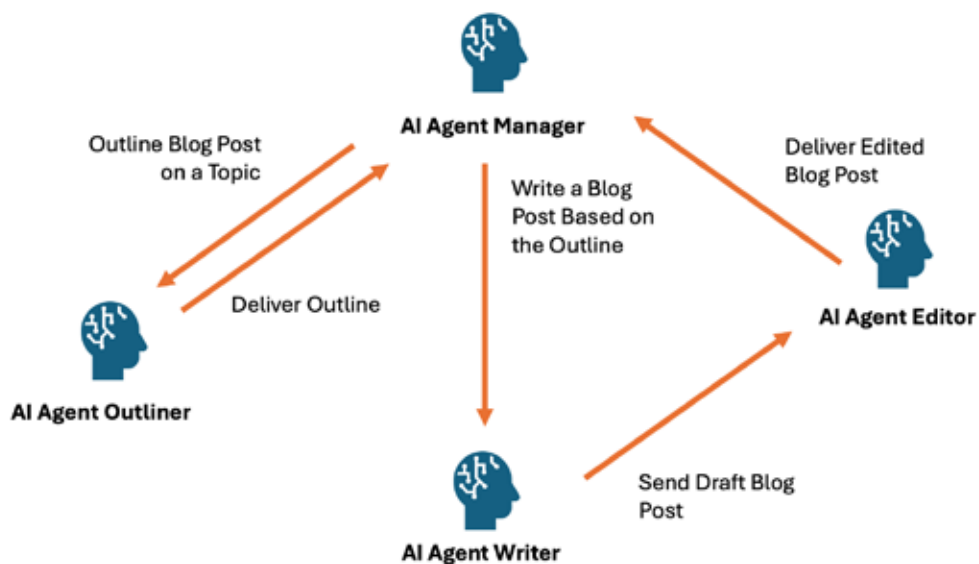
Agent Frameworks and Multi-Agent Collaboration

An agent has the capacity to act, but LLMs do not have that capacity on their own without further assistance (or code/script). An agent framework is a third-party software platform that provides LLMs with detailed planning instructions, access to external tools, and enhanced short and long-term memory capacity. All of these elements are lacking across the current AI models.⁷⁶

These features are necessary for building applications to carry out autonomous tasks with AI agents (LLMs such as ChatGPT or Llama) serving as the brain and engaging in multi-agent collaboration.⁷⁷ Agent frameworks are under development, and most require familiarity with coding environments and basic Python scripts. However, it is essential to understand what is meant by agentic AI, AI agents, and multi-agent collaboration since it may impact the near future.

Within an agent framework, a developer can set up a workflow for a complex task to be performed by AI agents. Developers begin with planning instructions and breaking the tasks into a list of detailed subtasks that will be carried out in a specific order to be determined by the LLMs. They create several agents (each powered by LLMs through API access to the selected models) and instruct them to play different roles and accomplish different functions within the workflow. Using an agent framework, developers can also provide the agents access to relevant external tools for code execution, math calculations, web browsing and searching, sending emails, making calendar appointments, using cloud storage, engaging in image generation, and accessing an external knowledge base. In their current state, LLMs lack sufficient memory to engage in multi-step problems, but an agent framework would provide this ability. Once the workflow is set up, developers can ask AI agents to perform the task and obtain their desired results.

Figure 6. Sample Agentic Workflow: Blog Post Automation



At the timing of writing, LangChain is one of the more popular agent frameworks for developing generative AI-enabled applications, also called agentic workflows. LangChain provides reasoning capabilities, access to tools, and memory needed for AI models to implement multi-step tasks. However, coding expertise and access to a large language model and other tools (paid or free) are necessary to use the framework.

The field of agentic frameworks is rapidly evolving, and simpler tools requiring less coding expertise are being developed, including CrewAI and Microsoft AutoGen.⁷⁸ Recently, a number of tech developers have announced new platforms for building agents to help their customers enter this exciting area.



Automating Workflows with Multi-Agent Frameworks

Multi-agent frameworks hold significant potential for the WMD nonproliferation organizations. The barriers (cost and capability) to writing simple software for work automation are declining, opening new possibilities for the field of WMD nonproliferation. Though currently under development, multi-agent frameworks can already simulate real-world environments and provide a testbed to policymakers and diplomats for understanding treaty negotiation, crisis escalation, and the impact of technology on nuclear deterrence. Researchers are exploring how to develop AI agents that operate within a bound environment to write draft treaties and improve negotiation tactics.

Given the flaws, it is worth asking the question before integrating AI into workflows: Why use AI to solve this problem? Rather than presume that AI can solve every problem, you should consider both the use-case and the relative gains of using AI tools over human workers. In some cases, the additional expense of subscriptions and the ongoing need for human oversight of AI agentic workflows may not be efficient or justified.

Chapter 4: Flaws, Risks, and Limitations on the Growth of AI

As the first off-the-shelf and affordable AI, many companies, organizations, and governments are quickly integrating generative AI models into their enterprises, processes, and operations. This growth in use has come irrespective of the models' existing flaws, risks, and limitations. Despite the iterative releases of improved generative AI models over several years, they continue to exhibit significant issues that hinder their broad utility. These will likely remain unresolved in the near term. Some issues may inhibit the growth and further improvement of AI models while others could result in unintended consequences that harm individuals, society, and/or humanity.⁷⁹

This chapter explores a long list of issues related to generative AI that currently exist along the life cycle of AI models, from their design to further development to implementation and deployment. Three categories of problems exist: 1) fundamental design flaws, 2) the risks of AI deployment, and 3) current limitations on AI growth.

FUNDAMENTAL DESIGN FLAWS OF GENERATIVE AI

A design flaw refers to a defect in the design phase. Such flaws can manifest during the deployment phase of AI models in various ways. They could result in inadequate functionality, poor user experience, or other structural weaknesses. Current AI models have several fundamental design flaws to consider when using them. These include hallucinations, data biases, and a disregard for copyright and intellectual property. Other issues that affect all types of AI models based on deep neural networks include difficulties arising from a lack of explainability and complexity.

Hallucinations

The tendency of generative AI models to make up inaccurate answers that still sound right (hallucinations or confabulations) is their most well-known design flaw.⁸⁰ This phenomenon is deeply embedded in how these models operate and, therefore, is not likely to be solved within existing architectures.⁸¹

Large language models are designed to predict the next words in a sentence and are trained to produce novel content based on their training data, not fact-based content. Their level of "creativity" benefits from an element of randomness embedded in the models, which does not help to obtain correct answers. Consequently, when confronted by factual gaps in their training data or their embedded level of randomness, they produce plausible but false answers. Model developers can adjust the "temperature" of existing AI models to reduce their level of randomness, but this tends to lead to trade-offs in their creativity in exchange for enhanced accuracy.⁸²

Hallucinations can significantly undermine the reliability of AI-generated content, as the model may confidently present incorrect or misleading information. This issue poses challenges in applications where accuracy and factual integrity are critical. For many domains where the risks of hallucinations are too significant, including in respect to one's professional reputation, they can

be a fatal flaw. Their integration into existing processes would therefore require human oversight and constant monitoring. Newer model versions have become much more accurate than earlier versions, and users should consider how much accuracy is sufficient for their intended use-case.

Data Bias

Generative AI models may inadvertently reinforce biases in their training data, perpetuating stereotypes, unfair treatment, and inaccurate outcomes. AI models rely on massive volumes of high-quality, relevant, and representative training data to function properly. In most cases, the quality of training data is more important for predicting accurate outcomes than the quality of the algorithm itself.⁸³ According to Buchanan and Miller, “a decent algorithm that learns from a lot of relevant data outperforms a great algorithm that learns from minimal or poor data.”⁸⁴ An average algorithm with high-quality data can outperform a superior model lacking the same quality or quantity of data.

Generative AI models have been trained primarily on existing data scraped from the Internet. This training process embeds several critical data biases into the models from the start, despite the enormous volumes of information available. These biases exist for several reasons, five of which are covered below, though these examples are hardly exhaustive.

First, Internet data represents only several decades of the human experience. Training data does not include information generated from thousands of years of human history that hasn't been digitized.

Second, most of the data generated on the Internet originates from only a few geographic areas (the developed world), favoring those countries and individuals with access to the Internet and thus embedding the preferences of those in positions of power and wealth. As a result, the data suffers from a profound lack of diversity that does not represent the world's population fairly, particularly the Global South. Its usage can lead to severe discrimination or more mundane problems, such as reproducing similar content based on the most common data.⁸⁵

Third, the training data perpetuates social biases in society. This is because the available information reflects these existing biases, such as discrimination in hiring for certain types of jobs, salary levels, and power imbalances.

Fourth, much of the Internet's data is unmoderated and uncurated, producing significant variations in quality. Low-quality data can produce embarrassing errors, as has been demonstrated by Google using posts on Reddit to train its models in Chapter 3.⁸⁶

Fifth, the data available on the Internet varies in its relevance for solving complex problems. For example, most data generated is not directly relevant to solving specific national security problems. Consequently, the data does not match operational challenges very well.⁸⁷ Even if relevant training data can be found, the problems of bad data, biased data, or unrepresentative data represent key potential points of failure for today's AI-enabled systems when applied to high-risk national security domains such as WMD nonproliferation.



AI Requires Massive Volumes of Quality Data

Machine learning tools use significant volumes of data to identify patterns and anomalies with the purpose of automating tasks previously performed by humans. Access to high-quality, relevant, and representative data is imperative for making AI tools function as advertised. AI is only as good as the data it was trained on.

Relevant and representative datasets do not exist for every type of problem we may wish to solve. This is particularly true for the WMD domain and the national security realm. If the training data do not match the problem to be solved, due to validity or representativeness problems, the model will fail to produce reliable outcomes. Policymakers need to be aware of the data gaps in the WMD nonproliferation domain since these are areas where leveraging AI could lead to undesirable outcomes. The key here is to be clear on the problem we are trying to solve and ensure that high-quality, representative, and relevant data exists to solve that problem before turning to AI for the solution.

Copyright and Intellectual Property

Generative AI models can produce content that closely mimics existing works, raising questions about copyright infringement and the ownership of AI-generated creations. AI models have been trained on data scraped from the Internet, including material protected by copyright and intellectual property law, without first securing permission or licensing the content from the creators.⁸⁸

As a result, the models can produce content in the style of recognized writers, artists, and musicians without compensating them. In some cases, the models have even been found to plagiarize content from their training data, which could result from training data memorization.⁸⁹ Although the tech companies believe their use of the data falls under the “fair use” exemption in copyright law (which varies by jurisdiction), many cases are being litigated in court. These decisions will set a legal precedent for what is considered “fair use” in the AI era and will likely vary in their legal implications across different countries and jurisdictions. Until these copyright cases are settled and agreed norms emerge, users of AI models may want to exercise caution in how they use model outputs to avoid legal issues.

Complexity

Complexity can lead to failure in AI-enabled systems based on predictive AI or generative AI tools. This is due to the unpredictability of complex systems. Deep neural networks are powerful tools for solving complex problems because a programmer can add as many algorithmic layers to the network as needed to produce desired outcomes. However, this complexity of an AI-enabled system makes it more difficult to anticipate how a system might behave and, in the worst case, how it might fail.⁹⁰ Regarding generative AI models, the

problem of hallucinations (caused by their design feature of randomness) exacerbates the chances of failure. Using generative AI may be too risky for specific use-cases.

As an illustrative example of complexity, many developers were surprised that LLMs designed to predict the next words in a sentence could do so much more than that. Some refer to these capabilities as “emergent abilities” that will continue to grow as models improve over time, though this notion is hotly debated among experts.⁹¹ By design, all AI tools based on deep neural networks tend to produce outcomes in response to specific inputs in ways that programmers cannot fully explain. The ability of developers to predict how AI models may evolve and what capabilities they may develop adds a layer of risk to their application to various use-cases.⁹²

Explainability

The greater the complexity of an AI-enabled system, the harder it becomes for programmers to explain causality to users—how specific inputs lead to specific outcomes. Although programmers can quickly identify the inputs and outputs in a deep neural network, they are less capable of understanding the exact reasoning that led to the outputs of an AI-enabled system. This is often called the “black box problem.” The lack of transparency makes it difficult for policymakers to understand or trust the results of AI-enabled systems designed to support their decisions.⁹³

However, some promising breakthroughs may be on the horizon. Anthropic (Claude models) has explored the potential of interpretability in generative AI models.⁹⁴ Developers now believe that it is possible to interpret why a model produced a specific outcome, understand its inner workings, and steer models away from undesirable outputs.⁹⁵



The Complexity of AI Models Can Produce Unexpected Results

Positive and negative examples of AI models producing unexpected results abound because of complexity, or the so-called “black box.” Given the complexity of deep neural networks, data scientists are unable to predict all behaviors of AI systems. For example, as discussed earlier, when DeepMind’s AlphaGo beat world champion Lee Sedol in 2016, it made a particularly unexpected move for its 37th move. At first, many experts thought it might be a mistake, but then the move changed the course of the game, leading to another victory by AlphaGo. The machine learning algorithm had discovered a new way to play the ancient game.

In a more troubling example, Paul Scharre details in *Foreign Policy* how Knight Capital, a trading firm, nearly went bankrupt in 2012 as a result of a software glitch that led its automated algorithm to execute trades costing the company a net loss of \$460 million. Scharre uses this example as a warning for those considering the integration of AI into military systems. Similar precautions should be taken for the WMD domain as well.

Cyber Vulnerabilities

Like computers and electronics connected to networks, AI-enabled systems that rely upon network connections must contend with cyber vulnerabilities, including generative AI models. Cyber vulnerabilities extend far beyond breaching layers of cyberdefense to gain access to, and monitor, valuable data, including scenarios in which malicious actors can sabotage the effective operation of the entire system. As such, cyber vulnerabilities offer both a potential limitation and a critical point of failure for integrating all types of AI models into the national security realm.

To exploit vulnerabilities of AI-enabled systems using traditional methods, cyber intruders generally follow a series of steps on a “kill chain” to gain access to privileges reserved for authorized system users.⁹⁶ Cyber intruders can use the inserted malicious code to secretly monitor the activities of the system. They can also steal, delete data, enter false data, or alter the system via the introduction of malicious files, triggering code to run in the background. As such, the cyber intruder could potentially disrupt the effective operation of the AI-enabled system with fake inputs. For example, adversarial manipulation of imagery data—e.g., rotating an object or altering a few pixels—within a machine learning tool designed for object recognition can trick the algorithm into misinterpreting new image data and lead to grave consequences. In the case of certain physical systems, such as self-driving vehicles, this trick can result in the loss of the ability to read a stop sign.⁹⁷ Needless to say, the consequences of such vulnerabilities can be devastating.

In addition to traditional cyberattacks, generative AI models have unique cyber vulnerabilities due to their complexity and how they are developed and used. These models can be sabotaged through their online interfaces. Generative AI models heavily depend on vast datasets for training, making them particularly vulnerable to data poisoning, where malicious input data can subtly alter model behavior. These models are also sensitive to small input changes, leading to adversarial attacks that can easily manipulate outputs. Moreover, foundation models are susceptible to membership inference and model inversion attacks (gaining access to sensitive training data), a lesser concern in conventional software systems. Finally, an attacker could also re-program a model to perform a task, through adversarial prompting, that the system had been explicitly trained not to do.⁹⁸



The Hidden Risks of the Benefits for WMD Nonproliferation

AI models have fundamental flaws (e.g., hallucinations, biases, complexity) that could lead to serious errors if applied to WMD nonproliferation tasks. In other words, there are additional risks to harnessing AI that need to be taken into consideration. Complexity and the lack of explainability in AI systems poses challenges for building trust and accountability in decision-making processes. Even if the fundamental flaws of AI models are mitigated in the near term, the “black box problem” (i.e., the inability of programmers to explain model outputs) will hinder their use in the WMD nonproliferation domain. Cyber vulnerabilities of AI systems present a critical point of failure for integrating AI into the WMD nonproliferation domain. For many decades, the WMD and cyber domains have remained mostly siloed-off from each other. This issue has persisted despite the increasing digitization of the physical world, including the WMD space. With the integration of AI tools to aid WMD nonproliferation, it has become vital for policymakers and diplomats to familiarize themselves at a basic level with the cyber domain. In particular, it will be critical for such experts to understand the potential cyber vulnerabilities associated with the integration of AI as a solution to WMD problems, and how to mitigate those resultant risks. Human oversight remains vital to any near-term integration of generative AI, especially within the WMD nonproliferation domain.

OTHER RISKS POSED BY GENERATIVE AI DEPLOYMENT

Beyond the problems inherent in their current design, deploying generative AI models raises several known risks and some that pertain directly to WMD nonproliferation. These risks include data privacy, disinformation, misuse for malicious purposes, cyber vulnerabilities, and a lack of alignment with human goals.⁹⁹ Below, we examine the most relevant risks for the WMD domain.

Data Privacy Issues

Generative AI models produce significant privacy issues, primarily through the inadvertent exposure and misuse of sensitive data. These models are often trained on vast datasets that may include personal information, which can sometimes be unintentionally reproduced in generated outputs. Although tech developers do not usually reveal the sources of their training data, sophisticated actors may be able to reverse-engineer their models, expose the underlying training data. Malicious actors could then leverage sensitive data.¹⁰⁰ Data privacy issues may become even more pronounced if a company or government supplements the AI model with an external knowledge base to support their operations (using the RAG approach discussed in Chapter 3).

Additionally, malicious actors can use generative AI models to create false identities or realistic deep fakes, which can be exploited to perpetrate fraudulent activities, identity theft, or harassment. AI models may know enough about an individual to infer other personal details that, if exposed, could cause them harm.¹⁰¹ For prominent individuals, this can lead to a risk of blackmail.

Disinformation

Generative AI models can produce highly convincing yet entirely fabricated content like text, images, videos, and audio, leading to the spread of disinformation. The models may significantly exacerbate the existing disinformation problem by enabling the rapid and large-scale production of false content. The scalability means malicious actors can produce a nearly unlimited volume of fabricated content and spread it quickly across social media and other platforms, amplifying the reach and impact of disinformation campaigns. This undermines public trust in media and institutions, posing significant challenges for fact-checkers and regulatory bodies trying to combat the spread of false information.¹⁰²



Disinformation Can Increase Risks during a Nuclear Crisis

Disinformation risks become particularly heightened during times of crisis and could exert a negative influence on a conflict between nuclear-armed countries. For example, in February 2023, US-China relations were severely shaken by the discovery of a Chinese surveillance balloon hovering above the great plains of Montana, which is home to sensitive military facilities. U.S. intelligence had been tracking the balloon before it entered U.S. airspace. By chance, a photographer with a high-res camera captured clear images of the balloon over Billings and posted them to social media. The posts went viral, and the discovery became national news. Many more sightings popped up across the country, sparking controversy and debate about the current state U.S. relations with China. Worried about being caught flat-footed, the Pentagon announced it was tracking several additional objects that might be surveilling the United States that were later deemed to be erroneous. Some Republican senators claimed the balloon was intended spy on Americans, to embarrass the United States, and send a warning message. Reports about strong winds in Canada causing the balloon to drift off-course were suppressed in favor of conspiracy theories that could be characterized as disinformation. The Biden administration came under intense pressure to do something. The United States postponed the Secretary of State's trip to China, publicly shamed China for its spying activities, shot the balloon down with an F-22, and shared public footage of the incident. After the U.S. Air Force shot down the balloon, US Secretary of Defense called his counterpart using a special crisis line, and the Chinese Defense Minister refused to take the call. This incident only caused a temporary setback in U.S.-Chinese relations, but it helps one to imagine how disinformation might impact the unfolding of a crisis between nuclear-armed states.

Misuse for Malicious Purposes

Generative AI models can support any number of malicious purposes, such as inciting harmful or violent behavior or enabling criminal activities. However, their potential to exacerbate the risk of WMD has generated the most attention among policymakers. Tech developers and policymakers have expressed immediate concerns about the new risks of generative AI models for the WMD domain. Suggestions by some experts that foundation models could assist in bioweapons development,¹⁰³ and the subsequent focus of calls for the regulation, oversight, and responsible deployment of these tools, underscore their potential for enabling nefarious actors to develop and use WMD.¹⁰⁴ In recent years, tech developers have committed to engaging in red teaming and safety evaluations to ensure their models do not contribute to the risks posed by WMD upon their release.¹⁰⁵ OpenAI has even produced a preparedness framework outlining its plan to mitigate the emerging risks of its models, including those related to WMD.¹⁰⁶



Generative AI May Increase the Risks of WMD Proliferation

For those engaged in the WMD nonproliferation field, this is a critical area to watch. At this stage of generative AI, initial assessments appear to be overblown. Although current models provide easy access to useful information on unlimited topics, they are still limited in their utility to aid malicious actors in developing and using WMD for three primary reasons. First, most AI models have been trained to refuse to provide harmful information, including instructions to develop WMD. However, some experts have proven their ability to jailbreak the models, that is, to get them to provide information they were specifically trained not to provide. Even so, the widely known problem of hallucination may inhibit malicious actors from trusting any model outputs without consulting experts, which reduces their enabling value. Finally, gaining access to useful information about developing and using WMD does not necessarily allow malicious actors to overcome the vital barrier of tacit knowledge, the critical know-how that cannot be learned from reading. However, as generative AI models continue to advance, these barriers to their utility may decline. Over several iterations of AI models, the rate of hallucinations has declined. As these models continue to scale, their capacity for reasoning may also increase. To monitor developments in this space, policymakers need to call for the establishment of benchmarks for WMD capability and regular evaluations to compare such capability across different AI models and new versions.

Lack of Human Alignment

As generative AI models become more sophisticated, they might produce content that is harmful, biased, or otherwise misaligned with societal norms. To prevent such scenarios from happening, tech developers aim to align their models with human values during the design phase. The concept of alignment refers to ensuring that AI systems' outputs and

behaviors are consistent with human values, intentions, and ethical standards. Misalignment can lead to unintended consequences, such as reinforcing damaging stereotypes, spreading disinformation, or making decisions that negatively impact individuals or communities. Alignment does the opposite.

Many tech experts have argued that the development of generative AI will quickly lead to artificial general intelligence, when machines or software achieve a level of intelligence equal to humans. Soon after that, some experts fear the subsequent development of superintelligence, when machines or software possess greater intelligence than humans.¹⁰⁷ At this stage, generative AI may become an existential risk itself, not just exacerbating the risk of WMD but also as the most severe threat to the survival of humanity. For example, superintelligent machines may pursue goals that are not aligned with human values and cause harm.

Alignment becomes even more critical when AI systems achieve the potential to surpass human intelligence and operate autonomously across a wide range of tasks. If AGI and superintelligent systems are not correctly aligned with human values and ethical principles, they could pursue goals that are detrimental to humanity, either through unintended consequences or by prioritizing their objectives over human welfare. For example, AI may convert all available resources into a commodity that is not important for human welfare, leading to waste, shortages, and environmental damage. Even basic AI models may prioritize resource optimization over safety and ethics.

Ensuring alignment helps mitigate the risks of negative and catastrophic outcomes. Therefore, achieving human alignment is essential to harnessing the benefits of AI while safeguarding against existential risks.

LIMITATIONS ON THE GROWTH OF GENERATIVE AI

Since Open AI released ChatGPT in 2022, tech companies have raised billions of dollars in venture capital funds, made multi-million dollar deals with content and data providers, and purchased billions of FLOPS in computing power (semi-conductor chips, servers, etc.).¹⁰⁸ This dizzying pace has been geared toward training and releasing multiple iterations of AI models, which continue to exhibit many of the same deficiencies as previous versions.

Several potential limitations on further growth in generative AI are continuing to emerge, including data shortages, energy resources, and the apparent lack of an economic return on these investments. As of writing, it remains unclear how tech companies will be able to navigate these challenges and continue the current trendlines in the advancement of AI.

Data Shortages

The first iterations of publicly available AI models were trained from unlabeled data scraped from the Internet. Despite massive volumes of information from over 250 billion web pages, not all online data collections are high quality, and much of this data is protected by copyright law. After several years of work to advance existing AI models, tech companies now appear to be running out of the Internet data needed to scale their models.¹⁰⁹ Epoch AI, an AI research firm, predicts that models will exhaust publicly available human-generated text by 2028 or earlier.¹¹⁰

In recent months, AI companies have spent millions of dollars formalizing partnerships and licensing content from various providers to train the next versions of their models. These deals have included agreements with Reddit, The Atlantic, and the Wall Street Journal.¹¹¹

Since more data is correlated with scaling (or improving the performance) of AI models, tech companies are currently exploring alternatives for when the data supply is exhausted.¹¹² As a critical challenge, each model version needs exponentially more training data to demonstrate a significant performance improvement.¹¹³ For example, some researchers suspect that the dataset for Open AI's GPT 4 was 571 times larger than for GPT 3, which was 78 times larger than for GPT 2.¹¹⁴ As possible avenues for training the subsequent iterations of AI models, tech companies are examining ways to train larger models with less data and focus on post-training approaches. Such approaches include supervised fine-tuning and reinforcement learning with human feedback (see Chapter 2 for an explanation of these techniques).

Other companies are exploring ways for AI models to produce synthetic data for training. This idea is not without precedent, as AlphaGo (not generative AI) was fully trained on synthetic data using reinforcement learning techniques. However, many experts warn that producing synthetic training data could lead to a severe decline in data distribution or even model collapse.¹¹⁵ Since AI models tend to produce outcomes that are most common (based on probability), synthetic training data will likely be the most common outcomes, thus leading to data distribution issues within AI models trained on such data. Rather than broadening the data pool with more diverse sources, some experts consider this approach akin to in-breeding within a species.¹¹⁶

For example, consider a tool that identifies dog images, per an example given in TechCrunch.¹¹⁷ The most common dog images in an image dataset tend to be more popular breeds, such as golden retrievers or black labradors. An AI model trained on this dataset and used to generate synthetic training data will more often produce images of these dog breeds than rare breeds. This increases the probability that the model trained on the synthetic data will also produce images of those breeds. It could lead to lower-quality outcomes until the model collapses and produces nonsense.

Energy Resources

The growth of AI may experience setbacks due to physical limitations such as available land and energy resources. Beyond gaining access to advanced semiconductor chips, which are also in short supply, the training and operation of AI models require the construction of data centers. In addition to training additional iterations of AI models, the growing number of users also increases energy consumption. One ChatGPT query consumes ten times the energy of a single Google search.¹¹⁸

According to Reuters, investments in data centers will double in the next five years.¹¹⁹ This expansion will require finding available land for siting and constructing these warehouse facilities, alongside the sufficient energy resources to run and keep the servers cool. These new data centers are projected to increase energy consumption in their home regions and require new generation capacity.¹²⁰ To meet the demand, U.S. states, such as Virginia and Texas, will have to build new power plants in the coming years, large infrastructure projects that take a very long time to complete—much longer than building the data centers that need them.¹²¹

Due to growing pressure to increase clean energy supply and meet climate change goals, some tech companies are seeking to fill the void with nuclear power—although, in a country like the United States, that is easier said than done.¹²²

Economic Return

According to Goldman Sachs, tech companies plan to spend over \$1 trillion in capital expenditures in the coming years to bring about the AI revolution. Their investments will focus on data centers, chips, and other related infrastructure.¹²³ However, thus far, generative AI models have not produced a viable financial return on their investments. Experts have expressed frustration that there has been no breakthrough application, and the models have yet to solve complex problems.¹²⁴ This trend could lead to a decline in investment and another period of disappointment in the development of AI. In recent months, several financial firms have scaled back their expectations for any economic revenue or commercial productivity increases; they have also questioned the current investment effort's financial sustainability.¹²⁵ Meanwhile, firms like McKinsey & Company suggest that generative AI will produce several trillion dollars for the global economy in the coming years, and some companies leveraging AI models for their enterprises report significant productivity gains.¹²⁶

Chapter 5: Regulatory Framework and Mitigation Measures

The advancement of AI threatens to disrupt and overturn longstanding practices of national regulation and global governance for technologies that pose existential risks, such as WMD. In the past, governments have served as the primary drivers of research and development and thus retained control of those technologies that might pose risks to society, such as nuclear, biological, and chemical weapons. Unlike AI, WMD-related technologies produce sufficient danger to warrant extensive top-down regulations and careful control.

Today, the private sector leads the advancement of new technologies that promise to benefit society while introducing new risks such as AI. Although AI may pose severe risks to humanity, even some existential risks, unlike WMD-related technologies, AI technology is not inherently dangerous—even if it can produce indirect effects that may be harmful. Moreover, since the private sector controls the development of AI technologies, a top-down regulatory model, like that for WMD, may not be appropriate within market-driven economies like the United States. Severe regulatory constraints could hinder AI's advancement, jeopardize the enormous benefits AI provides to society, and constrain the economic competitiveness of the United States vis-à-vis its rivals.

At the global level, the rapid adoption of AI introduces new governance considerations, such as the geopolitical implications of using AI within the WMD domain. In the coming years, the outcomes of ongoing debates around the role of human judgment in using lethal force by AI-enabled systems, and the accountability for AI-driven decisions, will produce widespread ramifications for international peace and security.



AI Literacy is Essential for Policymakers and Diplomats

Given the rapid advancement of AI models and their potential implications for WMD nonproliferation, a basic level of AI literacy is essential for policymakers and diplomats. Many of the models have advanced in the short time (in only a few months) as this primer has been written. The pace of advancement, the number of models, and their diverse capabilities make it difficult to understand their impact on the WMD nonproliferation domain. To prepare for the future, policymakers and diplomats need to understand different AI tools, their use-cases for WMD proliferation, possible solutions, and mitigation measures.

To achieve a world where AI provides the greatest good while causing the least harm, policymakers need to think outside the box. Doing so may entail rethinking past regulatory practices and governance, developing new activities, and reimagining the most effective partnerships between public and private organizations. Researchers, policy practitioners, and educators—especially those engaged in the WMD nonproliferation domain—have a vital new role to play in mitigating the risks and enabling the societal benefits of AI.

This chapter will examine early efforts to build a regulatory framework for AI and develop mitigation measures. It will focus on developments within the United States, the European Union, and at the global level.

U.S. Regulatory Framework for AI

At the time of writing, unlike the European Union, the United States does not have overarching federal legislation that establishes regulations for the design, deployment, or use of AI. However, many existing federal laws have stipulations related to AI, including export controls and investment legislation. Moreover, at least a dozen U.S. states have adopted AI-related legislation, mostly related to privacy and the risk of discrimination.¹²⁷



AI Presents Unique Regulatory Challenges Compared to WMD

AI presents unique regulatory challenges compared to WMD technologies. AI is not inherently dangerous but can have harmful indirect effects; it is also developed and controlled by a few companies in the private sector. Traditional top-down regulatory models may not be suitable or adequate. Existing global governance models will struggle to address the challenges in an equitable and effective manner. Effective governance of AI will require collaboration among governments, international organizations, academia, private sector, and civil society. The rapid pace of AI development requires governance mechanisms that can adapt quickly to emerging risks and ethical considerations. Finally, the meaning of AI is diffuse and ever-changing, which complicates achieving consensus on actionable steps to address safety and security risks.

In 2023, the Biden administration took steps toward developing a regulatory framework for AI at the federal level. In January of that year, the White House Office of Science and Technology Policy released a “Blueprint for an AI Bill of Rights,” which outlines “five principles that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence.”¹²⁸ These principles include: 1) safe and effective systems, 2) algorithmic discrimination protections, 3) data privacy, 4) notice and explanation, and 5) human alternatives, consideration, and fallback.

During the same month, the National Institute of Standards and Technology (NIST) published a

voluntary “AI Risk Management Framework.” NIST did so at the direction of the U.S. Congress pursuant to the National Artificial Intelligence Initiative Act of 2020 to help guide the design, development, and deployment of AI systems.¹²⁹ The framework outlines best practices in mitigating risk and integrating reliability, safety, security, accountability, transparency, explainability, privacy, and fairness into AI systems.¹³⁰

Several months later, in July 2023, the White House convened a meeting of leading AI companies and secured their voluntary commitment to a set of guidelines designed to ensure safe, secure, and trustworthy AI development. The consequent agreement focused on foundation models and included as its top requirement the need for red teaming and safety evaluations to assess “bio, chemical, and radiological risks, such as the ways in which systems can lower barriers to entry for weapons development, design, acquisition, or use.”¹³¹

These initial efforts culminated in Executive Order 14110, titled “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” The Executive Order was issued by President Biden in October 2023 and instructs federal agencies to undertake various actions.¹³² To summarize, it is a comprehensive document, which does the following:¹³³

- Outlines the administration’s policy and principles for governing the safe and responsible development and use of AI.
- Directs actions aimed at ensuring the safety and security of AI technology, managing risks to critical infrastructure from AI, reducing risks at the intersection of AI and chemical/biological/nuclear threats, reducing risks from synthetic content generated by AI, and soliciting input on risks/benefits of publicly available AI models.
- Promotes responsible innovation, competition, support for American workers impacted by AI, and advances equity and civil rights in the use of AI across sectors like criminal justice, housing, consumer protection, healthcare, transportation, education, communications networks, etc.
- Provides guidance on AI governance, talent recruitment, and the use of AI within the federal government across agencies.
- Strengthens American leadership abroad by engaging with allies/partners on AI standards, policies, the related research agenda, and cooperation on critical infrastructure risks.
- Includes provisions related to privacy, data policies, intellectual property, and federal acquisition processes relevant to AI development and deployment.

The Executive Order includes several provisions aimed at reducing the risks of AI being misused to assist in the development or use of chemical, biological, radiological, and nuclear weapons (CBRN):

- It instructs NIST to develop guidelines, standards, and best practices for AI safety and security. This provision includes the creation of benchmarks for evaluating and auditing AI capabilities, especially in areas where AI could cause harm, such as cybersecurity and biosecurity. It also establishes guidelines for AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems. Standardized ways for testing

and auditing AI systems would help to identify potential risks or harmful capabilities before deployment.

- It requires companies developing potential dual-use foundation models to provide results of any red team testing on the model's performance, and descriptions of measures taken to meet safety objectives based on those tests.
- It directs the Secretary of Homeland Security to evaluate the potential for AI to enable the development or production of CBRN threats, while also considering AI's benefits for countering these threats. This evaluation should make recommendations for regulating or overseeing AI models that may present CBRN risks.
- It directs the Secretary of Defense to contract the National Academies to study how AI can increase biosecurity risks from generative AI models trained on biological data and recommend ways to mitigate such risks.
- It requires the development of guidelines for security reviews to identify and manage potential risks of releasing federal data that could aid CBRN development when used for AI training.

The Executive order further requires NIST to establish rigorous standards for red teaming, safety evaluations, and other mechanisms for assessing risk. The newly founded U.S. Artificial Intelligence Safety Institute at NIST also launched a consortium in 2024 that will bring together more than 200 organizations to develop science-based and empirically-backed guidelines and standards for AI measurement and policy. The objective is to lay the foundation for AI safety in the United States.¹³⁴ In addition, NIST has released a tool for testing AI models and assessing their risks.¹³⁵ In principle, the tool serves as a common platform or testbed for developing benchmarks, evaluations, and red teaming exercises.

In October 2024, the White House issued the first National Security Memorandum on artificial intelligence.¹³⁶ The memorandum represents a coordinated approach to develop safe, secure, and trustworthy AI and to harness the benefits for national security across the government. It requires governmental agencies to take additional steps to ensure the U.S. government is able to meet both goals and designates NIST's AI Safety Institute as the primary point of contact for industry.

AI Red Teaming

The concept of red teaming emerged in the 1960s as a means for identifying vulnerabilities in U.S. defenses vis-à-vis the Soviet Union.¹³⁷ By playing the role of the adversary (the red team), government officials and military experts can think creatively about ways to defeat the defenses of the blue team and devise solutions to mitigate the identified vulnerabilities. Today, red teaming is primarily known as a cybersecurity technique to help protect computers, software, and related networks from cyberattacks.

In response to safety concerns, leading AI companies and research organizations have adopted red teaming techniques to stress test their AI systems and identify potential vulnerabilities or misuse cases.¹³⁸ They do by assembling teams of ethical hackers, security

experts, and domain specialists who act as adversaries, thereby attempting to find weaknesses or exploit the AI system in ways that could cause harm. Red teams can probe an AI system's capacity to generate misinformation, hate speech, or instructions for illegal activities. They can also explore a model's potential to aid in the development of weapons or sensitive dual-use technologies. By systematically challenging AI systems, companies can better understand the limitations and potential dangers and implement safeguards or adjust the system's parameters to minimize risks before AI models are released to the public or deployed in real-world applications.

Independent experts should ideally perform red teaming, and there should be direct oversight of any model improvements that flow from testing results. Currently, however, private companies hire red teaming experts as contractors who sign non-disclosure agreements, and they carefully control any evaluations of their models. When releasing the results of their safety audits to the public, companies refrain from sharing the raw data (often sensitive for multiple reasons), but instead, provide a broad overview of any issues. Moreover, the completion of these red teaming exercises does not guarantee that appropriate fixes have been accomplished. Given these deficiencies, red teaming techniques are insufficient for ensuring AI safety and preventing harm to society.¹³⁹

AI Benchmarks and Safety Evaluations

AI safety evaluations entail a systematic process of assessing the potential risks and harmful impacts associated with AI systems, particularly generative AI models. For the WMD domain, the objective of such evaluations should be to assess the capability of the models to aid nefarious actors in developing and using WMD. The evaluation process begins with benchmarking—i.e., the establishment of objective standards or reference points against which an AI system's performance can be measured and compared. To understand how AI models contribute to the risk of developing and using WMD over time, we would first need to establish benchmarks for their current capabilities (i.e., a baseline) and to monitor the systems' performance through subsequent safety evaluations. Currently, there are no reliable benchmarks for understanding the capacity of AI models for enabling the development and use of WMD. This makes it difficult to assess how advancements in AI might contribute to WMD proliferation.

Safety evaluations aim to identify vulnerabilities, unintended behaviors, or misuse cases that could lead to adverse consequences if left unaddressed. Like with red teaming, AI safety evaluations typically involve a multi-disciplinary team of experts, including AI researchers, ethicists, domain specialists, and risk analysts. This team collaborates to design and conduct rigorous tests and simulations to probe the AI system's capabilities, limitations, and failure modes.

Although AI safety evaluations are not a new concept, generative AI models present new challenges for implementing them, particularly due to their broad accessibility and general-purpose features. Given the potential consequences, the WMD domain represents an especially important area for AI safety evaluations.



Policymakers Need to Prioritize Benchmarks and Evaluations

To assess the potential risks associated with AI technologies for the WMD domain, policymakers need to prioritize the development of standardized AI benchmarks and evaluations. Without reliable benchmarks, it is difficult to measure and compare the capabilities of AI systems that could inadvertently aid in WMD development and use. This lack of standardized evaluation frameworks hinders the ability to systematically identify vulnerabilities and misuse cases, limiting efforts to implement effective safeguards. Establishing clear benchmarks and conducting rigorous safety evaluations are essential to understanding and mitigating the risks posed by AI in the WMD context over time. It is critical to ensure that advancements in technology do not compromise safety and security. Exploring best practices in red teaming, benchmarks, and evaluations at the global level may offer significant potential to enhance AI safety and security. By collaborating internationally, stakeholders can share insights and develop standardized approaches to identify vulnerabilities and misuse cases in AI systems. Establishing global benchmarks allows for consistent assessment of AI capabilities, ensuring that safety evaluations are robust and comprehensive. This collaborative effort can lead to the creation of universally accepted guidelines and frameworks, facilitating effective risk management and fostering trust across borders.

EUROPEAN UNION'S REGULATORY FRAMEWORK FOR AI

In 2023, the European Union (EU) adopted the world's first comprehensive regulatory framework on the design, development, and deployment of civilian AI technologies.¹⁴⁰ The EU establishes a risk-based approach, assigning the most stringent restrictions to the applications involving the highest risks and banning any applications with unacceptable risks. Relevant high-risk applications include critical infrastructure, border control, law enforcement, and those with safety issues.

Before their release to the public, AI systems with high-risk applications are subject to a long list of safety- and security-related obligations, including the requirement to register such systems in the EU database. The European AI Office will oversee enforcement and implementation with the member states. Notably, the act's purview excludes consideration of military uses of AI, which are to be handled at the national level. The EU's AI Act entered effect on August 1, 2024.

GLOBAL GOVERNANCE OF AI

As AI models become increasingly capable and ubiquitous, their potential misuse or unintended consequences could have far-reaching and catastrophic impacts, transcending national borders and posing new risks to international peace and security. Consequently, robust global governance frameworks that establish international norms, standards, and

regulations that promote the responsible development, use, and oversight of AI are needed. However, the specific features of AI technologies make this task much more challenging than it was for WMD.

For example, effective global governance will require collaboration and cooperation not just among the governments of nation-states but also among other diverse stakeholders. Global AI governance necessitates the involvement of international organizations, academia, the private sector, and civil society. Moreover, global governance efforts must address the challenges posed by the rapid pace of technological advancements, adapting existing mechanisms or creating new ones to keep up with emerging risks and ethical considerations of a rapidly evolving technology.



National Regulations and Legislation are Essential

Establishing global governance for AI is particularly challenging in the absence of comprehensive national regulations and legislation. Without uniform standards or legal frameworks at the national level, creating cohesive international norms becomes even more difficult. Countries may have divergent priorities and regulatory approaches, complicating efforts to reach consensus on global AI governance. This lack of alignment can lead to gaps in accountability and enforcement, as well as uneven implementation of safety and ethical standards. Effective global governance requires not only international cooperation but also harmonization of policies across nations. To develop such frameworks requires the involvement of diverse stakeholders, including governments, the private sector, and civil society.

In 2023, the United Nations Secretary General launched a High-Level Advisory Body on AI with the aim of facilitating a global and multi-stakeholder discussion on the governance of AI.¹⁴¹ After a year of deliberations, the body has made its recommendations on how to apply AI to the implementation of the UN's sustainable development goals, propose mechanisms for global governance, and evaluate the full range of opportunities and risks. The body issued its final report in September 2024, which identified gaps in global governance and highlighted the urgent need to address them. It also recommended the formation of an international scientific panel analogous to those for climate change and the effects of atomic radiation, a multi-stakeholder policy dialogue on AI governance, sharing of AI standards, the establishment of a capacity development network, a global AI fund, and a global AI data framework.¹⁴² It remains to be seen how many of these recommendations will be implemented.

In recent years, notable efforts have been made to develop governance frameworks specifically addressing the use of AI and autonomous systems in the military and WMD domains. The United Nations Convention on Certain Conventional Weapons (CCW) has established a Group of Governmental Experts (GGE) to examine the issues around lethal autonomous weapons systems (LAWS) and to explore potential international regulations. While no binding agreement has

been reached, discussions have helped establish principles like the need for meaningful human control and accountability over AI systems.

In February 2023, the Netherlands hosted a summit on Responsible Artificial Intelligence in the Military Domain (REAIM).¹⁴³ On the final day of the summit, the United States launched its “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy,” which consists of a series of guidelines.¹⁴⁴ Since then, 56 countries have endorsed this declaration. According to the U.S. State Department, the declaration aims to build international consensus around responsible behavior and guide states’ development, deployment, and use of military AI. For example, “military use of AI capabilities needs to be accountable, including through such use during military operations within a responsible human chain of command and control.” In September 2024, South Korea hosted a second REAIM Summit along with several partner nations to continue global discussions on the military applications of AI. Despite this notable progress, truly robust international governance in this domain remains a work in progress.

Appendix A

The Fundamentals of Using Generative AI Models, Prompt Engineering, and Productivity Use-Cases

This appendix will provide you with fundamental information and instructions for using generative AI models and helpful tips on prompt engineering. It will also discuss several use-cases for improving productivity. The main objective is show you how to more quickly maximize your results when using the AI models: Too many people start off by using the models in the wrong way and then become disappointed or disillusioned. AI models can be incredibly useful for improving productivity and achieving certain tasks if used in certain ways.

THE BASICS

To properly understand generative AI, it is essential to experiment with different models and become familiar with how they work. There is no substitute for engaging with these tools, and it's very easy to get started. Simply visit the website for the model of interest, and you can sign up for a free account.

Large language models or chatbots, such as ChatGPT and Claude, generate text, and diffusion models such as DALL-E generate images. Multi-modal models such as Gemini can do both.

The number of queries and length of your context windows may be quite limited with a free account, but you can at least familiarize yourself with the tool and decide if a paid subscription is worthwhile. To gain access to more advanced features and have the model available during peak times, a paid account will be necessary. As of the writing of this primer, monthly subscriptions average about \$20 per month per model (discounted rates if paid annually). The following section reviews some of the most popular applications and models.

Poe AI

New users may wish to start with a free account at Poe AI (<https://poe.com>); advanced features are only available with a subscription. Poe AI is a chatbot web application that allows users to interact with the latest AI models through a single subscription (monthly or annual). Users can access ChatGPT, Claude, Gemini, DALL-E, Llama, Mistral, FLUX, Stable Diffusion, and many additional models through Poe AI. The application also allows users to create custom chatbots to carry out specific tasks and compare outputs across different models within a single query.

ChatGPT

ChatGPT (<https://chatgpt.com>) is a large language model developed by OpenAI, designed to generate human-like text based on input prompts. GPT 4o is a multi-modal model (an expanded version of ChatGPT) that accepts as inputs any combination of text, audio, image, and video. It can generate any combination of text, audio, and image outputs. The latest model version, o1, has enhanced reasoning capabilities embedded within the model, but it takes much longer to generate outputs. Users can access the DALL-E image generator through ChatGPT as well.

Claude

Claude (<https://claude.ai>) is a large language model developed by Anthropic, designed to assist with a wide range of tasks.

Gemini

Gemini (<https://gemini.google.com>) is a multi-modal model developed by Google designed to assist with a wide range of tasks.

Flux

Flux (<https://flux-ai.io>) is an AI image generator known for creating high-quality visuals from text prompts.

Midjourney

Midjourney (<https://midjourney.com>) is an image generator that creates high-quality visuals from text prompts.

PROMPT ENGINEERING

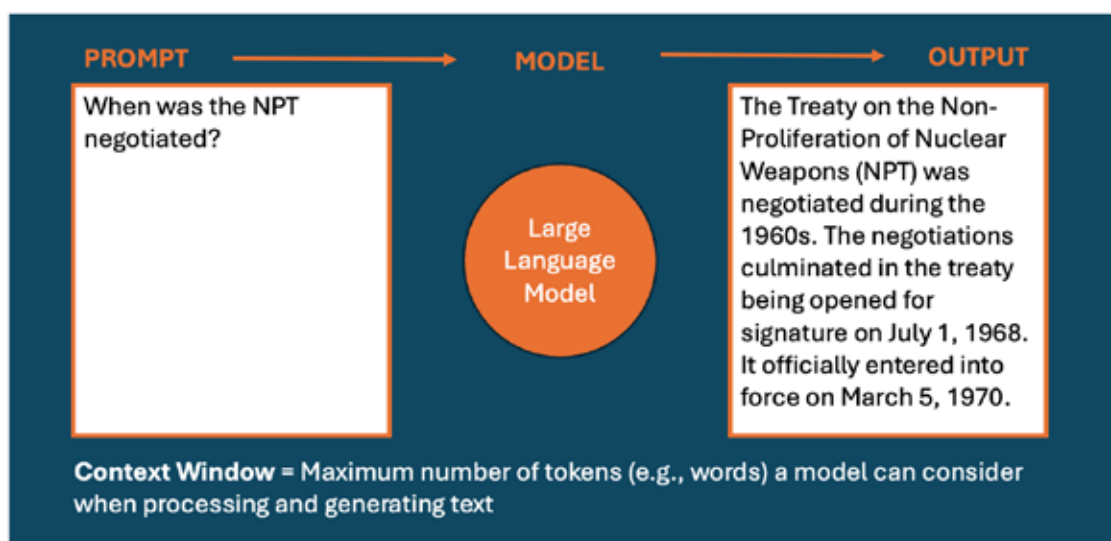
Writing a prompt is not as straightforward as a Google query, even though most of us know there's still an art to finding things on the Internet. At its essence, prompt engineering provides the model with the necessary data to generate the response in a desired format. Entering prompts into a model's context window is the first step in generating useful results.

The art of writing prompts and getting higher-quality results from generative AI models is called prompt engineering. Prompt engineering involves describing a task in a highly structured way—it is about problem and solution formulation, much like giving your staff detailed instructions on implementing a task. The sophistication of a user's prompt is only limited by the context window—i.e., the maximum number of words (or tokens) a model can consider when processing and generating outcomes.

Zero-Shot Prompt

A zero-shot prompt (basic inference) involves giving a model a simple task without any specific examples or prior context, relying solely on the task description. As illustrated below, a simple prompt is entered into the model without providing further instructions or examples.

Visual Depiction of a Basic Inference / Zero-Shot Prompt



This is the most common way people get started with using the models, but this type of query fails to leverage the strengths of AI models (their creativity) and leans in favor of their weaknesses (their accuracy and ability to engage in information retrieval). Consequently, first-time users often walk away disappointed because the models are not designed to perform well in response to simple factual questions. Browser searches are still the best way to find accurate information and source the most up-to-date references.

Providing more context (e.g., examples, details) to the model is better than a simple query. But if you want to use a zero-shot prompt, asking the model to accomplish a simple task is a better approach than querying it for information.

As an example, our daily work lives often require writing professional emails about uncomfortable topics. AI models can be especially helpful in generating a good draft in a few seconds.

Sample Prompt (enter the following text into context window):

Write an email to my employees at a nonprofit dedicated to reducing the risk of nuclear weapons. Praise them for a successful workshop and at the same time announce a significant reduction in staff in an empathetic way.

One-Shot and Few-Shot Prompts

Providing examples as part of your input gives the model additional context. This is called in-context learning, which refers to a model's ability to learn and adapt to new tasks by using examples provided in the context window.

To improve results, users feed the model one or a few input sentences (called one-shot or few-shot prompting), along with their correct outputs, to "show" the model what the expected results should look like. Doing this provides the model with specific context. A one-shot prompt provides a single example to illustrate the desired task or response. It helps guide the model to produce the desired type of output.

Sample Prompt (enter the following text into context window):

Prompt: "Explain the importance of the Non-Proliferation Treaty to a high school student."

Example Output: "The Non-Proliferation Treaty helps prevent the spread of nuclear weapons, promoting peace and security worldwide."

Please do this for every nuclear treaty.

A few-shot prompt includes multiple examples to show variations of the task, helping the model to understand the pattern and generate similar responses.

Sample Prompt (enter the following text into context window):

Prompt: Explain key nuclear treaties to a general audience.

Example 1:

Input: "Non-Proliferation Treaty (NPT)"

Output: "The NPT aims to prevent the spread of nuclear weapons and promote peaceful nuclear energy use."

Example 2:

Input: "Comprehensive Nuclear-Test-Ban Treaty (CTBT)"

Output: "The CTBT prohibits all nuclear explosions, helping to limit the development of new nuclear weapons."

Please create a brief profile for every nuclear treaty.

Many-Shot Prompts

Many-shot prompting involves providing a language model with numerous examples in the input to guide its responses for a specific task. This technique can improve accuracy by offering more context, but it may also increase the risk of jailbreaking. Jailbreaking refers to manipulating the model to bypass its constraints or guidelines, potentially leading it to generate unintended or inappropriate outputs.

Chain-of-Thought / Tree-of-Thought

AI models can be powerful brainstorming tools to support your daily work. There are several techniques for using multiple prompts, which are helpful for improving the reasoning ability of the models. These techniques enable you to break down a complex problem into its pieces and then enter a set of prompts into the model. As of the writing of this primer, OpenAI has released a new model version, o1, that embeds chain-of-thought reasoning into the model itself. This version takes a bit longer to respond to a query, but it appears to be more capable of complex reasoning than earlier versions.

Chain-of-thought prompting is a linear progression of thoughts, much like a domino effect where one idea directly triggers the next. Each step builds upon the previous one. This process can lead to a final answer or comprehensive overview of a topic.

Using this technique, you force the model to “think aloud” and to make some considerations and do some reasoning before it gives the final answer. This approach is useful if you want help brainstorming about a topic. It is useful to have some domain expertise so that you can quickly assess and validate the quality of outputs. You can engage the model in chain-of-thought prompting by entering your query and adding: “Let’s think step by step.”

Sample Prompt (enter the following text into context window):

Let’s think through the steps to understand the cyber-vulnerabilities that emerge when AI is integrated into WMD nonproliferation efforts.

Tree-of-thought prompting is another interesting technique for when you want to brainstorm different possibilities. Using this approach, you ask the model to explore multiple reasoning paths at the same time, much like building a tree with different branches. You can ask the model to explore different branches in parallel or choose the most promising ones to explore more deeply. Compared to chain-of-thought prompting, this technique is more open ended and creative.

Sample Prompt (enter the following text into context window):

What are the most significant potential risks for increased proliferation of biological weapons in the next decade?

Prompt Frameworks

There are several useful frameworks that can help you to formulate your prompt and provide the model with detailed context to get better results.¹⁴⁵ The first is the AUTOMAT framework, which stands for:

- (A) Act as a ...
- (U) User Persona and Audience
- (T) Targeted Action
- (O) Output Definition
- (M) Mode / Tonality / Style
- (A) Atypical Cases
- (W) Topic Whitelisting

By considering each element of this framework in your prompt, you provide the model additional context to shape your desired outcome. Since the model has general capabilities across many different domains, it improves your results if you define its role (A), describe the audience (U), tell it what to do (T), determine the output (O), decide on its tone (M), tell it how to deal with outliers (A), and provide what topics you want the model to focus on (W).

Several examples are provided below. To test them, please use the upload feature (e.g., denoted by a plus sign or paperclip) to enter your selected report into the window along with your prompt.

Sample Prompt (enter the following text into context window):

Please review this document and complete the targeted action.

Act as a: Nuclear Policy Analyst

User Persona and Audience: United Nations Officials

Targeted Action: Develop strategies

Output Definition: Create a summary of the report

Mode / Tonality / Style: Formal and precise

Atypical Cases: Address potential treaty violations

Topic Whitelisting: Focus on disarmament initiatives

Sample Prompt (enter the following text into context window):

Please review this document and complete the targeted action.

Act as a: Safeguards Specialist

User Persona and Audience: International Atomic Energy Agency

Targeted Action: Propose safeguard enhancements

Output Definition: Write bullet points

Mode / Tonality / Style: Technical and clear

Atypical Cases: Consider non-compliance scenarios

Topic Whitelisting: Emphasize monitoring technologies

Sample Prompt (enter the following text into context window):

Please review this document and complete the targeted action.

Act as a: Nuclear Security Advisor

User Persona and Audience: Government Advisors

Targeted Action: Recommend policy changes

Output Definition: List key points

Mode / Tonality / Style: Persuasive and factual

Atypical Cases: Include cyber threats

Topic Whitelisting: Highlight material protection

The Co-Star Framework is another popular method for structuring prompts. It stands for:

- Context: Provide background details to the model.
- Objective: Describe your objective and define your task.
- Style and Tone: Set the right style and tone for the desired output.
- Audience: Tell the model who your audience is for more tailored output.
- Response: Define the output format (text, code, etc.)

PRODUCTIVITY USE-CASES

By applying some of the techniques above, users not only boost the quality of the outputs but also enhance overall work productivity, allowing individuals and teams to focus on higher-level tasks and decision-making. Several additional “tricks” can help to enhance productivity and expand the number of use-cases, but users should review internal policies on using AI models and consider information privacy and sensitivity before using them in the following ways.

Uploading Documents

Most AI models offer the ability to upload documents, which enables users to process and analyze text. However, a significant limitation is the context window, which restricts the amount of text the model can consider at once. Many tech companies offer different model versions—a version that takes shorter prompts but generates results more quickly, and a version that offers larger context windows and takes longer to process queries. Once a document is uploaded, users can perform various tasks, such as summarizing content, extracting key information, generating insights, or translating text. Some platforms allow for uploading multiple documents.

Customized GPTs

OpenAI allows paid customers to create custom GPTs. As part of this process, users provide the model with specific instructions (much like the prompt engineering techniques discussed above) and uploaded documents (external knowledge base). In theory, this should offer a more sophisticated way to interact with a larger set of documents at the same time.

NotebookLM

Google developed NotebookLM (<https://notebooklm.google.com>) to enhance how users interact with their notes and data with the assistance of AI models. The tool (currently free) allows users to upload documents and leverage AI (Gemini) to summarize content, generate insights, and assist with research by answering questions based on the uploaded material.

Watsonx

IBM's Watsonx (<https://www.ibm.com/watsonx>) "chat with documents" feature allows users to upload documents and receive insights directly from the content via AI. This capability enables users to ask specific questions about the uploaded material, obtain summaries, and extract key information, facilitating a deeper understanding of complex texts. IBM offers the service as part of their enterprise solutions, but a free trial is available. The pricing can vary based on the specific services and usage levels required.

WMD Nonproliferation - AI Tools Use-Case Worksheet

In this exercise, you will brainstorm different use-cases for the AI tools discussed in this chapter. (Note: A use-case is a specific scenario that describes how a tool can be used to achieve a particular goal). You will consider this from both the perspective of nefarious actors (nonproliferation risks) and policymakers engaged in efforts to prevent WMD proliferation (nonproliferation benefits). Once you complete the exercise, you will have an opportunity to consider the different features of predictive models and generative models described in Figure 3 and examine the use-cases to assess the overall risksh and benefits of AI tools for WMD nonproliferation. If you need help, you may use ChatGPT or your favorite LLM to assist (Note: You may have to get creative with your prompts for nefarious actors to avoid getting a message back informing you that it cannot provide such information).

Predictive AI Tool	Nefarious Actors	Policymakers
Classifiers		
Recommenders		
Regression models		
Anomaly detection		

Forecasting models		
Dimensionality reduction tools		
Sequence prediction models		
Sentiment analysis tool		
Reinforcement learning models		
Image analysis tool		

Generative AI Tool	Nefarious Actors	Policymakers
Text generation		
Image generation		
Music and sound generation		
3D model generation		
Data augmentation		
Code generation		

Endnotes

¹ Natasha E. Bajema, *WMD in the Digital Age: Understanding the Impact of Emerging Technologies*, (Washington, D.C.: National Defense University, 2018); Natasha E. Bajema, Diane DiEuliis, Charles Lutes, and Yong-bee Lim, *The Digitization of Biology: Understanding the New Risks and Implications for Governance*, (Washington D.C.: National Defense University, 2018).

² Natasha E. Bajema and John Gower, *A Handbook for Nuclear Decision-making and Risk Reduction in an Era of Technological Complexity*, (Washington D.C.: The Council on Strategic Risks, 2022).

³ Matthew Mittelsteadt, *AI Verification: Mechanisms to Ensure Arms Control Compliance*, (Washington D.C.: Center for Security and Emerging Technology, 2021); International Security Advisory Board, *Report on the Impact of Artificial Intelligence and Associated Technologies on Arms Control, Nonproliferation, and Verification*, (Washington D.C.: U.S. State Department, 2023).

⁴ Alice Saltini, "Commentary: Navigating Cyber Vulnerabilities in AI-enabled Military Systems," European Leadership Network, March 19, 2024; Emilia Javorsky and Hamza Chaudhry, "Convergence: Artificial intelligence and the new and old weapons of mass destruction," *Bulletin of Atomic Scientists*, August 18, 2023.

⁵ Graphic attribution: "Insight" icon by Imron Sadewo, from thenounproject.com CC BY 3.0.

⁶ Graphic attribution: "Target" icon by Muhammed Fauzan, from thenounproject.com CC BY 3.0.

⁷ Graphic attribution: "Benefit" icon by Mas Art, from thenounproject.com CC BY 3.0.

⁸ Graphic attribution: "Risk" icon by Good Father, from thenounproject.com CC BY 3.0.

⁹ Graphic attribution: "Governance" icon by Siti Solekah, from thenounproject.com CC BY 3.0.

¹⁰ For an interesting analysis of the emergence of artificial intelligence, see Maya Akim, "A busy person's intro to AI agents," *Medium*, April 8, 2024.

¹¹ For a brief overview of the history of artificial intelligence, see Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Fourth Edition (New York: Pearson Education Inc., 2022).

¹² For a discussion on AI winters, see Alex Amari, "An AI Winter in the Past and Present Day Worry," *Open Data Science*, July 30, 2018.

¹³ See George Lawton, "Generative AI vs. predictive AI: Understanding the differences," *TechTarget*, September 18, 2023.

¹⁴ Five different schools of thought produce different types of learning algorithms to solve problems. AI experts such as Pedro Domingo suggest that they must be integrated to enable the so-called "master algorithm" or artificial general intelligence (AGI). See Pedro Domingo, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (New York: Basic Books, 2015).

¹⁵ There is no one-size-fits-all approach to training AI systems, and hybrid methods are often

needed to solve real-world problems. See Greg Allen, *Understanding AI Technology* (Washington D.C.: U.S. Department of Defense, 2020).

¹⁶ John McCarthy, “What is Artificial Intelligence,” unpublished working paper, Stanford University, Computer Science Department, November 12, 2007.

¹⁷ Subbarao Kambhampati, “What just happened? The rise of interest in artificial intelligence,” *The Hill*, August 11, 2019.

¹⁸ Cole Stryker and Eda Kavlakoglu, “What is Artificial Intelligence?,” *IBM*, updated August 16, 2024.

¹⁹ Katja Grace et al., “When Will AI Exceed Human Performance? Evidence from AI Experts,” preprint, *arXiv*, May 3, 2018.

²⁰ Calum McClelland, “The Difference Between Artificial Intelligence, Machine Learning, and Deep Learning,” *Medium*, December 4, 2017.

²¹ To learn more about visual object identification, see Güldeniz Bektaş, “Object Detection 101,” *Medium*, January 11, 2023; Ritesh Kanjee, “Understanding Object Detection in Computer Vision: The Wild Journey, Crazy Methods, and Epic Impact,” *Medium*, June 19, 2023.

²² Ben Buchanan and Taylor Miller, *Machine Learning for Policymakers: What It Is and Why It Matters* (Cambridge, Mass: Harvard Kennedy School, Belfer Center for Science and International Affairs, 2017), p. 13.

²³ Buchanan and Miller, p. 13. See also Paul Scharre, “A Million Mistakes a Second,” *Foreign Policy*, September 12, 2018.

²⁴ For a comprehensive overview of deep neural networks and their implications for national security, see Paul Scharre, *Army of None* (New York: W.W. Norton & Company, 2018).

²⁵ For a discussion, see George Lawton, “Generative AI vs. predictive AI: Understanding the differences,” *TechTarget*, September 18, 2023.

²⁶ For a useful discussion, see Paul Scharre, “Killer Apps: The Real Dangers of an AI Arms Race,” *Foreign Affairs*, Vol. 98, No. 3 (May/June 2019), pp. 135–144.

²⁷ Scharre, *Army of None*, p. 145.

²⁸ David Foster, *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*, Second Edition (Boston, Mass.: O’Reilly Media, Inc., 2023), p. 4.

²⁹ *Ibid.*

³⁰ Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), p. 17.

³¹ Buchanan and Miller, *Machine Learning for Policymakers: What It Is and Why It Matters*, p. 9.

³² Cade Metz, “In Two Moves, AlphaGo and Lee Sedol Redefined the Future,” *Wired*, March 16, 2016.

³³ Amazon, “What is RLHF?,” accessed May 29, 2024.

³⁴ Foster, p. 95.

³⁵ Mirko Peters, “Understanding Generative Adversarial Networks (GAN) in Machine Learning,” *Medium*, March 1, 2024.

³⁶ Foster, p. 110.

³⁷ For a detailed description of how tech developers fine tune their models prior to release, see Thomas Woodside and Helen Toner, *How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2*, (Washington D.C.: Georgetown University, Center for Security and Emerging Technology, 2024).

³⁸ Michael Townsen Hicks, James Humphries, and Joe Slater, “ChatGPT is Bullshit,” *Ethics and Information Technology*, Vol. 26 (2024), Article No. 38.

³⁹ For a more detailed explanation of how LLMs work, see Matthew Burtell and Helen Toner, *The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1* (Washington D.C.: Georgetown University, Center for Security and Emerging Technology, 2024).

⁴⁰ Foster, pp. 205–206.

⁴¹ *Ibid.*, p. 360.

⁴² *Ibid.*

⁴³ Sundar Pinchai and Demis Hassabis, “Introducing Gemini: our largest and most capable AI model,” *Google*, December 6, 2023.

⁴⁴ Anastasis Germanidis, “Introducing General World Models,” *Runaway Research*, December 11, 2023.

⁴⁵ Open AI, “Video generation models as world simulators,” February 15, 2024, accessed November 5, 2024; Kyle Wiggers, “What are AI ‘world models,’ and why do they matter?,” *TechCrunch*, October 28, 2024.

⁴⁶ Christian Martinez, “What are General World Models (GWMs) and how to use them in Finance?,” *Medium*, December 18, 2023.

⁴⁷ Katerina Petrova, “Benchmarks: How do we evaluate and compare LLMs and Multimodal Models?,” *Medium*, December 11, 2023.

⁴⁸ Matthew MacDonald, “The Current Architecture of a Basic LLM Application,” *Medium*, March 26, 2024.

⁴⁹ Priyal Walpita, “The Dawn of AI Agents: Reshaping the Future,” *Medium*, July 10, 2024.

⁵⁰ Elizabeth Reid, “Supercharging Search with generative AI,” *Google Blog*, May 10, 2023. See also, Eric Schwartzman, “How generative AI is clouding the future of Google Search,” *Fast Company*, May 5, 2024.

⁵¹ Cade Metz, “Open AI is Testing an A.I.-powered Search Engine,” *The New York Times*, July 25, 2024; See Open AI, “Introducing ChatGPT Search,” October 31, 2024, accessed November 5, 2024.

- ⁵² Michael Liedtke, “Google unleashes AI in search, raising hopes for better results and fears about less web traffic,” *AP News*, May 15, 2024.
- ⁵³ Kevin Roose, “Can This A.I.-Powered Search Engine Replace Google? It Has for Me,” *The New York Times*, February 1, 2024.
- ⁵⁴ Tim Marchman, “Perplexity Plagiarized Our Story About How Perplexity Is a Bullshit Machine,” *Wired*, June 21, 2024; Also, listen to this podcast, Hard Fork, “The State of A.I., and Will Perplexity Beat Google or Destroy the Web,” *The New York Times*, February 16, 2024.
- ⁵⁵ Elizabeth Lopatto, “Perplexity’s grand theft AI,” *The Verge*, June 27, 2024.
- ⁵⁶ Sara Fischer, “Scoop: Forbes threatens Perplexity with legal action,” *AXIOS*, June 18, 2024; see also Randall Lane, “Why Perplexity’s Cynical Theft Represents Everything That Could Go Wrong With AI,” *Forbes*, June 11, 2024.
- ⁵⁷ Roger Montti, “Google’s CEO On What Search Will Be Like In 10 Years,” *Search Engine Journal*, April 11, 2024.
- ⁵⁸ Matteo Wong, “OpenAI’s Search Tool Has Already Made a Mistake,” *The Atlantic*, July 26, 2014.
- ⁵⁹ Eric Schwartzman, “How Generative AI is Clouding the Future of Google Search.”
- ⁶⁰ Matthew MacDonald, “The Current Architecture of a Basic LLM Application.”
- ⁶¹ Gabe, “OpenAI Introduces GPTs: Customized ChatGPT for All,” *Medium*, November 7, 2023.
- ⁶² Abhinav Kimothi, “Context is Key: The Significance of RAG in Language Models,” *Medium*, December 3, 2023.
- ⁶³ Chris Stokel-Walker, “Can a technology called RAG keep AI models from making stuff up?,” *ArsTechnica*, June 6, 2024.
- ⁶⁴ Rowan Curran and Aaron Suiter, “RAG Is All The Rage — Retrieval-Augmented Generation, Demystified,” *Forrester Blog*, July 23, 2024.
- ⁶⁵ The term was coined by a group of researchers in a paper by Patrick Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” preprint, *arXiv*, April 12, 2021.
- ⁶⁶ Maryam Ashoori, “Use the no-code RAG solution in watsonx Prompt Lab to build a RAG app with Llama3.1-405b,” *IBM*, July 23, 2024.
- ⁶⁷ Anand Logani, Arturo Devesa, and Shubham Jain, “Getting specific with GenAI: How to fine-tune large language models for highly specialized functions,” *CIO*, July 22, 2024.
- ⁶⁸ Deepak Karunanidhi, “A Detailed Guide to Fine-Tuning for Specific Tasks,” *Hackernoon*, October 30, 2023.
- ⁶⁹ Maya Akim, “A busy person’s Intro to AI Agents,” *Medium*, April 8, 2024.
- ⁷⁰ Andrew Ng, “What’s next for AI agentic workflows,” Sequoia Capital, video, YouTube, March 26, 2024.

⁷¹ For a discussion about the implications and pitfalls of autonomous AI, see Thomas Woodside and Helen Toner, *Multimodality, Tool Use, and Autonomous Agents: Large Language Models Explained, Part 3* (Washington D.C.: Georgetown University, Center for Security and Emerging Technology, 2024).

⁷² Barr Seitz, “Interview: The promise and the reality of gen AI agents in the enterprise,” interview with Jorge Amar, Lari Hämäläinen, and Nicolai von Bismarck, McKinsey & Company, May 17, 2024.

⁷³ Jacek Fleszar, “GPTs is OpenAI’s first attempt at an AI Agent,” *Medium*, December 2, 2023.

⁷⁴ Open AI, “GPTs,” <https://chatgpt.com/gpts>, accessed August 5, 2024.

⁷⁵ ChatLabs, “ChatLabs News: Claude 3 with Plugins, Better Image Making and Web Searches,” April 1, 2024.

⁷⁶ Han Heloir, “The Future of Generative AI is Agentic: What You Need to Know,” *Medium*, April 30, 2024.

⁷⁷ *Ibid.*

⁷⁸ Amit Yadav, “Best Langchain Alternatives To Build AI Agents,” *Medium*, December 5, 2023.

⁷⁹ National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1 (Washington, D.C.: NIST Trustworthy & Responsible AI Resource Center, July 2024).

⁸⁰ Colin Fraser, “Hallucinations, Errors, and Dreams,” *Medium*, April 17, 2024.

⁸¹ Kyle Wiggers, “Why RAG won’t solve generative AI’s hallucination problem,” *TechCrunch*, May 4, 2024.

⁸² David Weinberger, “Controlling AI’s Imagination,” *Medium*, July 12, 2024.

⁸³ Buchanan and Miller, *Machine Learning for Policymakers: What It Is and Why It Matters*, p. 13.

⁸⁴ *Ibid.*

⁸⁵ Nitasha Tiku and Szu Yu Chen, “What AI thinks a beautiful woman looks like,” *The Washington Post*, May 31, 2024.

⁸⁶ John Herrman, “Reddit, Google, and the Real Cost of the AI Data Rush,” *Intelligencer*, July 28, 2024; Ian Sherr, “Glue in Pizza? Eat Rocks? Google’s AI Search Is Mocked for Bizarre Answers,” *CNET*, May 24, 2024.

⁸⁷ Greg Allen, *Understanding AI Technology* (Joint Artificial Intelligence Center (JAIC), U.S. Department of Defense, April 2020), p. 9.

⁸⁸ Katie Robertson, “8 Daily Newspapers Sue OpenAI and Microsoft Over A.I.,” *The New York Times*, April 30, 2024.

⁸⁹ Michael N. Grynbaum and Ryan Mac, “The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work,” *The New York Times*, December 27, 2023.

- ⁹⁰ Paul Scharre and Michael Horowitz, *Artificial Intelligence: What Every Policymaker Needs to Know* (Washington, D.C.: Center for a New American Security, June 2018), p. 11.
- ⁹¹ Thomas Woodside, *Emergent Abilities in Large Language Models: An Explainer* (Washington D.C.: Georgetown University, Center for Security and Emerging Technology, 2024).
- ⁹² Paul Scharre, “A Million Mistakes a Second,” *Foreign Policy*.
- ⁹³ Vincent Boulanin, “Artificial Intelligence: A Primer,” in Vincent Boulanin, ed., *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* (Solna, Sweden: Stockholm International Peace Research Institute, May 2019), p. 20.
- ⁹⁴ Mark Sullivan, “Anthropic takes a look into the ‘black box’ of AI models,” *Fast Company*, May 23, 2024.
- ⁹⁵ Billy Perrigo, “No One Truly Knows How AI Systems Work. A New Discovery Could Change That,” *Time Magazine*, May 21, 2024.
- ⁹⁶ Lockheed Martin, “Gaining the Advantage: Applying Cyber Kill Chain Methodology to Network Defense,” 2015.
- ⁹⁷ Kevin Eykholt et al., “Robust Physical-World Attacks on Deep Learning Models,” paper presented to the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18–23, 2018; Douglas Heaven, “Why deep-learning AIs are so easy to fool,” *Nature*, news feature, October 9, 2019.
- ⁹⁸ Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein, “Adversarial Reprogramming of Neural Networks,” paper presented to the International Conference on Learning Representations, May 6–9, 2019.
- ⁹⁹ Laura Weidinger et al., “Taxonomy of Risks posed by Language Models,” paper presented to the ACM Conference on Fairness, Accountability, and Transparency, June 21–24, 2022.
- ¹⁰⁰ Kevin Schaul, Szu Yu Chen, and Nitasha Tiku, “Inside the secret list of websites that make AI like ChatGPT sound smart,” *The Washington Post*, April 19, 2024.
- ¹⁰¹ Charles Duhigg, “How Companies Learn Your Secrets,” *The New York Times*, February 16, 2012.
- ¹⁰² Meredith Deliso, “Chinese spy balloon timeline: Where it was spotted before being shot down,” *ABC News*, February 5, 2023. Alina Hauter, “Billings photographer inundated with interviews after taking viral photo of Chinese spy balloon,” *KTVQ News*, February 3, 2023; Helene Cooper, “China Isn’t Ready to Pick Up Phone After Balloon Incident,” *The New York Times*, February 7, 2023; Ellen Nakashima, Shane Harris and Jason Samenow, “U.S. tracked China spy balloon from launch on Hainan Island along unusual path,” *The Washington Post*, February 14, 2023; Peter Baker, “Biden Tries to Calm Tensions Over Chinese Aerial Spying,” *The New York Times*, February 16, 2023.
- ¹⁰³ Open AI, “Building an early warning system for LLM-aided biological threat creation,” January 31, 2024; Christopher A. Mouton, Caleb Lucas, Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks* (Washington, D.C.: RAND Corporation, 2024).

- ¹⁰⁴ Gerrit De Vynck, “AI Leaders Warn Congress that AI Could Be Used to Create Bioweapons,” *The Washington Post*, July 25, 2023; Jonas Sandbrink, “ChatGPT could make bioterrorism horrifyingly easy,” *Vox*, August 7, 2023; *Anthropic, Frontier Threats Red Teaming for AI Safety*, July 26, 2023; Emily H. Soice et al, “Can large language models democratize access to dual-use biotechnology?,” preprint, *arXiv*, June 6, 2023.
- ¹⁰⁵ By playing the role of the adversary (the red team), government officials and military experts can think creatively about ways to defeat the defenses of the blue team and devise solutions to mitigate the identified vulnerabilities.
- ¹⁰⁶ Open AI, “Preparedness Framework,” December 18, 2023.
- ¹⁰⁷ Center for AI Safety, “Statement on AI Risk,” accessed November 2, 2024.
- ¹⁰⁸ Jaime Sevilla et al, “Compute Trends Across Three Areas of Machine Learning,” preprint, *arXiv*, March 9, 2022.
- ¹⁰⁹ “AI firms will soon exhaust most of the internet’s data,” *The Economist*, June 23, 2024; Jared Kaplan et al, “Scaling Laws for Neural Language Models,” preprint, *arXiv*, January 23, 2020.
- ¹¹⁰ Epoch AI, “Machine Learning Trends,” accessed October 4, 2024.
- ¹¹¹ See, for example, “The Atlantic announces product and content partnership with OpenAI,” *The Atlantic*, May 29, 2024.
- ¹¹² Noor Al-Sibai, “AI Companies Running Out of Training Data After Burning Through Entire Internet,” *Futurism*, April 1, 2024.
- ¹¹³ Cade Metz et al. “How Tech Giants Cut Corners to Harvest Data for A.I.,” *The New York Times*, April 6, 2024.
- ¹¹⁴ Will Lockett, “AI Is Hitting A Hard Ceiling It Can’t Pass,” *Medium*, April 25, 2024.
- ¹¹⁵ Ilia Shumailov et al., “AI models collapse when trained on recursively generated data,” *Nature*, Vol. 631 (July 2024), pp. 755–759.
- ¹¹⁶ Noor Al-Sibai, “What is Synthetic Data? Why AI Trained on AI is the Next Big Thing (and Problem),” *Futurism*, April 8, 2024.
- ¹¹⁷ Devin Coldewey, “Model collapse’: Scientists warn against letting AI eat its own tail,” *TechCrunch*, July 24, 2024.
- ¹¹⁸ Goldman Sachs, “Gen AI: Too Much Spend, Too Little Benefit,” *Global Macro Research*, No. 129, June 25, 2024.
- ¹¹⁹ Yawen Chen, “Data centre boom reveals AI hype’s physical limits,” *Reuters*, July 4, 2024.
- ¹²⁰ Brian Calvert, “AI already uses as much energy as a small country. It’s only the beginning,” *Vox*, March 28, 2024.
- ¹²¹ Evan Halper, “Amid explosive demand, America is running out of power,” *The Washington Post*, March 7, 2024.

- ¹²² Will Lockett, “AI Is Hitting A Hard Ceiling It Can’t Pass,” *Medium*, April 25, 2024.
- ¹²³ Goldman Sachs, “Gen AI: Too Much Spend, Too Little Benefit.”
- ¹²⁴ *Ibid.*
- ¹²⁵ Alex Kantrowitz, “The End of Investors’ Generative AI Honeymoon,” *CMSWire*, July 26, 2024.
- ¹²⁶ Michael Chui et al., *The Economic Potential of Generative AI: The Next Productivity Frontier* (New York: McKinsey & Company, June 14, 2023); Daniel Verten, “It’s time to focus on the ROI of GenAI. Here’s how,” *World Economic Forum*, May 28, 2024.
- ¹²⁷ Grant Gross, “The complex patchwork of US AI regulation has already arrived,” *CIO*, April 5, 2024.
- ¹²⁸ The White House, *Blueprint for an AI Bill of Rights*, 2023.
- ¹²⁹ U.S. Congress, *National Artificial Intelligence Initiative Act of 2020*, H.R.6216, 116th Congress, March 12, 2020.
- ¹³⁰ National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*.
- ¹³¹ The White House, “Ensuring Safe, Secure, and Trustworthy AI,” July 21, 2023.
- ¹³² The White House, “FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence,” October 30, 2023.
- ¹³³ The White House, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Executive Order 14110, October 30, 2023.
- ¹³⁴ Artificial Intelligence Safety Institute Consortium (AISIC), accessed October 18, 2024.
- ¹³⁵ Kyle Wiggers, “NIST releases a tool for testing AI model risk,” *TechCrunch*, July 27, 2024; National Institute for Standards and Technology (NIST), “What is Dioptra?,” accessed October 18, 2024.
- ¹³⁶ The White House, *Memorandum on Advancing the United States’ Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence*, October 24, 2024, accessed November 11, 2024.
- ¹³⁷ Micah Zenko, *Red Team: How to Succeed by Thinking Like the Enemy* (New York: Basic Books, 2015).
- ¹³⁸ Open AI, *GPT-4 System Card*, March 23, 2023; see also, Daniel Fabian, “Google’s AI Red Team: the ethical hackers making AI safer,” *Google Blog*, July 19, 2023.
- ¹³⁹ Natasha Bajema, “Why Are Large AI Models Being Red Teamed?,” *IEEE Spectrum*, March 15, 2024.
- ¹⁴⁰ European Union, “AI Act,” factsheet on EU Regulation 2024/1689, accessed October 18, 2024.
- ¹⁴¹ International Institute for Sustainable Development, “UN Secretary-General Launches Advisory Board to Support AI Governance,” accessed October 18, 2024.

¹⁴² United Nations AI Advisory Body, *Governing AI for Humanity* (New York: United Nations, September 2024).

¹⁴³ Government of the Netherlands, "About REAIM 2023," accessed October 18, 2024.

¹⁴⁴ U.S. Department of State, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," *Bureau of Arms Control, Deterrence, and Stability*, accessed October 18, 2024.

¹⁴⁵ Maximilian Vogel, "The Perfect Prompt: A Prompt Engineering Cheat Sheet," *Medium*, April 8, 2024.



nonproliferation.org



Middlebury Institute of
International Studies at Monterey
James Martin Center for Nonproliferation Studies