November 2023

State of California

Benefits and Risks of Generative Artificial Intelligence Report

CalGovOps

CALIFORNIA GOVERNMENT OPERATIONS AGENCY

Table of Contents

I. Introduction
- a. What makes GenAI Different?
- b. Economic Backdrop of GenAI

II. Beneficial Use Cases for GenAI in State Government
- a. Use Case Analysis for GenAI in California State Government
- b. The Unique Benefits and Applications of GenAI

III. GenAI Risk Analysis
- a. Unique and Shared Risks of GenAI
- b. Identifying GenAI High-Risk Use Cases

IV. Ongoing Engagement

V. Conclusion

VI. Appendix
- a. Policy Landscape Assessment

I. Introduction

The diversity of the nearly 40 million people who call California home – and the strength of its multifaceted economy – have made California a global leader in technology and innovation. With the proper guardrails in place, the revolutionary technology of Generative Artificial Intelligence (GenAI) can be responsibly used to spur innovation, support the State workforce, and improve Californians' lives.

This report on the use of GenAI in State government is the first major product of Governor Newsom's Executive Order N-12-23 on Generative Artificial Intelligence (Executive Order), and it is the first step in an ongoing process of engagement with stakeholders and across State agencies. The report presents an initial analysis of the potential benefits to individuals, communities, government and State government workers, with a focus on where GenAI may be used to improve access to essential goods and services. Additionally, the report assesses the risks of GenAI, including but not limited to risks stemming from bad actors, insufficiently guarded governmental systems, unintended or emergent effects, and potential risks toward democratic and legal processes, public health and safety, and the economy.

When used ethically and transparently, GenAI has the potential to dramatically improve service delivery outcomes and increase access to and utilization of government programs. This report offers an analysis for State government leaders to explore the potential benefits and risks of GenAI thoughtfully, including how it can be used to empower California's workers. An examination of the research and feedback from academia, industry, local, state and federal government, and community organizations found the following common themes:

1. *GenAI is unique from conventional forms of AI*, and it necessitates a different state approach to implementing and evaluating this technology.
2. *GenAI enables significant, beneficial use cases* for state government through its unique capabilities.
3. *GenAI raises novel risks compared to conventional AI* across critical areas such as democratic and legal processes, biases and equity, public

health and safety, and the economy, and requires measures to address insufficiently guarded governmental systems and unintended or emergent harmful effects from this technology.

Additionally, as humans have explicit and implicit biases built into our society, GenAI has the capacity to amplify these biases as it learns from input data. As such, it's imperative to consider the implications on Californians of different regions, income, races, ethnicities, gender, ages, religions, abilities, sexual orientation and more for all GenAI inputs, outputs, and products–for both prioritizing implementations that may promote equity and guarding against bias and other negative impacts.

Acknowledging the unprecedented nature of GenAI requires a collaborative effort between states, the federal government, and international partners, this analysis relies on learnings from the National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF) and international policies and governance frameworks. The federal NIST AI RMF was developed to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

The State's commitment to transparency is a foundation for ongoing GenAI work and collaboration. This report is only the first step in a multi-year and iterative process as part of the Governor's Executive Order, which also:

- Directs state agencies and departments to perform a joint risk-analysis of potential threats to and vulnerabilities of California's critical energy infrastructure using GenAI.
- Supports a safe, ethical, and responsible innovation ecosystem inside state government by requiring general guidelines for public sector procurement, uses, and required training for application of GenAI.
- Provides for guidelines to analyze the impact that adopting GenAI tools may have on vulnerable communities.
- Prepares California's state government workforce through training for workers to use state-approved GenAI.
- Requires evaluation for potential impact of GenAI on regulatory issues.

California government will continually engage academic leaders and researchers, labor organizations, community organizations, and industry experts as the State pilots GenAI use cases and creates guardrails to protect Californians and their data.

What makes GenAI different?

GenAI builds on advances in conventional AI and uses very large quantities of data to output unique written, audio, and/or visual content in response to free-form text requests from its users and programmers. GenAI tools have the capacity to produce entirely new content instead of simply regurgitating inputted data. Unlike conventional AI systems designed for specific tasks, GenAI models are designed to be flexible and multifunctional.

GenAI products are already available as standalone applications such as ChatGPT, Dall-E, and Bard, and are being integrated into many other consumer-facing technology products, such as chatbots on websites.

Conventional AI models, on the other hand, are usually designed for just a few specific tasks and are often limited by the scope of the inputted training data as well as the technical expertise of the programmer. Model training is the process by which AI models ingest input datasets to learn the underlying patterns within the data and produce predictions for the context that the model was trained on.

Conventional AI is already widely used in products across government and society. Some examples of conventional AI include robotic process automation, fraud detection tools, image classification systems, recommendation engines, and interactive voice assistants.

Table 1: Comparison Between Conventional AI and GenAI Technology

Criteria	Conventional AI	Generative AI
What is the intended purpose?	Solve specific problems or accomplish predefined tasks using a predefined dataset.	Generate new content (text, images, music, etc.) and produce novel outputs not seen within input datasets.
How is the AI model trained?	Learns patterns from large amounts of structured data for training and uses them to make predictions or perform tasks.	Learns patterns using unstructured data sets. Ongoing training can be performed for fine-tuning of model for specific business uses.
What kind of algorithm does the AI model use to learn from its input data?	Typically runs on rule-based systems, decision trees, and similar models. Can learn underlying patterns in the data but requires more pre-processing for the algorithm to perform well.	Uses flexible neural network algorithms that can process different inputs and learn the underlying relationships and patterns within the data.

Criteria	Conventional AI	Generative AI
How is the AI model typically used?	Image recognition, recommender systems, anomaly detection, text classification, and risk prediction systems.	Creative tasks like art, music, storytelling, content generation, image synthesis, text generation, video creation, style transfer, and logical reasoning.
How is the AI model evaluated?	Typically, task-specific performance measures that assess accuracy, precision, recall metrics.	GenAI outputs can be more subjective and dependent on human judgment. Quality assurance of output is important.

GenAI technology function using foundation models, which are large-scale machine learning models with general purpose capabilities. These models are trained on datasets that can span the entirety of the internet, and they can become the foundation for applications that can help address specific business, policy, or social needs. As they are built and grow, foundation models require larger quantities of computing power and human capital resources than conventional AI development.

GenAI models use human-generated content as part of their underlying data, and they can respond to free-text human queries with human-sounding output. However, despite the capacity of GenAI to produce coherent, intelligent-sounding output, there is no guarantee that the output is accurate. In fact, many of the most widely available GenAI models were designed as a demonstration of what is possible, rather than to solve a specific use case or business purpose. As a result, free consumer models can produce outputs that are inaccurate, fabricated, potentially inappropriate, and/or biased.

These products demonstrate the unprecedented power of GenAI, and enterprise models continue to improve in approximating how humans write, draw, and speak. Simultaneously, the rapid development and availability of GenAI has accelerated policy, business, and social risks that are more urgent than previous AI technologies.

Economic Backdrop of GenAI

California stands at the forefront of the burgeoning AI economy. Home to 35 of the world's top 50 AI companies, California leads the world in GenAI innovation and research.

Our higher education institutions – including UC Berkeley's College of Computing, Data Science, and Society, and Stanford University's Institute for Human-Centered Artificial Intelligence – are among the most advanced AI research institutions in the world. Coupled with the State's unparalleled access to venture capital, our culture of innovation, and history of new, world-changing technologies, California sits at the epicenter of an industry that is experiencing exponential growth and development.

Although GDP growth and productivity gains are predicted, Goldman Sachs has also warned that 300 million jobs worldwide could be affected by GenAI. As such, the State must lead in training and supporting workers, allowing them to participate in the AI economy and creating the demand for businesses to locate and hire here in California. Starting with our world-class higher education institutions and vocational schools, California is well positioned to provide workers with relevant skills and businesses with the talent needed to drive job growth in the GenAI economy.

The global GenAI market is significant. According to Pitchbook, it is expected to reach $42.6 billion in 2023. Like all new technologies, particularly of this scale, GenAI offers immense economic opportunities, as well as new risks. As the industries of GenAI are developed, California, the U.S., and other nations must develop coordinated and thoughtful public policies to mitigate risks and maintain public trust through ethical use guidelines, accountability, and transparency, while still realizing the potential economic benefits of GenAI.

II. Beneficial Use Cases for GenAI in State Government

Use Case Analysis for GenAI in California State Government

Government leaders should prioritize GenAI proposals that offer the highest potential benefits, along with the appropriate risk mitigations, over those where benefits are not significant compared to existing work processes. This technology offers possibilities to improve the lives of Californians, such as by summarizing benefits enrollment policies in plain language, translating government communications into multiple languages, and providing interactive tax assistance.

Under the Governor's Executive Order, agencies are tasked with soliciting stakeholder input and crafting guidelines for state use of GenAI. That work has begun and will be completed in January 2024, but in the interim, basic principles that should apply:

- To protect the safety and privacy of Californians' data, and consistent with state policy–state employees should only use state-provided, enterprise GenAI tools on State-approved equipment for their work.
- Under no circumstances should state employees provide state or Californians' resident data to a free, publicly available GenAI solution like ChatGPT or Google Bard, or use these unapproved GenAI applications or services on a State computing device.
- It is important to provide a plain-language explanation of how GenAI systems factor into delivering a state service and disclose when content is generated by GenAI.
- State supervisors and employees should also review GenAI products for accuracy and make sure to paraphrase rather than use AI-generated text, audio, or images verbatim.

Through consultation with practitioners and researchers, California state government compiled an inventory of potential GenAI use cases that could improve state services and programs. High-level categories within the use cases were extracted and are enumerated in this section as potential areas of benefit from GenAI.

Looking ahead, with the appropriate pilot infrastructure and risk mitigations in place, California will evaluate potential use cases by prioritizing the following benefits:

1. ***Improve the performance, capacity, and efficiency of ongoing work, research, and analysis through summarization and classification.***

   By analyzing hundreds of millions of data points simultaneously, GenAI can create comprehensive summaries of any collection of artifacts, irrespective of whether the content is in a text, audio, or video format. As GenAI learns, it can also categorize and classify information by topic, format, tone, or theme.

   Example Use Cases include:

   - Conduct sentiment analysis of public feedback on state policies, using GenAI to recommend opportunities for process and service delivery improvement. This can help government understand public experience and improve policies and communication to better serve constituents.
   - Summarize meetings, work, and public outreach documentation, leveraging GenAI to find insights in the analyzed data. GenAI can find the key topics, conclusions, action items, and insights without needing to read everything word for word.

2. ***Personalize and customize work products to California's diversity of people with the potential to improve access to services and outcomes for all.***

   GenAI's capacity to learn makes it easier for the State to design services and products to be responsive to Californians' diverse needs, across geography and demography. GenAI solutions can recommend ways to display complex information in a way that resonates best with various audiences or highlight information from multiple sources that is relevant to an individual person. These functions can further California's goals as they allow for optimized government experiences allowing Californians greater access to state information and services, and by advancing equity, inclusion, and accessibility in outcomes.

   Example Use Cases include:

   - Apply GenAI on government service data to identify specific groups or subsets of participants that may benefit from additional outreach, support services, and resources based on their circumstances and needs (for example, local job training for people claiming EITC).

- GenAI can identify groups that, for language or other reasons, are disproportionately not accessing services by analyzing feedback surveys or comments for language that indicate accessibility difficulties. This can help determine opportunities to improve access.

3. *Improve language and communications access in multiple languages and formats.*

GenAI can create unique content in a variety of formats. Based on a single prompt, a GenAI solution can easily construct a video or image that a user can refine. These products can be in multiple languages, allowing the State to make its videos, recordings, and other documents more accessible to and inclusive of all Californians. These translated outputs can be refined through a quality control process to ensure accuracy and inclusivity before reaching Californians.

Accessible communications are a critical part of ensuring that government services can meet Californians where they are. The ability to meet the varying communication needs of persons with disabilities and reach Californians in their primary languages is a priority for improving government service delivery.

Example Use Cases include:

- Using GenAI to help experts convert educational materials into formats like audio books, large print text, or braille documents. Can also generate captions for video materials, and make information more accessible for those with visual, hearing, or learning disabilities.
- Leveraging GenAI to help experts translate government websites, public documents, policies, forms, and other materials into the various languages spoken in the State. This expands access to important information and services to non-native English speakers.

4. *Optimize software coding and explain and categorize unfamiliar code.*

Summarization, classification, and translation features make GenAI a powerful tool for state coders and the developer community at large. GenAI can generate code in multiple computing languages and translate code from one language to another. This can improve state operations if a state system is using code that is written in an obsolete language. Moreover, GenAI has the potential to explain and categorize unfamiliar or uncertain code so that the State can better understand the exact technical architecture of agency applications.

Example Use Cases include:

- Powerful code conversion tools based on foundation models can accurately translate legacy codebases (e.g., COBOL mainframe apps) into modern programming languages. This automates time-consuming and error-prone manual conversions.
- Powerful GenAI development tools auto-generate quality code, spin up test environments, and generate synthetic datasets to train machine learning models. This can slash timelines, reduce bugs, and democratize development. Low-code solutions also enable non-programmers to build applications.

5. *Find insights and predict key outcomes in complex datasets to empower and support decision-makers.*

Without specific training or pre-set rules, GenAI models can analyze multiple datasets to find meaningful insights for users. The conversational aspects of GenAI solutions can empower workers with a range of technical expertise to ask questions in plain language to get at findings that may be relevant to their work. Significantly, Californians could also use a GenAI solution to ask data-driven questions that are important to them.

Example Use Cases include:

- Cyber protection systems powered by foundation models can rapidly analyze network activity logs, identify anomalies and threats, generate explanations of the attacks, and propose remediation actions. This can enable security teams to detect and respond to sophisticated cyberattacks in real-time before major damage occurs.
- GenAI analyzes data streams from drones, satellites, and sensors monitoring public infrastructure. It generates detailed damage and deterioration assessments via techniques like visual inspection, anomaly detection, etc. This enables improved forecasting of maintenance needs.

6. *Optimize workloads for environmental sustainability.*

Incorporating GenAI in government can drive environmental sustainability by optimizing resource allocation, maximizing energy efficiency and demand flexibility, and promoting eco-friendly policies. For instance, this technology can enhance operational efficiency, decrease paper usage and waste, and support environmentally conscious governance. Notably, stakeholders also highlighted the need for reducing environmental impacts of GenAI use and ensuring environmental costs are equitably distributed.

Example Use Cases include:

- GenAI could analyze traffic patterns, ride requests, and vehicle telemetry data to optimize routing and scheduling for state-managed transportation fleets like buses, waste collection trucks, or maintenance vehicles. By minimizing mileage and unnecessary trips, GenAI could reduce associated fuel use, emissions, and costs.
- GenAI simulation tools could model the carbon footprint, water usage, and other environmental impacts of major infrastructure projects. By running millions of scenarios, GenAI can identify potentially the most sustainable options for planning agencies and permit reviewers.

Across all use case opportunities, potential use cases will need to be customized to the case-by-case needs of state departments and evaluated through a coordinated, standardized benefits and risks assessment process through pilot programs. Through pilot testing and experimentation in GenAI sandbox environments, the State will document learnings to refine and scale its GenAI community of practice.

The Unique Benefits and Applications of GenAI

GenAI has the potential to improve the delivery of government services and operations. Feedback from academic, industry, and community stakeholders highlights the unique benefits and applications of this novel technology compared to conventional AI and manual workflows.

The following table lists high-level categories for the wide variety of functionality for GenAI with sampled public sector use cases. The example use cases are only intended to help illustrate the potential uses of state government adoption of GenAI tools.

Table 2: A Typology for GenAI Tasks

GenAI Task	Unique Benefits	Example of Public Sector Use Cases
Content generation (text, image, video)	Generates completely novel content, instead of remixing and modifying existing content. Few-shot learning allows high-quality output with minimal data.	• Generate public awareness campaign materials like fliers, website content, posters, and videos. • Generate visualizations of transportation data.

GenAI Task	Unique Benefits	Example of Public Sector Use Cases
Chatbots	Leverages conversational models trained on massive dialogue datasets. Can have coherent discussions and execute tasks via conversation naturally.	• Build virtual assistant for common constituent questions. • Create chatbot to guide users through services in their preferred language. • Increase first-call resolution for state service centers. • Reduce call wait and handle time at state customer service centers. • Create greater language access equity for program beneficiaries.
Data analysis	Finds insights and relationships in data through learned knowledge about the world, without hand-coded rules or labeled training data.	• Analyze healthcare claims or tax filing data to detect fraud. • Analyze network activity logs, identify cybersecurity anomalies and threats, and propose remediation actions.
Explanations and Tutoring	Generates natural language explanations and tutoring through dialogue without hand-authored content.	• Explain program eligibility to potential enrollees. • Provide interactive tax assistance.
Personalized Content	Leverages user models to adaptively generate personalized content without explicit rules or large amounts of user data. User models learned via few-shot interaction.	• Auto-populate tax information and filing instructions based on a person's needs. • Help auto-populate public program applications based on a person's situation and household composition.
Search and Recommendation	Understands meaning and context to improve search relevance and provide useful recommendations.	• Searching or matching state code regulations concerning specific topics. • Recommend government services based on eligibility.
Software code generation	Generates code by learning underlying structure and patterns of code, without need for human written examples. Can expand short descriptions into full programs.	• Translate policy specifications such as Web Content Accessibility Guidelines (WCAG) and Americans with Disability Act (ADA) requirements, into software code. • Generate data transformation scripts from instructions.

GenAI Task	Unique Benefits	Example of Public Sector Use Cases
		• Accelerate adoption of human-centered design in state web-based forms and pages. • Reduce administrative cost and burden to developing and maintaining best-in-class state government websites.
Summarization	Does not require human-written summaries as training data. Can learn underlying patterns of language to generate summaries.	• Summarize public comments to identify key themes. • Summarize public research to inform policymakers. • Summarize statutory or administrative codes.
Synthetic data generation	Allows generation of new diverse, anonymized data from existing datasets for analysis and experimentation.	• Generate synthetic patient data for training healthcare AI. • Generate simulated tax records for training tax auditing AI.

GenAI offers a wide variety of potential applications, with varying impacts. Any application of GenAI tools within California state government will follow the appropriate protocols and testing procedures, as well as incorporating feedback from impacted stakeholders as guidance on the use of this technology. Looking ahead, California state government will evaluate potential use cases that will provide maximum benefit to Californians, and in line with updated guidelines and criteria as directed by the Executive Order.

III. GenAI Risk Analysis

Research conducted within state government, informed by feedback from subject matter experts and community groups, has developed an emerging picture of the specific risk factors of GenAI compared to those posed by conventional AI. As with conventional AI, GenAI poses risks both from bad actors using the technology to cause harm as well as from unintended, emergent capabilities of GenAI that can be misused.

The NIST AI RMF divides risks into seven categories: Validity & Reliability, Safety, Accountability & Transparency, Security & Resiliency, Explainability & Interpretability, Privacy, and Fairness. In no particular order or weight, these seven NIST AI RMF categories have been analyzed as they apply to GenAI adoption in California. Although the NIST AI RMF provides a helpful framework to illustrate key risk areas, it does not specifically address GenAI, and it is not specific to California's values or use case context. To bridge this gap, and as identified through research and stakeholder engagement, the additional category of Workforce & Labor Impacts is included below.

Given the rapidly evolving capabilities, integrations, and standards of GenAI products, the following analysis represents an initial evaluation of GenAI risks, which delineates risks based on being a shared risk of conventional AI, an amplified risk, or a new risk associated with GenAI.

- **Shared risks**: Known risks of GenAI shared by earlier types of AI models without significant differences in severity or scale.
- **Amplified risks**: Risks of GenAI tools shared by earlier types of AI models that are enhanced due to any of the following factors:
    - Reduced technical or cost barriers to using GenAI.
    - Increased speed or scale of impact by GenAI tools.
    - Increased scope of systems or processes impacted by GenAI.
    - Increased exposure to bad actors via larger, more diverse training datasets.
    - Higher complexity of GenAI technology architectures with multiple producers and consumers.
- **New risks**: Novel risks surfaced by GenAI's unique capabilities to generate high-quality outputs across a diversity of modalities such as text, images, audio, and video.

Unique and Shared Risks of GenAI

1. Validity & Reliability

AI systems that are inaccurate or unreliable increase risks and reduce trustworthiness.

- *Validation* is the "confirmation through evidence that the requirements for a specific intended use or application have been fulfilled."
- *Reliability* is the "ability of an item to perform as required, without failure, for a given time interval, under given conditions."

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Amplified risks	AI models that rely on static datasets can become outdated. This can lead to less relevant outputs and model degradation over time.
	Third-party providers of conventional AI models commonly release minor software updates without notice, which in turn can impact performance.
	Automated "testing" of Large Language Model (LLM) outputs; unlike in traditional software testing, the output of AI models can differ, even with the same prompt or input.
	GenAI models are normally pre-trained using a vast amount of unbalanced, incomplete, and potentially harmful content, which may not be directly relevant to the target application.
New risks	"Hallucinating," or creating misleading, false, or fabricated information and presenting it as if it were true.
	Worsening model performance through training feedback loops, when new GenAI models are trained on self-generated, synthetic data.
	Appearance of causal reasoning under standard tests and benchmarks for AI models.
	The qualitative elements of many GenAI evaluation processes such as coherence, fluency, and creativity can make it challenging to evaluate GenAI outputs in a standardized way.

GenAI models are more complex than conventional AI models, and as a result, they are more susceptible to model degradation and collapse, where the AI model's performance will worsen over time as the data used to teach it becomes more outdated. This is because GenAI models are trained on a large body of data and can produce their own synthetic data. This means that they can become biased towards their own synthetic data and become less accurate over time (a process known as "model collapse"). GenAI outputs can also be non-deterministic and inconsistent, making it difficult to embed into critical systems where performance stability is a key requirement.

The risk of over-reliance on automated GenAI recommendations to make decisions (automation bias), related to validity concerns on hallucinations, poses concerns around GenAI outputs given the ability to generate answers that "sound right" without having factual accuracy. Without proper safeguards, Californians may believe hallucinations inadvertently created by government GenAI tools, which could lead to additional downstream misinformation. This could reasonably erode Californians' trust in their government and its services.

2. Safety

AI systems "should not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered."

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Amplified risks	Misuse in critical applications such as systems affecting housing or accommodations, education, employment, financial credit, health care, or criminal justice for example.
	Using GenAI in tasks where precision and accuracy are paramount.
	GenAI tools can lower technical barriers for influential accounts to personalize content on platforms like social media, potentially amplifying the risk of mental health impacts or political polarization.
New risks	Input prompts crafted to push the GenAI model to make or recommend hazardous decisions.
	Creating harmful or inappropriate misinformation or disinformation material (e.g., cybersecurity, warfare, promoting violence, and harassment).
	GenAI tools may enable bad actors to design, synthesize, or acquire dangerous chemical, biological, radiological, or nuclear (CBRN) weapons.
	The output of GenAI systems may unintentionally contain inappropriate or harmful content such as violence, profanity, racism, or sexism.
	As models are increasingly able to learn and apply human psychology, models could be used to create outputs to influence human beliefs, addict people to specific platforms, or manipulate people to spread disinformation.

GenAI tools can pose significant risks to public health and safety–whether employed by people with malicious intent, or simply because of a lack of quality controls. For example, bad actors can leverage AI to engineer dangerous biological materials, AI chatbots could give consumers incorrect or dangerous medical advice, or GenAI systems used for drug discovery could create harmful substances. In sensitive domains like healthcare and public safety, GenAI requires careful governance to mitigate the risk of harm.

Additionally, GenAI can utilize better and more realistic text generation capabilities to simulate human text and opinions, leading to novel scaling capabilities for spreading misinformation or disinformation on public forums. Bad actors could weaponize misinformation and disinformation, amplifying it through GenAI to interfere in democratic processes. This includes the generation of disinformation campaign material to disseminate on social media, generating deepfakes of political representatives or candidates, or submitting large volumes of fake public comments for proposed rules.

Given these risks, the use of GenAI technology should always be evaluated to determine if this tool is necessary and beneficial to solve a problem compared to the status quo. GenAI should center on the needs of the human workforce, support the carrying out of responsibilities to Californians, and avoid contributing to additional bureaucracy, process, or safety risks.

3. Accountability & Transparency

Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with the system. Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle.

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Shared risks	Lack of standardized audit trail documentation when tracing the provenance of predictions from an AI system.
	Reproducibility concerns when auditing poorly documented AI models.
	Governance concerns with open-source AI models; third-parties able to host models without transparent safety guardrails.
Amplified risks	Lack of disclosure around the usage of AI models within a system or when embedded in a third-party vendor.
	Difficulty in receiving model decision explanations from third-party hosted model providers.
	Difficulty in auditing large volumes of training data for GenAI models.
	Gen AI systems are typically pre-trained and provide limited explainability or control to the end-users.
New risks	Difficulty in tracing the original citation sources for references within the generated content.
	Uncertainty over liability for harmful or misleading content generated by the AI.

The GenAI model lifecycle is typically more complex than that of conventional AI and raises novel challenges in ensuring transparency and accountability along the AI value chain. Building a GenAI model may involve multiple organizations that all may contribute data to the base foundation model or within the fine-tuning process.

California state government must be cautious about over-automating decisions or removing human oversight entirely with GenAI chatbots and text generators. There are risks in over-trusting these and other tools that rely on GenAI without proper review and evaluation of GenAI outputs, such as inaccurate information being provided to constituents or inaccurate public program determinations. Such inaccurate determinations, especially if made repeatedly, could pose particular risks severely undermine California's progress in creating a California for All by emphasizing to diversity, equity, inclusion, and accessibility. It will be critical to have a human reviewer of any GenAI-supported workflow or output that results in a decision about program eligibility or social safety net benefits.

4. Security & Resiliency

Security and resiliency are defined in the following ways:

- *Secure* AI systems can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use.
- *Resilient* AI systems can withstand unexpected adverse events or unexpected changes in their environment or use.

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Shared risks	Unauthorized user access of AI models.
	Data breaches or leaks tied to the AI model.
	Theft of AI models leading to misuse or malicious content generation.
Amplified risks	Data poisoning, when low quality or biased data is intentionally or unintentionally leaked into a training dataset for an AI model.
	Model inversion, when malicious actors can steal sensitive personal data through the AI model's outputs.
	Model skewing, when malicious actors intentionally amplify biased training data to skew model decisions.
	Adversarial attacks, when malicious actors can supply inputs to the AI model designed to break the system.
	Supply chain vulnerabilities through third-party services, plug-ins, and libraries.

Type of Risk	Description of GenAI Risks
New risks	Adversarial prompt attacks that can cause the GenAI model to produce unwanted content.
	Remote execution of harmful code through the GenAI model to modify access permissions, delete, or steal data.
	Prompt injection attacks, which can manipulate the model into taking undesirable actions.
	Generated content may be indistinguishable from content created by a human, which could enable the scope of harm caused by bad actors across sectors.

There are some shared data security risks across conventional AI and GenAI models. Data can be vulnerable to unauthorized access, low-quality data can be injected into training datasets to impact overall model performance, and crafted inputs can cause AI and GenAI models to exhibit inconsistent performance.

As members of Cal OES's Cybersecurity Integration Center (Cal-CSIC), CDT's Office of Information Security works collaboratively with the California Highway Patrol (CHP), California Military Department (CMD), Office of Health Information Integrity, and other essential agencies on mitigating, identifying, responding to, and reporting security incidents.

GenAI systems can be susceptible to unique attacks and manipulations, such as poisoning of AI training datasets, evasion attacks, and interference attacks. As with any other technology-driven threat to state security, when a state employee suspects one of these GenAI related incidents such as a GenAI-generated or -impacted incident has occurred, to the degree they're known, the employee should report it immediately for central tracking and coordination. Consistent with State Information Management Manual (SIMM) section and current practice for other technology-driven threats, it is the responsibility of the state entity Information Security Officer (ISO) or authorized user to immediately report the incident through the California Compliance and Security Incident Reporting System (Cal-CSIRS) so that further pattern analysis can be conducted for correction and safeguarding.

The capabilities of GenAI generally raise concerns about enabling bad actors and undermining government security if not properly governed.

A few general examples include:

- Augmenting criminal activities by generating more convincing scams, malicious code, and deception.
- Tricking consumers into sharing personal data for advertising or manipulation through enhanced phishing capabilities with voice, image, and video deepfakes.
- Enabling scammers to efficiently produce high volumes of convincing text.

New capabilities created by GenAI will pose new security risks, threatening existing systems around both physical and digital infrastructure. Robust, new security controls, monitoring, and validation techniques will be needed to guard against potential attacks. GenAI has a wider security risk surface exposed via their natural language interfaces. It is easier for adversarial attacks to occur and less intuitive to place security controls on the model weights that produce recommendations and decisions by the GenAI model.

To that end, the Governor's Executive Order requires a classified joint risk analysis of potential threats to and vulnerabilities of California's energy infrastructure and directs development of a strategy to assess threats to other critical infrastructure by the use of GenAI.

5. Explainability & Interpretability

Explainability and interpretability are defined in the following ways:

- *Explainability* refers to a representation of the mechanisms underlying AI systems' operation.
- *Interpretability* refers to the meaning of AI systems' output in the context of their designed functional purposes.

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Shared risks	Black-box decision-making that makes AI model recommendations unexplainable.
Amplified risks	Complexity and opaqueness of AI model architectures.
New risks	Users or stakeholders misunderstanding or misinterpreting generated content.
	Users attributing logical thinking to GenAI models when models are asked to give explanations for how the output was generated.

GenAI models are similar to certain types of conventional AI models like neural networks, which are black-box algorithms that cannot provide direct explanations for their predictions. Without the ability to explain model predictions and outputs, it becomes more difficult to address cases where this technology produces an unexpected result that impacts the validity and consistency of the answers. There is ongoing research to gain better explainability capabilities for these types of algorithms. However, GenAI models amplify these concerns because they are built from much larger and more complex neural networks than conventional AI models.

The difficulty in extracting human-interpretable explanations from GenAI technology is an important factor to consider for government to provide sufficient information about decisions that concern constituents.

Additionally, GenAI models can be prompted to "explain their reasoning" through prompting techniques. However, these techniques can be inconsistent because GenAI models have been shown to misrepresent their stated reasoning. These techniques can be unreliable in extracting a GenAI model's true logical reasoning for an output, compared to the model's stated reasoning.

6. Privacy

Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities.

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Shared risks	Unauthorized data access or usage by users.
	Insufficient data anonymization on training data leading to the leakage of sensitive information.
	Collection of more data than necessary to train the AI model.
	Proprietary data may be used in training third-party AI models.
Amplified risks	Over-reliance on vast amounts of data for generation, risking privacy.
	Difficulty in erasing personal information embedded within the model features (known as algorithmic disgorgement).
New risks	GenAI unintentionally recreating, inferring, or falsifying private or sensitive details.
	AI-generated content potentially revealing or alluding to training data details.
	Nonconsensual use of people's likeness (e.g., deepfakes, voice impersonation, biometric data like gaze direction, gait analysis, and hand motions).

GenAI models can leak personal data if they are not properly anonymized or if their training data is not properly secured. For example, if a GenAI model is trained on a dataset of medical records, it could potentially generate text that includes personal information about patients, such as their names, medical conditions, or medications. This information could be used to identify individuals, even if the model was trained on an anonymized dataset.

GenAI also raises novel privacy issues such as:

- *Re-identification risk:* GenAI models can also be used to synthesize new datasets from previously unintegrated data sources that can be used to re-identify individuals. For example, if a GenAI model is trained on a dataset of images of people, it could potentially generate new images that are similar to the images of real people in the training dataset. These new GenAI images could then be used to identify real individuals in the training dataset, even if the original images were anonymized. This re-identification risk is particularly critical in regard to sensitive personal data, where individuals could be exposed to unsafe conditions if unintentionally disclosed.
- *Third-party plug-ins and browser extensions:* Third-party plug-ins and browser extensions that interact with GenAI models can also pose privacy risks. For example, a plug-in could collect data about the user's interactions with a GenAI model, such as the text that they generate or the images that they create. This data could then be shared with the plug-in's developer or with third-party companies without the user's knowledge or consent.
- *Government's ability to respond to consumer privacy requests:* As Californians' right to remove their personal data online becomes more widely practiced, extracting and destroying their information embedded within GenAI models may become difficult or administratively unsustainable.
- *Bad actors accessing and sharing government database content:* The state of California maintains secure databases with records of individuals' data, such as census data and the program-specific data minimally necessary to make eligibility determinations. If a bad actor were able to gain illegal access to a state database, GenAI could power the rapid capture and leak of Californians' private data. Data leaks and data loss from data centers also pose an ongoing risk, which will need to be addressed through improved controls.

7. Fairness

Fairness in AI includes concerns for equality and equity by addressing issues such as bias and discrimination.

When applied to GenAI, California identified the following risks:

Type of Risk	Description of GenAI Risks
Shared risks	Services using AI systems may not be accessible across different parts of the digital divide, where some communities may not have equitable access to digital technology platforms.
	AI systems may not work the same way or with similar level of accuracy for all subsets of the population, in particular for under-represented, vulnerable or protected groups. Such algorithmic discrimination would exacerbate existing social inequities.
	On-demand pricing for GenAI tools can result in large costs to institutions that under-resourced communities may not be able to pay. This may limit or block constituents from using GenAI tools, further exacerbating inequities.
Amplified risks	Discriminatory or biased outcomes caused by model recommendations or predictions.
New risks	Generating content that reflects or amplifies racial, gender identity, sexual orientation, or other biases or stereotypes.
	Model performance results in disparities across languages, biasing towards English and high-resource languages.
	The output of GenAI does not reflect social or cultural nuances of sub-sets of the population.

GenAI models can perpetuate societal biases if the training data is imbalanced. For example, large language models often perform poorly for non-native English speakers. This could create inequity in access to certain government services. Government must also proactively assess for algorithmic discrimination, such as gender, racial, or other biases, particularly in high-impact areas like criminal justice, healthcare, mental health, social services, and employment decisions. Algorithmic bias in state systems can be especially harmful if the GenAI authorship of the content is not disclosed, leading human consumers to misattribute the biased or harmful content to the government.

In conventional AI models, bias can be mitigated by collecting and processing training data to correct for under-representation of historically marginalized groups. This is important because creating rules that intentionally bias model weights during model training could have legal implications. For example, if an AI model is used to make decisions about who gets a loan, and that model is biased against people of a certain race, then the company using that model could be sued for discrimination.

GenAI datasets however are much larger than conventional AI models, making it more difficult to resolve embedded bias. Common expressions of data bias in GenAI outputs can include gender and racial stereotypes. This is usually relevant when generating narrative examples, image generation, or creating synthetic data.

8. Workforce & Labor Impacts

The adoption of GenAI technologies into the economy and workforce will introduce many changes that will support workers in their daily responsibilities and tasks, but also will change or modify parts of their existing workflows.

California identified the following risks when analyzing the use of GenAI:

Type of Risk	Description of GenAI Risks
Shared risks	Ethical considerations for AI annotation work within the training process and ensuring safe, fair working conditions.
	Workers may require training programs to effectively use AI tools to facilitate their existing workloads.
Amplified risks	Certain industries may experience job displacement from AI, requiring proactive and comprehensive re-skilling programs to support workers through employment transitions.

Key areas for workforce impact considerations include:

- *Up-skilling, re-training, and job transition assistance:* With the integration of GenAI tools into the workplace, staff may require up-skilling programs to effectively use the technology in their daily responsibilities. For individuals that experience job displacement, private companies and public services need to prepare for proactive and thoughtful re-skilling and transition support services.
- *Labor exploitation:* GenAI could enable new forms of labor exploitation, such as in data labeling where contract workers in developing countries are employed to annotate datasets used for training AI models without labor rights guarantees.
  This can encourage unsafe working conditions, especially, for contract workers in sensitive fields like content moderation for graphic and inappropriate content.
- *Anticompetitive behavior:* Major firms could use GenAI to further concentrate power in anticompetitive ways, such as by replicating copyrighted data from artists or small businesses.

Identifying GenAI High-Risk Use Cases

Future state research and development of guidelines will continue to support agencies and departments in identifying the severity and scope of GenAI risks, so that state government can better align oversight to real-world impacts. The Governor's Executive Order instructs agencies to begin this work. But California's efforts in this regard will surely continue beyond the Executive Order's deliverables – including partnership between the Administration and the Legislature to identify risks and codify strategies to mitigate them.

A risk-based approach to AI aligns with global trends. Major governmental entities like the European Union and NIST already employ risk-based frameworks for AI evaluation and deployment. By identifying the level of risk associated with GenAI deployment, organizations can implement responsible GenAI systems consistent with international practices.

When a high-risk use case is identified and GenAI is being used, state entities will need to take additional precautions. Government Code § 11546.45.5[1] defines "high-risk automated decision systems" for state entities and serves as a basis to identify where these precautions should be defined.

The definition states:

> *"High-risk automated decision system" means an automated decision system that is used to assist or replace human discretionary decisions that have a legal or similarly significant effect, including decisions that materially impact access to, or approval for, housing or accommodations, education, employment, credit, health care, and criminal justice.*

Lower risk systems that fall outside of this high-risk classification may still benefit from risk mitigation and transparency measures. The following table displays initial considerations that may help determine actions needed to mitigate the risks presented by GenAI.

[1]/ Effective January 1, 2024.

Table 3: California's Oversight Considerations for GenAI Applications by Risk Level

Risk Level	Oversight Considerations
Low risk	For AI systems deemed low risk, standard monitoring and lightweight evaluations are sufficient. Monitoring efforts can track user uptake, feedback, time savings, and output quality. Voluntary adoption by staff also signals effectiveness.
Moderate risk	Moderate-risk systems warrant more involved oversight, such as systematic monitoring and rapid cycle evaluations. Monitoring remains important for moderate risk applications, tracking usage, feedback, efficiency gains, and outcome improvements. Short internal evaluations should compare processes and outputs with and without the AI tool.
High risk	High-risk systems require intensive evaluations, qualitative assessments, and risk mitigation measures. • Pre-deployment assessments and red-teaming of GenAI systems are crucial as guardrails to catch any issues with fairness, privacy, security, performance, and safety in the model beforehand. • Post-deployment monitoring is also a critical complementary piece in identifying security vulnerabilities, performance changes, and equity issues. • Community feedback collections should capture diverse perspectives and enable processes to appeal decisions, especially from historically marginalized groups. • High risk GenAI systems should undergo an evaluation of whether the tool is beneficial and necessary prior to release.

Use cases involving critical applications, tasks requiring empathy or compassion, contextual understanding, and tasks requiring extensive domain knowledge or experience, are likely inappropriate for GenAI algorithms without human operators and significant oversight.

IV. Ongoing Engagement

This report was not possible without extensive collaboration. The initial findings and recommendations of this report mark the beginning of a much broader and ongoing conversation about the benefits and risks of this potentially transformative technology.

The State of California will regularly assess and update the findings of this report with significant developments as appropriate. To do this, the State will continue strengthening collaborations with academia, other governmental entities, industry, policy experts, organizations representing employees, and community-based organizations.

GenAI has incredible potential, and it is the State's responsibility to create an opportunity where Californians can help to chart their own future with this new technology.

V. Conclusion

This report represents a preliminary analysis to a much bigger conversation around a technology that is emerging and still rapidly developing, from environmental considerations to training and education. Acknowledging GenAI presents new and unique risks that are still to be addressed by future updates, California has laid out initial, potential use cases and risks identified through literature research and feedback from stakeholders. The unique risks posed by novel GenAI capabilities require augmented governance for the added risks between conventional AI and GenAI.

Under the Governor's Executive Order, the State will undertake significant efforts to evaluate and update its current procurement methods, practices, and vendor terms and conditions to place protections to acquire GenAI tools safely. Moreover, through GenAI pilots and sandbox applications established in the Executive Order, the State will be able to continually adapt guidance in response to lessons learned. Carefully designed pilot cases will enable state leaders to assess the outcome, scale efforts that prove to be successful, and share learnings and best practices with public policy makers and stakeholders in other state governments, the federal government, and internationally. Additionally, knowing that GenAI may make changes to our technical landscape, the state will explore specialized training and development curricula for current state employees to work successfully with GenAI technologies. Importantly, as it is an emerging technology, the State will continually evaluate the most responsible ways to implement GenAI.

Looking forward, the State will continue building off considerations laid out in the NIST framework for conventional AI as well as developments from the Biden Administration's Executive Order on AI, and in ongoing partnership with the Legislature, community, academia, and technical experts. As the State acknowledges GenAI's capacity to perpetuate and exacerbate bias, California will continue to lead the nation in removing barriers to equal opportunity and, ensure our systems do not promote explicit or implicit biases when providing key benefits and governmental services, while continuing to advance diversity, equity, inclusion, and accessibility. In that way, California can continue to lead the nation through thoughtful, equitable, and innovative deployment of some of the most promising technology to become available in a generation.

VI. Appendix

Policy Landscape References

To gain a comprehensive and balanced understanding of the benefits and risks of generative AI, it is essential to gather insights from a diverse range of sources across academia, government, industry, and civil society.

Important sources that were critical in informing this report included:

- The White House AI Bill of Rights
- The NIST AI Risk Management Framework
- Internal guidance policies on generative AI usage from local and state governments
- International AI governance frameworks
- The National Telecommunications and Information Administration (NTIA) Request for Public Comment on AI accountability
- Academic and civil society research and recommendations

This context will help illustrate the complex array of considerations around responsible AI development and set the stage for further examination of how state governments can navigate this emerging policy domain.

NIST AI Risk Management Framework

The NIST AI Risk Management Framework (RMF) working group develops flexible policies and guidelines for responsible AI governance that align with other major frameworks like the White House's proposed AI Bill of Rights. The RMF provides detailed questions and checklists to systematically guide organizations in implementing best practices for accountable and observable AI systems. An important note is that the RMF is intended as a standards framework and set of benchmarks rather than an off-the-shelf governance toolkit - it serves as a reference for organizations to develop their own tailored operational policies and toolkits. Alignment with the consensus standards and benchmarks in the RMF enables greater interoperability between different organizations' governance processes and tools that build on the same foundations.

Government internal guidance policies

Internal government guidance policies at the state and local levels have provided an initial policy backstop for public sector use of generative AI technologies. These developing guidance policies commonly outline considerations around privacy risks between using enterprise versus public-facing generative AI tools, disclosure and transparency requirements when governmental bodies utilize generative AI capabilities and cautions around the potential for hallucinated or factually incorrect outputs if generative models are not carefully monitored and tuned. While limited in scope, these internal government policies demonstrate early governance attempts to balance public sector opportunities from leveraging cutting-edge generative AI with responsible oversight.

Citations

- Information Technology Department Generative AI Guidelines | City of San José
- City of Boston Interim Guidelines for Using Generative AI
- Seattle IT Interim Policy

International AI governance frameworks

International AI governance frameworks and regulatory policies from bodies like the EU, UK, Canada, and others occupy critical niches in the overall AI governance ecosystem that substantially impact the compliance requirements and burdens for private sector technology companies to fulfill if they wish to operate globally. Pioneering and influential frameworks like the EU's General Data Protection Regulation (GDPR) have set the tone and expectations for privacy standards around AI and data utilization worldwide - many companies are wary of having to comply with multiple substantially different sets of national or regional AI regulatory requirements. As California explores potential policies for accountable state-level AI governance, policymakers should be aware of the major international governance frameworks being adopted elsewhere in the world in order to minimize regulatory divergence and compliance burdens.

Citations

- Interim guidance for agencies on government use of generative AI platforms | Australian Public Service
- The European Union AI Act
- China's Generative AI Policy
- AI Foundation Models Initial Report - GOV.UK
- White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

NTIA Request for Public Comment

The National Telecommunications and Information Administration's (NTIA) recent Request for Comment on frameworks and best practices for AI accountability gathered input from a diverse range of stakeholder groups including academia, industry, civil society, all levels of government, and individual constituents. This public consultation offered a valuable venue for surfaced feedback across the AI governance spectrum on assessed risks from current AI systems, suggestions for responsible AI development and fielding, and proposed organizational and technical solutions for improved AI accountability and observability. Reviewing responses submitted by the public helps inform more comprehensive approaches to steering responsible generative AI development and mitigating potential harms from misuse.

Academic, industry, and civil society research

Academic studies, industry papers, and civil society reports have offered a number of valuable and relatively comprehensive surveys of assessed risks and benefits from deploying generative AI systems across different societal domains. These analyses help inform a broader taxonomy of potential benefits and risks from the application of generative AI technologies in areas like finance, healthcare, criminal justice, employment, and more. Such knowledge mapping exercises highlight domains of concern and foreground issues for governance approaches aimed at maximizing generative AI's benefits while curtailing foreseeable risks from irresponsible use or negative externalities. These findings help inform priorities and nuanced approaches for governance that enable accountable and ethical generative AI utilization across diverse contexts. In particular, the IBM AI Ethics framework that characterizes Traditional, Amplified, and New risks for GenAI helped inform the structure of our risk analysis framework.

Citations

- Overview of Generative AI
  - Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality
  - Capabilities and risks from frontier AI
  - Holistic Evaluation of Language Models
- Risk Level Assessment
  - AI Act: Risk Classification of AI Systems from a Practical Perspective
  - The EU AI Act: Adoption Through a Risk Management Framework.
  - The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment | Brookings
  - AI Act: Risk Classification of AI Systems from a Practical Perspective
  - Contentious areas in the EU AI Act trilogues

    - Regulating Foundation Models and Generative AI: The EU AI Act Approach
    - European Parliament Adopts AI Act Compromise Text Covering Foundation and Generative AI - Data Matters Privacy Blog').
    - Deployers of High-Risk AI Systems: What Will Be Your Obligations Under the EU AI Act? - Kluwer Competition Law Blog
    - The case of the EU AI Act: Why we need to return to a risk-based approach
    - Analyzing the European Union AI Act: What Works, What Needs Improvement
    - The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment | Brookings
    - Ethical and social risks of harm from Language Models
- Validity & Reliability
    - Survey of Hallucination in Natural Language Generation
    - Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment
    - Is GPT-4 getting worse over time?
    - Generative AI Risks & Considerations Whitepaper – Trustible
    - The Curse of Recursion: Training on Generated Data Makes Models Forget
- Safety
    - A Categorical Archive of ChatGPT Failures
    - Three lines of defense against risks from AI
    - Adding Structure to AI Harm - Center for Security and Emerging Technology
    - AI Ethics | IBM
    - Taxonomy of Risks posed by Language Models
    - Can large language models democratize access to dual-use biotechnology?
- Security & Resiliency
    - The Gradient of Generative AI Release: Methods and Considerations
    - Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned
    - Red Teaming GPT-4 Was Valuable. Violet Teaming Will Make It Better | WIRED
    - OWASP Machine Learning Security Top Ten Risks
    - OWASP Top 10 Security Risks for LLM
    - Understanding the risks of deploying LLMs in your enterprise
    - Greylock: Securing AI
    - AI Red-Teaming is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-teaming for AI Accountability
- Accountability & Transparency

    - Guidance for the Development of AI Risk and Impact Assessments
    - Assessing Language Model Deployment with Risk Cards
    - Auditing large language models: a three-layered approach
    - HAI Auditing Algorithms
    - Release Strategies and the Social Impacts of Language Models
    - The Foundation Model Transparency Index
    - When is automated decision making legitimate?
- Explainability & Interpretability
    - Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment.
    - Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting
    - Decomposing Language Models Into Understandable Components
- Privacy
    - Generative AI: Privacy and tech perspectives
    - AI Browser Extensions Are a Security Nightmare
- Fairness
    - [NIST] Towards a Standard for Identifying and Managing Bias in Artificial Intelligence
    - Who's Afraid of Disparate Impact? – The Markup
    - Quantifying ChatGPT's gender bias
    - Fairness in AI and Its Long-Term Implications on Society