

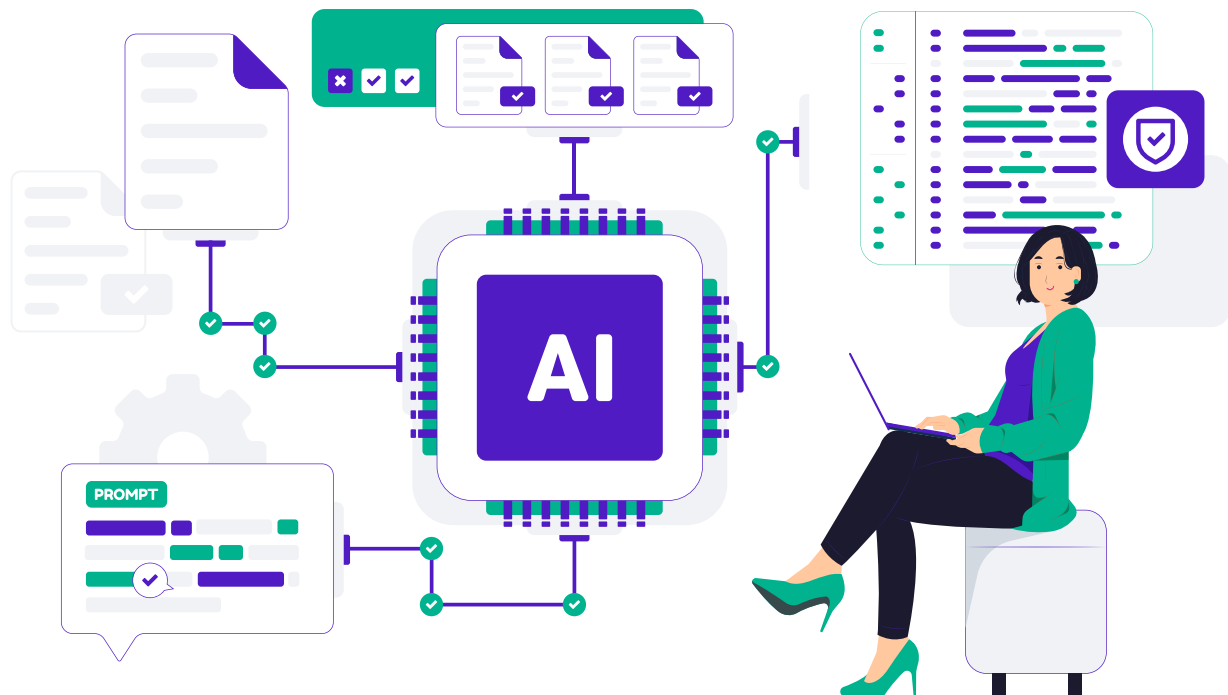
# Building a Secure Generative Artificial Intelligence Environment for Research Use

In today's rapidly evolving digital landscape, public agencies are increasingly turning to generative artificial intelligence (AI) technologies to enhance knowledge management, streamline operations, and deliver innovative solutions that benefit citizens. Innovative tools like ChatGPT and Microsoft's Copilot are already being deployed by some forward-thinking agencies.

While it is tempting to immediately begin integrating AI into data systems, it is important to remember that there are limitations to what AI can do. These powerful capabilities come with substantial responsibilities, particularly for public entities that operate under

stringent regulatory frameworks and are held to high standards of public trust. To harness the potential of AI while safeguarding against the misuse of sensitive data, inaccurate conclusions, or unfair biases, establishing a robust infrastructure with strict guidelines and governance is imperative.

This spotlight provides an overview of WestEd's implementation of a Secure AI Environment, including the risks, benefits, and effort of the implementation, to inform your organization's discussions on AI. The Data Integration Support Center (DISC) strongly recommends that prior to implementing any new technology, including AI tools, you have a solid understanding of



the risks, governance, benefits, challenges, staffing considerations, training, and specific use cases for your organization. To have productive conversations about how to effectively incorporate AI into data systems, it is crucial to begin with understanding the different types of AI; exploring the specific use cases that your organization may be considering; and developing a strategy that includes an honest assessment of the maturity of your data, systems, and staff skills to support and scale this technology.

DISC can support public agencies with facilitating a structured discussion and assist with strategy formation to ensure your organization builds a fundamentally sound foundation for AI.

## WestEd's Use Case

WestEd staff, both internal and client-facing, increasingly requested both access to and use of unvetted external tools to streamline their work and create efficiencies. These tools, however, had potential data integrity and security issues. While senior leadership saw value in the use of such tools, there were ongoing concerns around intellectual property, data use and ownership, and equity and bias in existing AI tools.

In one instance, a WestEd employee was traveling to facilitate an on-site meeting with a client. As the only project staff able to attend the meeting, she sought solutions that would allow her to both facilitate the meeting and take meeting notes. The employee requested the use of Zoom's AI note-taking assistant

to record notes while she facilitated the meeting. While WestEd's Information Technology (IT) department and Data Protection Office had not previously evaluated the use of Zoom's AI tools, this request prompted a review aligned with WestEd's established procedures. Following multiple requests for the use of various AI-based tools, WestEd's IT leadership recommended the establishment of an internal secure AI environment. To achieve this, it was necessary to embark on a process that would thoroughly assess the potential uses of AI tools and address ongoing concerns, a process that would clearly outline the benefits, efficiencies, costs, and risks associated with the use of AI in WestEd's work.

## Overview of WestEd Secure AI Environment

The formation of WestEd's Secure AI Environment is not just a technological upgrade; it is a strategic imperative that ensures responsible utilization of AI in the agency's work. The WestEd Secure AI Environment currently includes access to both an internal ChatGPT and an isolated data lake enclave with network-isolated, secure AI tools for research use.

The WestEd Secure AI Environment uses a two-pronged approach to implement data security, data compliance, and ethical AI use. This approach uses both a virtualized isolation computing environment and secure enterprise AI services deployments, as well as Azure Active Directory Single Sign for authentication.

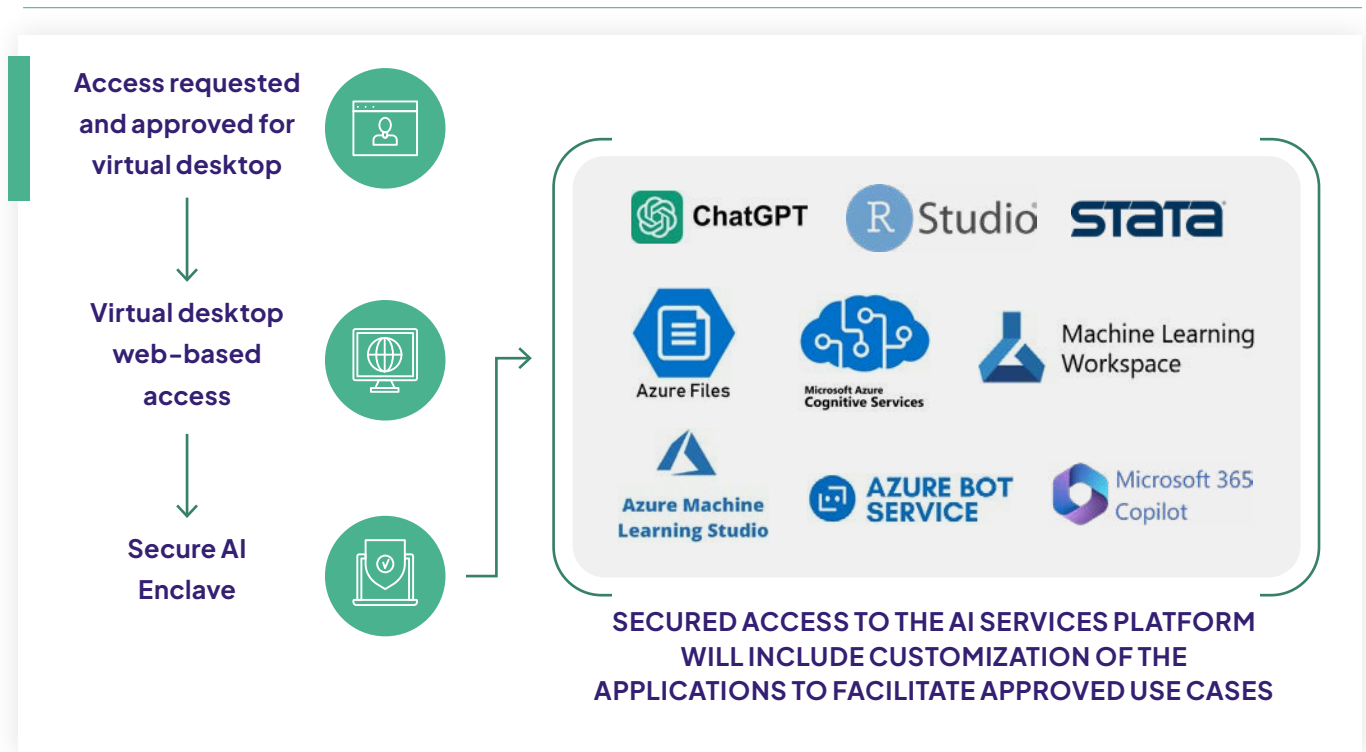
## WestEd Secure AI Environment Two-Pronged Approach

- 1. Azure Virtual Desktop** creates virtualized desktop environments that allow users to securely access WestEd's Secure AI Environment from various devices or locations.
- 2. Azure AI Service** provides a set of secured services, including the predominant Machine Learning models, with access isolated to authorized end users.

Authorized end users access the WestEd Secure AI Environment through Azure Virtual Desktop from their agency-issued, encrypted devices. The virtual desktop provides access to the WestEd Secure AI Environment, which is populated with AI services such as

- Azure OpenAI Bot Services for user interface configurations;
- Microsoft Machine Learning Studio for ML model creation;
- Artificial Intelligence Services, which include Computer Vision image recognition;
- Azure Cognitive Services for AI engine components;
- Stata and RStudio for data analysis;
- Azure Files for managed secure storage;
- Microsoft Copilot for AI Services Desktop Integration; and
- Secure ChatGPT models for prompt engineering and information services.

Figure 1. WestEd's Secure AI Environment



## Benefits

The advent of AI could transform the research landscape, offering unprecedented capabilities to analyze complex data sets and uncover new insights. The importance of building an environment that ensures accuracy, privacy, and reliability is paramount. The necessity of having a secure AI environment for researchers cannot be overstated because it stands as the safeguard against a multitude of risks and ethical concerns.

## Safety and Reliability

WestEd's Secure AI Environment provides a safe and reliable space for the experimentation and deployment of AI technologies to foster innovation and research due to a combination of technological safeguards, ethical frameworks, and governance protocols that are put in place to protect against various risks associated with AI deployment.

In one instance, an employee began using an unvetted AI note-taking and transcription tool for virtual meetings. While the employee only intended the AI tool to be used in one meeting, they soon found the AI-bot in all of their meetings, with access to their calendars and email and with no clear way to remove the AI. Even with seemingly benign tools for note-taking, the AI tool allowed the organization's proprietary data to be used in external AI programming that may be disseminated to unknown parties. While in this instance the organization's information was relatively benign, these issues are compounded when access includes confidential or private information.

**Figure 2. Components of Safety and Reliability**



### Data Security

#### Encryption

Data at rest and in transit is encrypted to prevent unauthorized access and to ensure that sensitive information is securely transmitted and stored.

#### Access control

Strong authentication mechanisms and role-based access control ensure that only authorized individuals can access the AI systems and data.

#### Data anonymization

When necessary, personally identifiable information is removed or obscured to protect individual privacy.



### Robust Infrastructure

#### Redundancy

Systems are designed with redundancy to avoid single points of failure and to ensure continuous operation even in the face of hardware or software failures.

#### Backup and disaster recovery

Regular backups and a well-defined disaster recovery plan ensure that systems can be quickly restored in the event of data loss or a catastrophic event.



## Model Integrity and Reliability

### Model validation

AI models are thoroughly tested and validated for accuracy and performance before deployment and are based on WestEd's own proprietary data.

### Bias detection and mitigation

Systems include processes for detecting and mitigating bias in data sets and algorithms to ensure fair and unbiased decision-making.

### Continuous monitoring

AI systems are monitored continuously for performance issues or anomalies, allowing for immediate corrective actions if needed.



## Compliance and Ethics

### Regulatory compliance

Secure AI Environments are designed to comply with all relevant laws and regulations, including data privacy laws like GDPR and sector-specific regulations like HIPAA for health care.

### Ethical guidelines

Ethical considerations are integrated into the design and operation of AI systems, aligning with principles such as transparency, accountability, and fairness.



## Transparency

### Interpretable models

Efforts are made to use or create AI models that are interpretable and explainable, providing insights into how decisions are made.

### Documentation

Comprehensive documentation of AI systems' design, data sources, and decision-making processes aids in understanding and trust.



## Human Oversight

### Human-in-the-loop

Systems are often designed to include human oversight where critical decisions are concerned, allowing for human intervention when necessary.

### Training and education

Personnel are trained to understand and manage AI systems responsibly, ensuring they can respond appropriately to any issues that arise.



## Quality Assurance

### Testing and auditing

Regular testing and auditing of AI systems help identify vulnerabilities, ensuring that the systems are secure and functioning as intended.

### Feedback loop

Mechanisms for feedback allow for continuous improvement of AI systems, increasing their reliability over time.

By integrating these practices into the design, implementation, and operation of AI systems, WestEd's Secure AI Environment is not only safe from cyber threats and data breaches but also reliable in terms of performance and ethical considerations. This holistic security approach is essential for fostering trust among users and interest holders and for ensuring the sustainable deployment of AI technologies.

### Protecting Intellectual Property

Public agencies also need to consider data that are not necessarily personally identifiable but that still need to be protected because they are proprietary in nature. Consider education assessment items, which include questions and prompts used in standardized assessments. Maintaining the confidentiality of assessment items helps maintain the integrity and validity of the standardized assessment, allows for reuse in future assessments, and prevents cheating. Utilizing publicly available generative AI tools to assist staff with modifying or enhancing assessment items may result in those items becoming widely available. This would result in damaging the integrity of the assessments or potentially allowing those items to be used for commercial purposes.

To mitigate that type of risk, WestEd's Secure AI Environment protects the agency's intellectual property by controlling the data entering the system, limiting access to users through established controls, and maintaining audit trails. Additionally, because the AI model itself is the agency's intellectual property, these safeguards protect their proprietary algorithms from unauthorized access or theft. This ensures that valuable intellectual assets remain confidential and protected, contributing to the long-term innovation and competitive advantage of WestEd.

**Figure 3. Components of Protecting Intellectual Property**



#### Data Protection

Secure AI environments implement advanced encryption and data anonymization techniques to protect the underlying data that could contain proprietary or confidential information.



#### Access Controls

Through strict access controls and authentication protocols, only authorized personnel can access AI models and the data they process, reducing the risk of IP theft.



#### Model Security

AI models themselves can be considered intellectual property. Secure environments use measures like code obfuscation and model encryption to prevent reverse engineering.



#### Audit Trails

Maintaining detailed logs of data access and model use helps in tracking any unauthorized attempts to access IP and provides evidence for legal recourse if necessary.

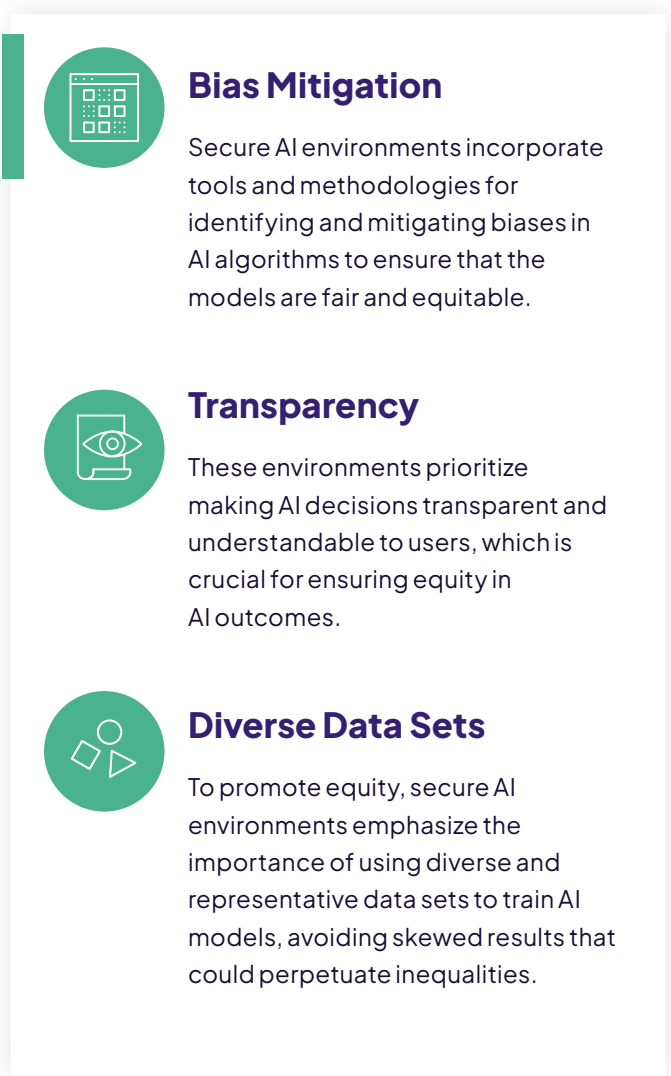
The combination of these measures protects WestEd’s intellectual property through the full life cycle associated with AI systems, from initial development to deployment and operational use.


### Supporting Equity


WestEd’s Secure AI Environment supports equity by ensuring that the benefits and protections provided by AI are distributed fairly and without bias. All the information used in WestEd’s Secure AI Environment was based on the agency’s own information. This limited quality errors arising from unauthenticated data and ensured that the foundational data for the AI algorithms limited biases. WestEd’s development team was transparent and inclusive in their decision-making, which allows users to understand how the AI model makes decisions and to ensure those decisions are fair and equitable.


For example, AI tools can be used to mitigate the cultural and racial biases in those same protected assessment items. Large language models can be used to identify phrases or words that have different connotations for different groups of students. Generative AI can be used to rephrase assessment items to meet culturally diverse test takers. Because the agency controls the foundational data for the underlying algorithms, staff can ensure the data used to train the AI model are as diverse and representative as possible. This helps prevent the model from learning and perpetuating biases present in the unauthenticated data. Additionally, applying de-biasing techniques during data preprocessing or model training, such as reweighting, resampling, or modifying the learning algorithm, can help reduce the influence of biases.

**Figure 4. Components to Support Equity**



- **Bias Mitigation**

Secure AI environments incorporate tools and methodologies for identifying and mitigating biases in AI algorithms to ensure that the models are fair and equitable.
- **Transparency**

These environments prioritize making AI decisions transparent and understandable to users, which is crucial for ensuring equity in AI outcomes.
- **Diverse Data Sets**

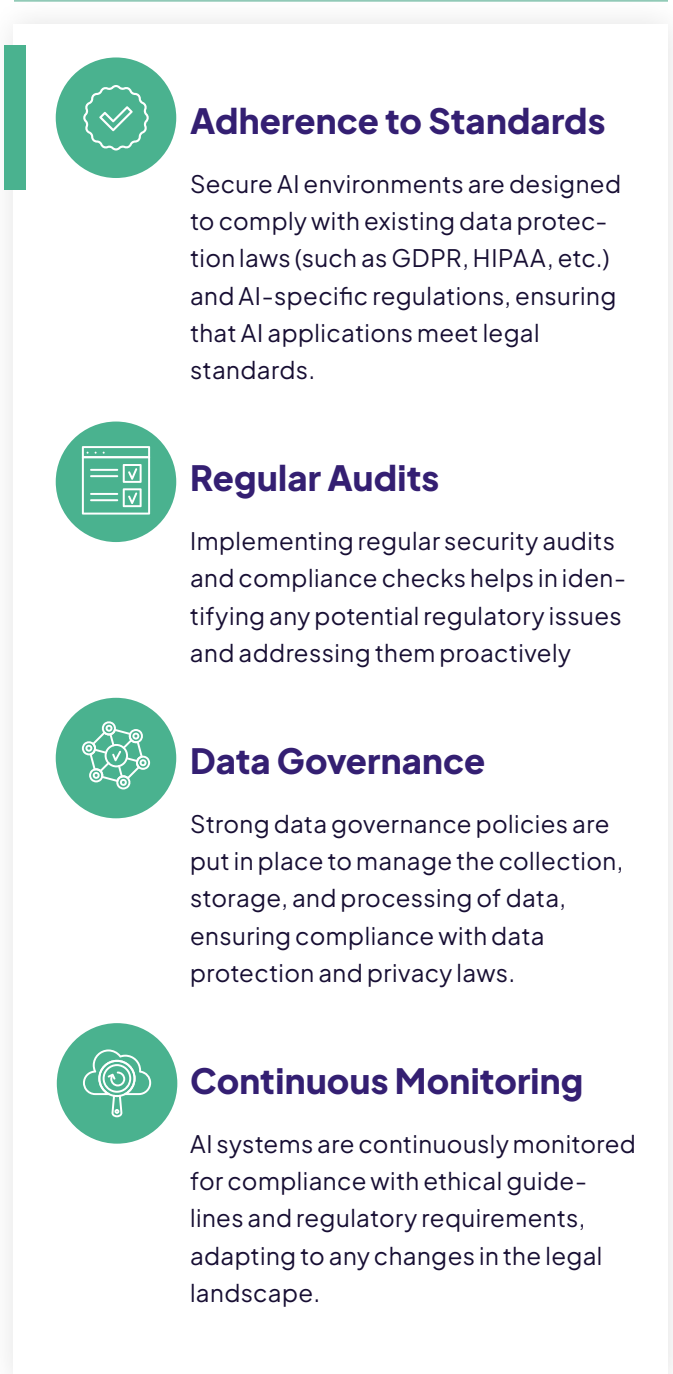
To promote equity, secure AI environments emphasize the importance of using diverse and representative data sets to train AI models, avoiding skewed results that could perpetuate inequalities.

By integrating these components, WestEd’s Secure AI Environment works to mitigate inequality and discrimination to support equitable decision-making.

## Ensuring Compliance With Regulations

Authorized and unauthorized disclosures are central concepts required in many privacy regulations. As articulated earlier in this brief, there are no guarantees of where data are shared or stored in standard AI configurations. Most public agencies, such as WestEd, house confidential data as required by state or federal reporting requirements. As such, compliance with regulatory requirements and standards, such as the Federal Risk and Authorization Management Program (FedRAMP), the Family Educational Rights and Privacy Act (FERPA), and the Health Insurance Portability and Accountability Act (HIPAA), were required features of WestEd's Secure AI Environment, ensuring that sensitive information is safeguarded and data are handled according to legal and ethical guidelines. Both technical and administrative measures designed to protect sensitive data and align with legal requirements are key components of WestEd's Secure AI Environment.

**Figure 5. Components of Compliance with Regulations**



This combination of components ensures that WestEd's Secure AI Environment operates within the legal framework established for the use and protection of the data within the model.



## How the Data Integration Support Center Can Help

The establishment of a secure AI environment, such as the WestEd Secure AI Environment, is crucial for public agencies to harness the potential of AI while ensuring data protection, privacy compliance, trust-building with interest holders, intellectual property safeguarding, and overall ethical utilization.

The Data Integration Support Center (DISC) at WestEd can facilitate productive conversations and assist in responsibly leveraging the power of AI for your integrated data system. DISC offers technical assistance to public agencies free of cost. Forms of technical assistance that DISC can provide include

- providing expertise on the scope and variety of AI to support data integration efforts;
- developing use cases that leverage AI tools aligned with the state's strategic priorities;
- conducting neutral, external assessment of the maturity of your integrated data system and capacity to support and scale AI technologies; and
- facilitating structured discussions to develop a foundation for the use of AI in your integrated data system.

© 2024 WestEd. All rights reserved.

Suggested citation: Rodriguez, B., El-Amin, A., Tiderman, L., (2024). *Building a secure generative artificial intelligence environment for research use*. WestEd.

WestEd is a nonpartisan, nonprofit agency that conducts and applies research, develops evidence-based solutions, and provides services and resources in the realms of education, human development, and related fields, with the end goal of improving outcomes and ensuring equity for individuals from infancy through adulthood. For more information, visit [WestEd.org](https://www.wested.org).

A project of **WestEd**   
WestEd.org