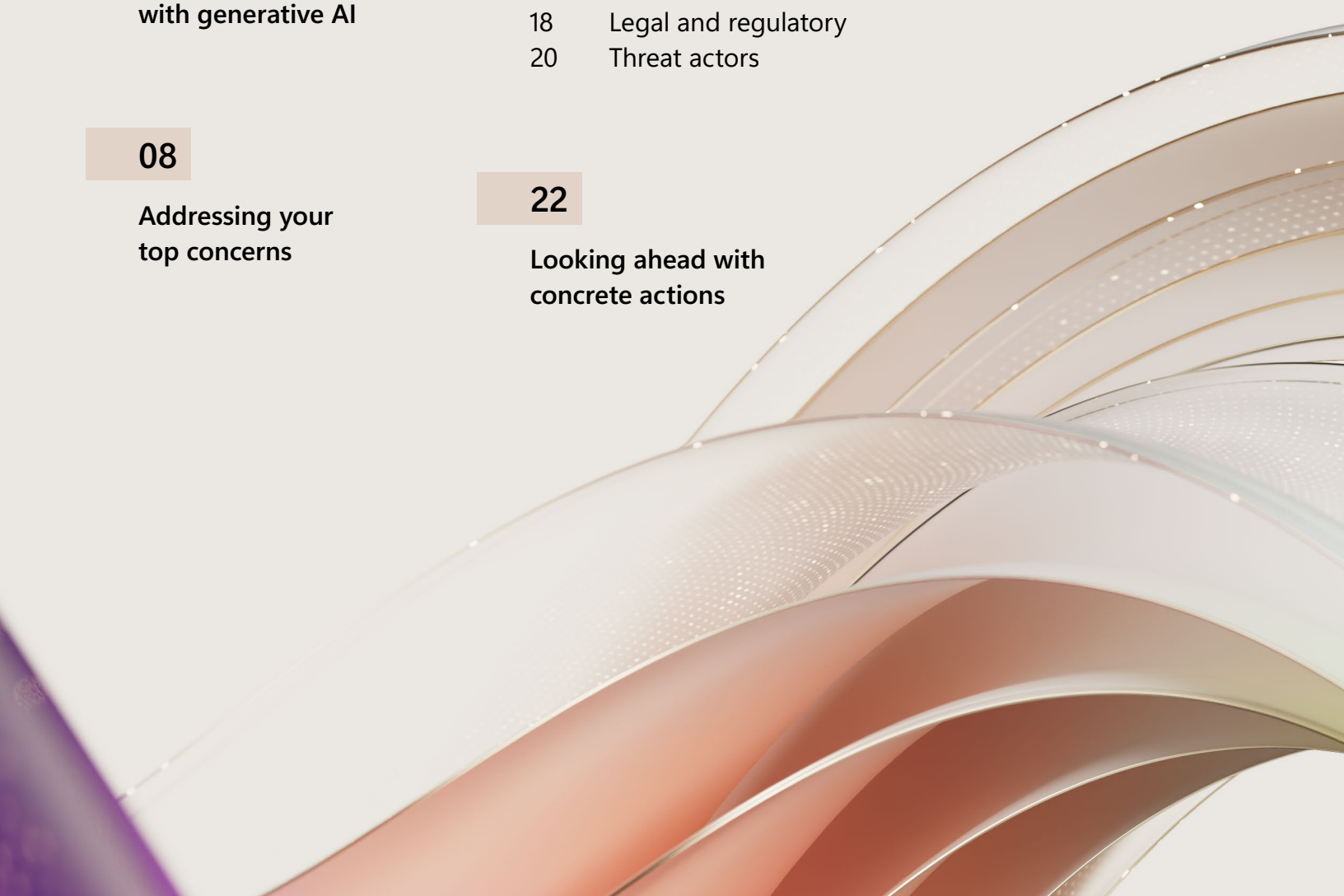
Microsoft

Minimize Risk and Reap the Benefits of AI

Addressing security concerns and implementing safeguards

Contents



Foreword

In the ever-evolving landscape of cybersecurity, one immutable truth persists: Change is the sole constant. Among the most promising and profound of these changes is the rise of generative AI. As security leaders we have an opportunity to drive innovation and foster a new age of enablement that is at the same time secure. We know that within the innovation hype cycle comes a realistic approach to adoption, and while we want to embrace innovation, we can also ensure that we're addressing the concerns surrounding data security and governance, transparency, regulatory compliance, and accountability to meet our customer requirements.

We know that to feel good about the AI solutions we adopt, we need to prepare our environment and even more importantly understand how security is applied across the AI supply chain. By understanding the inner workings of AI models, we can ensure that they operate ethically and responsibly, free from biases and hallucinations, and that our existing data and the data that is created by AI are kept safe and secure.

The goal of this paper is to provide best practices and guidance on how to address concerns around generative AI. We have been on technological transformation journeys before, and one thing that holds

Bret Arsenault, CVP,
Chief Cybersecurity Advisor

true is the importance of ensuring that we rely on security best practices and frameworks that we already apply in our existing environments like **Zero Trust**, National Institute of Standards Technology (NIST) Risk Management Framework, and data security and governance. Basic security hygiene still applies and is one of our best defenses against attacks. The fundamentals will always be important and apply to new technologies; however, with new technology we also see new threats and risks that we need to adapt to and create new security approaches. We outline and recommend some of these new security practices in the paper.

Becoming enablers of AI transformation

AI has the power to elevate human potential and solve some of the most serious challenges that we are facing—including threat intelligence, closing the skills gap in the cybersecurity workforce, and having a positive effect on diversity and inclusion in the industry. And while we know that with most transformative technology there will be risks and concerns, as security professionals, we need to be supportive and enable AI for our organization to be successful. I'm frequently approached with questions by customers and colleagues about how their company should approach AI adoption. From my perspective, initiating a safe environment for experimentation is key, and utilizing resources such as the responsible AI framework and principles is beneficial. At Microsoft, when we adopt new groundbreaking technology, I recommend a simple framework known as the four Es: Experiment, Effective, Efficient, Exploit. Start by creating an environment that encourages teams to experiment and learn about generative AI, ranging from user experiences to developing and operating AI applications. It's crucial to understand data protection requirements, classification, and necessary controls for broader AI adoption. Next, focus on effectiveness. Identify use cases that ensure the solution delivers

[1] Microsoft Digital Defense Report 2023

anticipated results, and then explore ways to increase the solution's efficiency. Consider alternative approaches that could improve both effectiveness and efficiency. Finally, exploit is when, equipped with practical knowledge and proven progress along the adoption curve, one can take a step back to assess how you can utilize and take full advantage of the capabilities.

This paper will discuss AI security risks and how to address them. We will share some of our learnings, safeguards, and recommended approaches. I hope you find our guidance valuable, and as always, I welcome feedback and would love to hear about your experiences in implementing generative AI.

I am excited for the future of cybersecurity as we enter a new phase of innovation. We have an opportunity as leaders in the business to lean in and enable our organization with a risk-based approach to adopting AI and to make sure that we use it in a responsible and secure way.

Bret Arsenault
CVP, Chief Cybersecurity Advisor

Our new world with generative AI

Generative AI is changing the world, and various groups in your organization may want to use different applications because of the various advantages it can offer the business. For example, you probably have seen some early adopters in your sales and marketing organizations who want to use generative AI to enhance communications with customers or provide better pricing data as part of the request for proposals process. Creating a proposal and calculating a price using different pricing models can be tedious and inefficient—however, with generative AI, it can be done within minutes!

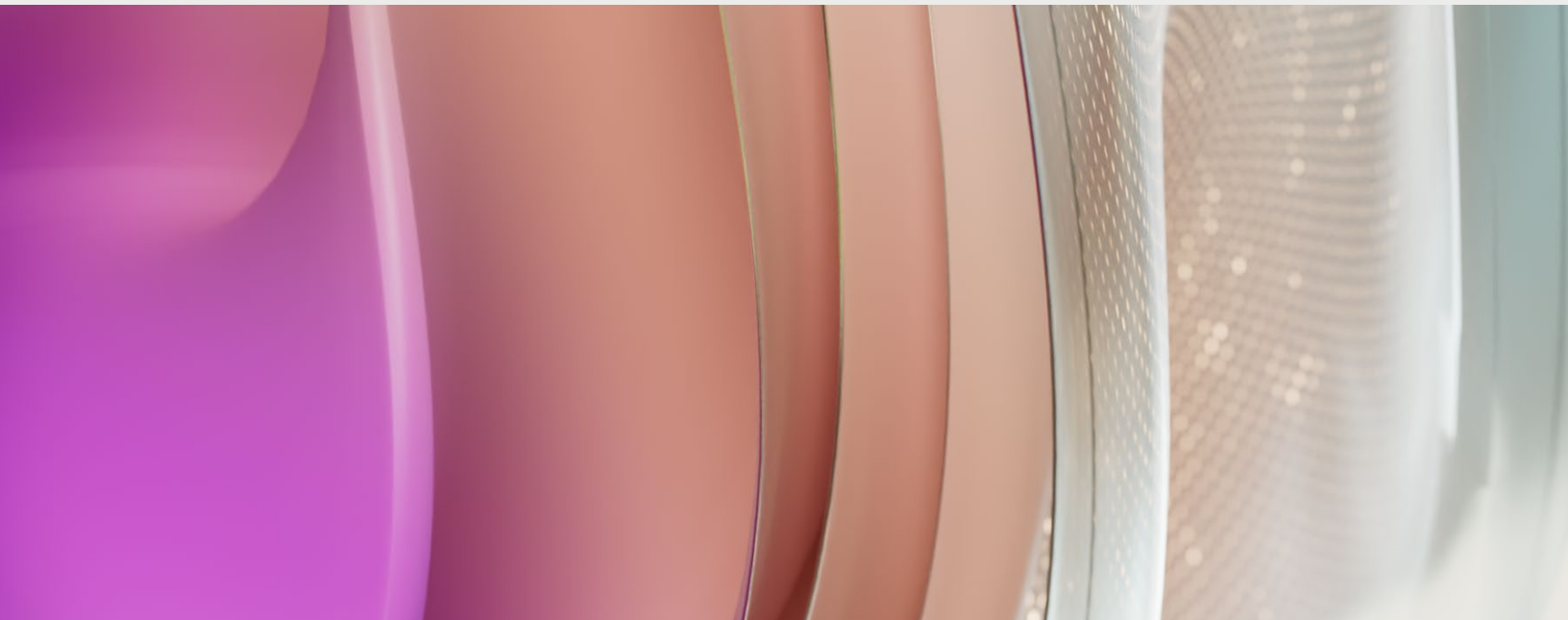
Although it may appear harmless for a sales manager to leverage generative AI to determine pricing for a proposal based on different Excel spreadsheets, the concern you may have is the quality of the data that is being retrieved to calculate the pricing. Is this data accurate and reliable? Is it a

hallucination, where incomplete or outdated data inputs could lead to ungrounded responses or overreliance on flawed insight? While the sales manager may be thrilled that within minutes they have created an attractive proposal with a strong price incentive, it's important to double-check the information generated to avoid business errors or misleading results.

As we explore how generative AI results impact the full business in this scenario, consider a global supply chain with international customers. You may ask yourself the following questions: Is the pricing information provided in accordance with the accounting and pricing standards of that country? Could this cause legal or regulatory challenges and does this cover the intricacies of the conversion rates? What about data access? Will all sales professionals have access to this pricing data? Will the pricing information leak out to the public or competitors?

What about your development teams that look to generative AI to help them build applications efficiently? How do you protect the code and ensure the data provided to AI is secure? If your customer support teams or your customers are relying on its answers, what will you do when it gives wrong answers? What about copyright infringement? These are just a few situations among many examples of teams wanting to realize the benefits immediately.

For that reason, we use this paper to address the main concerns and risks of implementing generative AI, describe what we have learned from our work developing Responsible AI, and provide insight into how to safeguard these risks and what to do to secure your future with AI.



Top security and business leader concerns[2]

1 Data security (82%)	Leakage and risks for data privacy, ownership of intellectual property, etc.
2 Hallucinations (73%)	Hallucinations (inaccurate outputs) and misunderstanding of process and algorithms
3 Threat actors (60%)	Using AI to expose confidential information, create vulnerabilities in security models, and cause data distortion
4 Biases (57%)	Reputational risks from possible erroneous, harmful, biased, or distressing output
5 Legal and regulatory (55%)	Lack of understanding of how it is and will be regulated, among different industries

Addressing your top concerns

According to a recent ISMG (Information Security Media Group) survey, **roughly 30% of respondents report holding back on adopting AI across their organizations**, based on security risks and concerns.

ISMG's survey with business and cybersecurity professionals shares five key AI implementation concerns: (1) data security, (2) hallucinations, (3) threat actors, (4) biases, and (5) legal and regulatory compliance.

ISMG researchers surveyed more than 400 business and cybersecurity professionals on their current and intended use cases, concerns, and challenges regarding generative AI. While business leaders appear more enthusiastic about adopting generative AI than their cybersecurity leader peers, both groups share AI implementation concerns.

[2] "First Annual Generative AI Study: Business Rewards vs. Security Risks," ISMG, 2023

Safeguards you can implement to mitigate risks of generative AI

The risks of generative AI split into two categories. First, AI is software, and needs to be secured just like any other application. Traditional software security principles and best practices should be applied, especially around access controls to the underlying data.

Second, there are novel threats that come when users interact with the AI as an AI application, by speaking to it. Because language is complex and infinitely varied, these are best understood as nondeterministic threats, ones which aren't subject to patches and remediations of the traditional security sort. But this isn't a showstopper—we already know how to build secure systems out of nondeterministically insecure components, namely people, and the approaches we use there. Training and having multiple sets of eyes look at systems apply very effectively to AI. In this paper we will look at five different risks and safeguards within these two categories.

1. Data security

Securing generative AI begins with data security because data is central to AI. Chief among the data security considerations facing leaders today is the issue of data leakage, which is when generative AI systems unintentionally reveal sensitive data to which they have access. Security concerns around data include (1) leakage of sensitive data, (2) over-permissioned data, and (3) inappropriate internal data exchange. Let's address how to safeguard these risks.

Data security risks and safeguards

Leakage of sensitive data

The data risks of generative AI are narrower than they seem. These systems can access data in only two ways. First, by way of an API call that searches for or otherwise fetches the data. This is often called "retrieval-augmented generation," or RAG, and is preferably done with the end-user's credentials. A second way is by using the models themselves, which can be fine-tuned on a private data repository. Both these paths can be secured through traditional methods. Finally, the outputs of the AI system should also be properly classified for use in the rest of your systems.

Securing RAG is an identical problem to securing enterprise search. If you have not done so already, you should expedite data security and governance. To protect your organization from data leakage, you should enforce data labeling and apply digital labels to content based on your classification policy. This helps ensure that sensitive data is appropriately tagged, making it easier to identify and control its distribution.

It's recommended that you implement stringent data permissions within your organization and assign user permissions based on their roles and group memberships. This means only authorized individuals will have access to sensitive information, reducing the risk of unauthorized exposure. By combining these measures, you can effectively manage and protect sensitive data in accordance with your security policies.

You should also use a governance solution that includes retaining and logging interactions with AI apps, detecting any regulatory or organizational policy violations when using those apps, and investigating incidents once they arise.

It's also recommended that you implement data lifecycle management and data minimization whenever possible. After all, the easiest data to secure is the data you don't even have.

When securing a fine-tuned model, the important thing to know is that generative AI systems cannot deterministically provide access controls among the information they know intrinsically. This means that anyone who has access to a fine-tuned model will have access to the underlying data that was used to train it. As models are most frequently exposed as program features, this translates to access controls on those features. While there are some methods in private data analysis that allow for aggregation and anonymization of data prior to training, these are very nuanced. Implementing them correctly requires specialist expertise on par with what is required to implement cryptographic primitives and should be avoided whenever possible.

Over-permissioned data

To deal with data over-permissions and users accessing and handling inappropriate data, organizations can use **data scanning** to find and sort data into categories based on how sensitive or frequently it's used. This helps adjust permission settings, so users get access to only what they need. Doing regular scans can keep you on top of managing permissions and lowers the chance of unauthorized access.

Inappropriate internal data exchange

To reduce the chances of data leaks in a multi-user environment, it's crucial to establish clear communication barriers based on **compliance boundaries**. These barriers serve as checkpoints that prevent sensitive data from crossing over between different groups. You can achieve this by implementing straightforward protocols and user-friendly technologies that restrict the movement of confidential information.

By setting up these structured compliance boundaries and utilizing encryption, access controls, and network segmentation, you can effectively contain data within its designated compliance areas. This approach not only ensures regulatory compliance but also minimizes the risk of unauthorized data leakage. Regular audits and monitoring will help to reinforce the effectiveness of these measures over time, ensuring ongoing protection of sensitive information.

Additional data security safeguards

Organizations can further fortify their data security posture with the following:

- Implementing a **Security Development Lifecycle** (SDL) to embed security measures into the code supporting access to AI models from the outset.
- Employing **content moderation** to promptly remove harmful AI-generated content, thereby mitigating risks.
- Utilizing **AI red teaming learnings and pen testing** to rigorously test systems for vulnerabilities, proactively addressing potential threats.

2. Hallucinations and overreliance

A common concern with generative AI is that the systems can hallucinate, which means to produce outputs that are not grounded in factual basis. This is an intersection of two problems that can occur: overreliance (when people overestimate the reliability of large language model (LLM) output) and ungroundedness (a combination of hallucinations and omissions—that is, false positives and false negatives).

Overreliance comes in four types:

1. Naive overreliance happens when people are unaware that the LLM is potentially wrong and that they should be cautious with its outputs.
2. Rushed overreliance happens when the operator doesn't have (or take) the time to check outputs —possibly because of a fast-paced environment, or because of confirmation blindness.
3. Forced overreliance happens when the operator physically can't check the outputs, e.g., when doing vision augmentation for people with visual impairments or when creating interactive websites for nonprogrammers.
4. Motivated overreliance happens when the AI is being used as an excuse to come to a conclusion that the operator wanted to reach, e.g., that a person is behaving suspiciously and should be stopped and searched.

The best mitigations to overreliance are generally a combination of UX and business process: using the systems in ways that will drive appropriate attention.

Ungroundedness happens when a generative AI system is set up so that the user expects outputs to be grounded in a source (e.g., a search of reputable data, or an input to be summarized) and it fails either by hallucination (producing data not in the input) or omission (dropping critical data from the input). This is a problem when this leads to either overreliance (bad outputs not understood as such) or making the entire system useless. It needs to be remedied in the software architecture of the generative AI system.

Hallucination risks and safeguards

Managing overreliance

Grounding will never be perfect, and the underlying data may itself be incorrect. When AI is used to drive critical decisions or create high-impact artifacts, it's important to have multiple people examine the result.

To minimize naive and rushed overreliance, consider the places where it makes sense to stop for human oversight. A useful rule is that it's best suited to tasks that are difficult for humans to do but easy for humans to verify. Consider ways to divide tasks into subsections whose intermediate outputs can easily be reasoned to maximize the efficacy of human checks. Also, pay close attention to framing within the user interface: For example, when a proposed draft email is written directly into the content window, it is far more likely to be quickly accepted than when it's in a secondary window, and the proposed action to the user is to keep the results rather than to send or save them.

In situations where naive, rushed, or forced overreliance is a risk, providing user training for AI is essential. Users should be educated about the concerns and risks associated with AI so that they can develop critical thinking skills to evaluate AI-generated content and engage in editorial review processes to ensure the accuracy and reliability of the content before dissemination.

Motivated overreliance is perhaps the most dangerous possibility, and it can be mitigated only through a wider look at business practices. For example, someone might use "because the AI said so" as an excuse for malfeasance.

Ungrounded and inaccurate responses

Overreliance manifests as a problem only when the LLM makes a mistake. For there to be a mistake, there needs to be some ground truth which the user was expecting the LLM to base itself on. Generally speaking, the LLM's own knowledge gleaned during training is not a good source because most LLMs are trained on the broad spectrum of human communication. LLMs do not have intrinsic knowledge of what is true or false, especially in your company's specialized environment. Talking directly to an LLM and expecting a usefully grounded result is guaranteed to fail. Custom AI agents are software that use LLMs as components to solve a problem. Within these, LLMs can be used repeatedly and in different ways to solve groundedness.

For example, a system intended to answer questions based on a repository of data (such as a documentation site) might proceed in stages. First, it translates the question to a set of search queries and performs the searches. Second, it examines each of the resulting documents and extracts relevant facts to the answers. Third, it combines those answers into a user-facing narrative. Fourth, it performs an editing pass, checking that every item in the final narrative can be footnoted with a link to the original documents, and that no critical item in the list of facts was omitted. This last step—asking the system to act as its own editor—is an example of how groundedness can be resolved in production systems.

When evaluating systems built by others for prospective deployment, it is useful to explicitly check for ungrounded responses, keeping an eye on both false positives and false negatives. Human oversight can provide feedback and rectify any errors or inaccuracies in the AI-generated content. Feedback loops where feedback is given through the user interface also foster correction of errors. When building your own generative AI application, you can find **new tools in Azure AI** to help you build more secure and trustworthy generative AI applications that solve for groundedness, and help protect your LLMs against prompt injection attacks with prompt shields and evaluate your LLMs for risks and safety.

Additional hallucination safeguard: AI red teaming

AI red teaming can help assess the accuracy and reliability of AI-generated content, identify vulnerabilities to malicious prompts, evaluate data access and processing for grounded responses, and mitigate the risk of overreliance on potentially inaccurate outputs.

AI red teams proactively conduct thorough assessments and simulations to enhance the overall reliability and effectiveness of AI systems, mitigate potential harm, and bolster trust in AI-generated outputs.

Depending on your situation, it's often possible to train existing people within your organization to do AI red teaming. The critical skills for this are the ability to think like an adversary and creatively, so a wide range of people may prove very good at it. Open-source tools like **Python Risk Identification Tool** (PyRIT) can greatly boost the efficacy of an AI red team.

For additional information on AI read teams, please see **Microsoft AI Red Team | Microsoft Learn**.

3. Biases

The risk from biases stems from AI producing erroneous, harmful, or embarrassing output, such as social profiling, emotional damage, and hate speech. This may also apply to custom data sets where a business requires accurate reporting but somehow ends up with biased answers due to the data constraints.

There are two areas where bias and discrimination come to play: (1) bias on training data and (2) overcontrolled generative AI systems.

Biases, risks, and safeguards

Bias always needs to be understood in the wider context of the way the system is being used. If it's being used to make decisions, will it implicitly consider factors that you did not want? If it's being used to create content, will it create content that harms or upsets people unintentionally?

Understanding and measuring bias

The first step in assessing bias is to figure out the kinds of bias that might be relevant to your system. Here it is useful to consider both a "target-first" approach (i.e., enumerating the potentially affected stakeholders and thinking of ways in which an output might affect them) and an "attacker-first" approach (i.e., enumerating people who might attempt to misuse the system to generate biased results for their own purposes). "Design from the Margins" is a highly recommended approach. By designing the system to work well for the most impacted and marginalized users, the resulting system is more likely to work well for all users.

With a sense of the relevant bias risks, the next step is to measure the system's actual behavior. Depending on the situation, this may involve test suites of inputs designed to elicit biased responses, or it may involve large sets of test inputs where good statistical behavior is required. In either case, making the measurement as automated as possible will save you work when the system is upgraded.

Mitigation is often most effectively done by adjusting the metaprompt, driving the system to take note of and account for its biases, much as we do with humans. However, you should always take care to test the mitigated behavior as well so as not to be embarrassed by its output.

Bias on training data

You need to be concerned with training data bias only if you're building your own models or fine-tuning models. If you do find yourself in such a situation, consider bringing in experts in debiasing models. Techniques include auditing datasets to ensure diversity, use of tools and frameworks such as AI Fairness 360 to measure bias, and a range of technical methods including data augmentation, resampling, preprocessing, and fairness regularization.

Over-controlled generative AI systems

Controls put in place to prevent harm in generative AI development can potentially cause additional problems. A vital way to avoid overcontrolling your generative AI system and block its efficacy is to allow active feedback loops. By enabling human feedback, users can provide real-time input on the effectiveness and impact of controls. This feedback mechanism not only empowers users to identify potential issues but also enables swift adjustments to controls to prevent unintended consequences. Additionally, incorporating feedback loops fosters a collaborative approach between users and developers, promoting continuous improvement and adaptation of security measures to strike the right balance between protection and usability.

Organizational components

There are three organizational components that can help eliminate these risks. The first is a strong ethics committee, empowered to keep an eye on the use of AI across the organization and raise concerns with sufficient political force to prevent bad outcomes. The second is investment in diversity and inclusion. This is critical because ideation about the ways in which the system might produce bad outcomes for particular stakeholder groups is difficult to do when you don't have personal experience of the issues diverse groups face. Empowering people to raise issues and having a wide range of voices in the room are a reliable way to stop problems before they happen. Finally, a strong listening system within your organization and its trusted users are key for discovering problems at an early stage. Many of these problems are apparent to nontechnical users and AI experts alike, so getting your tools in front of a wide range of people to test and providing a feedback loop is a fast way to address the risk.

4. Legal and regulatory

Unanswered questions about legal and regulatory obligations drive hesitancy for all of us—and it's even more significant with the board of directors. Ambiguous regulatory boundaries make it difficult to discern which aspects of AI development and deployment require oversight, and where protection against potential risks, such as privacy breaches and algorithmic bias, may be required. Moreover, the global disparities in AI regulation pose challenges for businesses operating across various jurisdictions, leading to inconsistencies in addressing ethical concerns and ensuring responsible AI use.

Decision-makers are cautious about striking a balance between fostering innovation in AI technologies and implementing adequate protection against privacy breaches, along with such concerns as 1) compromised ethics posture, 2) AI supply chain issues, and 3) human rights violations.

Legal and regulatory risks and safeguards

Compromised ethics posture

Inappropriate accountability and response to ethical concerns can eventually disrupt and damage an organization's reputation. To address ethical faults, it's crucial to incorporate the intended ethics policies and goals into the Security Development Lifecycle (SDL). Doing so protects the code and reflects the ethical posture into the security measures and outputs, as well as reinforces these actions through establishing an Ethics Committee dedicated to defining and upholding ethical standards in AI usage.

Companies should rely on established frameworks that can be applied to AI deployment and development, such as utilizing the NIST AI Risk Management Framework (RMF) to guide design choices, helping to identify and manage risks associated with AI implementation. By combining these measures, organizations can effectively mitigate ethics-related data security issues, reducing the likelihood of disruptions and reputational damage.

AI supply chain issues

Lack of transparency or accountability between partners and providers can cause supply chain issues in an AI environment. To address these risks, it's primarily important to integrate an SDL to protect code at every stage of development within the supply chain, ensuring that robust security measures are implemented according to all parties' standards, followed by human oversight for continuous feedback and response to errors.

It's also vital that organizations that are a part of the AI supply chain adopt an AI Shared Responsibility Model. This helps ensure accountability among all stakeholders and that they leverage the NIST AI RMF to guide design choices, enabling systematic identification and mitigation of risks throughout the supply chain.

Finally, organizations can establish an AI-focused user policy to provide clear guidance and rules for interactions with AI apps. This reduces vulnerabilities and enhances overall security posture. By implementing these measures, organizations can mitigate AI supply chain issues and improve transparency and accountability across the entire ecosystem.

Human rights violations

Within the fast-developing landscape of generative AI, international regulations will develop to protect human rights and the use of AI, and failure to comply will have serious consequences for the organization. To address the potential human rights violations stemming from AI use and ensure compliance with international regulations, organizations must establish an ethics committee composed of diverse experts. This committee will be tasked with defining and safeguarding the ethical use of AI, ensuring that data security concerns are addressed.

Additionally, implementing the NIST AI RMF will guide design choices, enabling organizations to proactively identify and mitigate potential risks to human rights. By integrating these safeguards into their AI development and deployment processes, organizations can uphold human rights standards and avoid serious consequences for noncompliance.

5. Threat actors

Many have concerns about threat actors trying to take over AI systems to access sensitive data or manipulate its output. There have been emerging AI threats and threat actors leveraging AI to orchestrate cyberattacks at scale, particularly with social engineering. Also, cybercrime-as-a-service (CaaS) providers can help anyone become a threat actor with a few tools.

Here are some safeguards to consider for these three areas of concern: 1) malicious model instructions, 2) jailbreaks into an AI system, and 3) defending against newly developed AI-driven attacks.

Threat actor risks and safeguards

Malicious model instructions

To mitigate the risks ranging from insider threats to external prompt injections that can lead to harmful instructions being processed by an LLM, we recommend a combination of measures. First, a prompt filter can inspect inbound user prompts and flag potential risks or inappropriate requests before they reach the AI system. Second, spotlighting software can mark external data in a special way and add metaprompt components that help the system ignore commands therein. Third, an output filter can scrutinize AI-generated content prior to delivery to users, ensuring that no harmful instructions are included.

To prevent unauthorized access and tampering, deploy strong authentication measures, including robust identity and access management policies for users, devices, and applications. Implementing a comprehensive insider risk program is essential for detecting and responding to internal threats effectively, thereby safeguarding against malicious actions within the organization and from external prompt injections.

Jailbreaks into an AI system

Threat actors can use AI to bypass or break through the security controls or restrictions placed on the AI system, potentially allowing the system to perform unauthorized actions or access unauthorized information in what could cause security breaches or misuse

of data. This scenario can be prevented by utilizing metaprompt guardrails and frameworks through prompt engineering. These guardrails help establish boundaries and guidelines for AI interactions, minimizing the risk of manipulation or exploitation by malicious actors.

Additionally, implementing a robust insider risk program is essential. This program should focus on detecting and responding to insider threats effectively, including potential attempts to jailbreak controls or misuse AI capabilities to access sensitive data.

Defending against newly developed AI-driven attacks

To counter the escalating speed and sophistication of AI attacks, organizations must start their defenses from the inside, establishing strict data permissions based on user roles and group memberships. By controlling access to sensitive data, organizations can mitigate the risk of unauthorized use or exploitation by malicious actors, including insiders. Additionally, implementing a robust insider risk program is essential. This program should focus on detecting and responding to insider threats effectively, as insiders can exploit their access to perpetrate AI attacks.

Another important concept to keep in mind when defending your organization against threat actors is threat modeling. AI threat modeling is about imagining the worst things that could happen if a bad actor took over the generative AI system or manipulated its data, and what steps can be taken to both prevent and detect if any stage of the attack path is successful. It helps understand the maximum damage that could occur and guides how to protect against such scenarios, while improving the communications, knowledge sharing, and critical thinking required for a strong security posture. To be successful, this process requires a combination of a diverse set of contributors, simplified and clear communications, and involvement from every key stakeholder across business and technical domains.

Looking ahead with concrete actions

We understand how generative AI can revolutionize the way we approach security. With this great reward comes a set of risks and concerns that need to be addressed so we can be enablers of innovation. Microsoft has been at the forefront of fostering a regulatory environment that promotes the safe and responsible deployment of AI through initiatives such as the **responsible AI principles** and the establishment of advisory boards dedicated to AI ethics. It's important to always use responsible AI principles like fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability when deploying or developing generative AI. One of the best ways to approach implementing responsible AI principles is to leverage the map, measure, and manage framework as a guide:

- **Map:** Mapping risks is a critical first—and iterative—step toward measuring and managing risks associated with AI, including generative AI. Mapping informs decisions about planning, mitigations, and the appropriateness of a generative application for a given context.
- **Measure:** Measuring AI risks and related impacts informs how we manage these considerations when developing and using generative applications.
- **Manage:** Once the first two steps are taken, you can then manage or mitigate identified risks at the platform and application levels.

Through Azure, Microsoft provides open-source tools, commercial tools and services, templates, and guidance to help organizations build, evaluate, deploy, and manage AI systems responsibly.

Lastly, as we do with any adoption of transformative technology, try to experiment with the new technology in a controlled setting to gain insights into the generative AI applications. This is where you will learn about data protection requirements, classification, and necessary controls. Through experimentation and learning, efficiency and effectiveness improve over many trials and time, enabling broader deployment within your organization.

To summarize, we recommend the following:

Data security and governance

Create a model for data access based on audience privileges. Establish policies to ensure data sensitivity is properly identified, content is labeled, permissions are reviewed, and data sharing is controlled by automated policy.

Use a data security and governance solution that includes retaining and logging interactions with AI apps, detecting any regulatory or organizational policy violations when using those apps, and investigating incidents once they arise.

- AI content anomaly detection
- AI activity logs and investigations
- Access governance for AI

Hallucinations

Should your organization plan on building a manual process with dedicated people to run AI governance, you should build strong prompt engineering and metaprompt framework capabilities to protect the input when sending requests to an LLM.

Maintain human monitoring to detect when potential failures have happened, and alert teams to unexpected or serious outcomes. Always have good mechanisms of user feedback in place.

Apply AI red team and penetration testing to proactively find vulnerabilities and unexpected behaviors before they pose a serious risk to the organization.

Leverage Microsoft Prompt Shields, groundedness detection, and other responsible AI tools to help customers build more secure and trustworthy applications.

Biases	Implement an ethics framework and committee or virtual team to define principles, and then monitor and respond to ongoing changes in perception about generative AI uses. Apply filters that can detect inputs likely to trigger problems as well as overtly bad output. Ensure human feedback is in place to detect when potential failures have happened, and alert teams to unexpected or serious outcomes.
Legal and regulatory	While new regulations can take time to materialize, take proactive steps to implement and build upon new government-led AI safety frameworks of the U.S. National Institute of Standards and Technology, or NIST. Update policies to call out specific behaviors and actions that may be prohibited when using generative AI.
Threat actors	Practice threat modeling and scenario mapping to help limit potential damage a single user interaction can have, without additional security checks in place. Make sure data loss prevention, privilege management, and insider risk indicators are in place. Implement conditional access policies for tailored security and enforce multifactor authentication to mitigate unauthorized access effectively. Separate internal IT systems from customer-facing AI systems, including the addition of operational components such as secure access workstations (SAW). Follow best practices for deploying secure AI systems.

To further mitigate the risks of generative AI, it is important to rely on proven security best practices and frameworks such as Zero Trust, NIST AI Risk Management Framework, Deploying AI Systems Securely, and Microsoft Purview Data Governance. By understanding the inner workings of AI models, we can help ensure ethical and responsible operation free from biases and unintended consequences.

At Microsoft, we are committed to partnering with the industry and our customers to advance the safe and beneficial adoption of generative AI. Through collaboration and ongoing research, we aim to develop and implement technologies and policies that promote transparency, accountability, and ethical use of generative AI. Together, let us benefit from the promise of AI and chart a course to a more resilient cyber-secure world.

Get more guidance in An Introduction to Generative AI and Safety and learn more about AI transformation at Microsoft.com/AI.

Microsoft