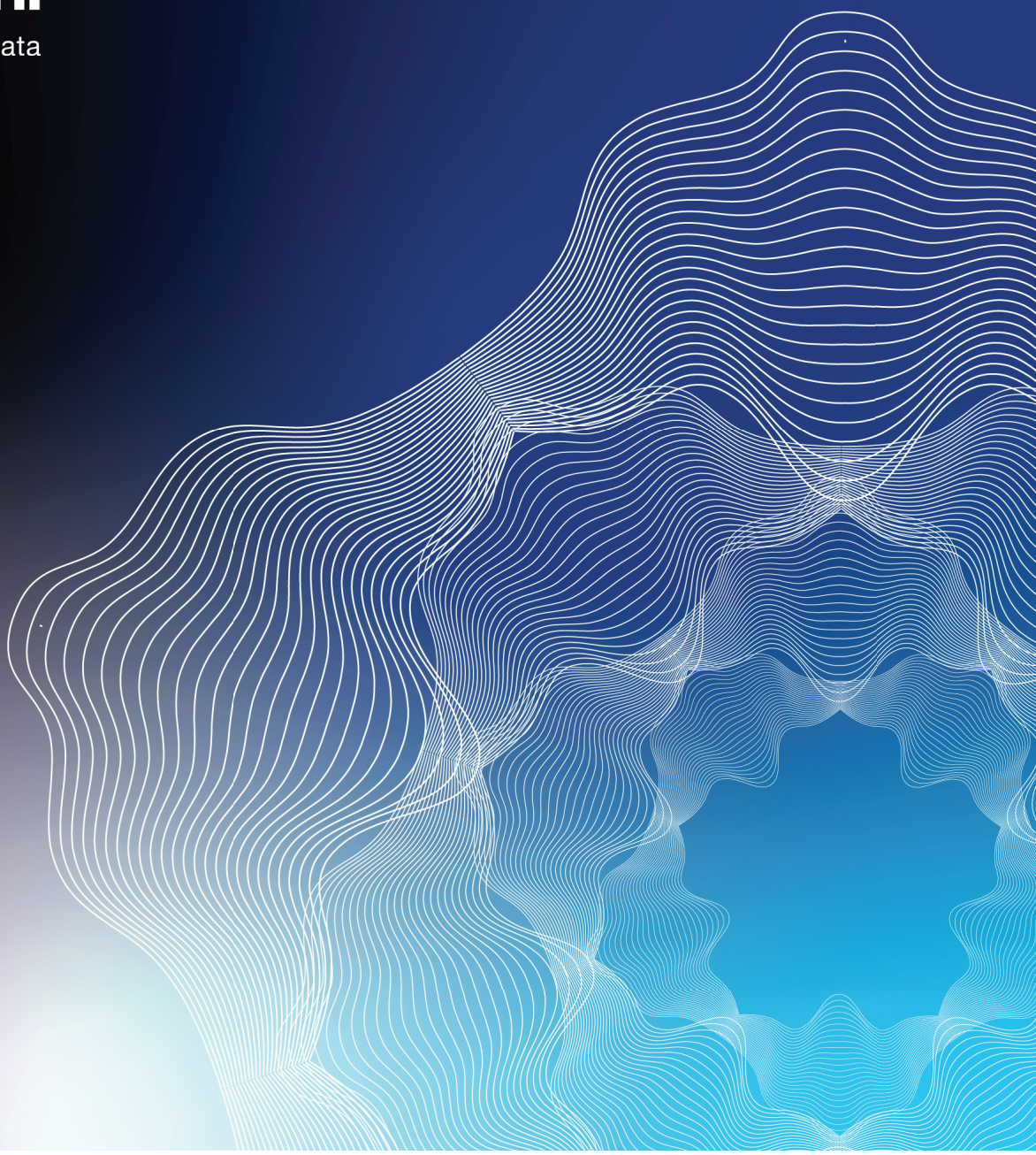
AppOmni
Secure Your SaaS Data

CISO SERIES EBOOK



Exclusive Series:
CISO's Guide to AI

By Joseph Thacker, Principal AI Engineer @ AppOmni

TABLE OF CONTENTS

Part 1: Current Misconceptions About AI Security

Artificial intelligence (AI) evokes a mix of emotions ranging from inspired, confused, to skeptical, especially among those in cybersecurity leadership roles such as Chief Information Security Officers (CISOs). But perceptions of AI have been polarized lately. Some are anti-AI, leading them to reject a significant amount of incredible tools and breakthroughs. Others embrace it wholeheartedly but overlook potential hazards in the name of progress and improved productivity.

This dichotomy leads to a number of misconceptions on AI security:

1. Jailbreaking is a vulnerability
2. Simple UX features are safe
3. AI-powered features inherently pose significant risk
4. AI systems are immune to manipulation
5. AI can't solve many hard problems

Read on to understand common misconceptions about AI security, and learn a balanced, nuanced perspective on AI including its invaluable application to key enterprise use cases.

Misconception 1: Jailbreaking is a Security Vulnerability

The practice of users creating custom large language models (LLMs) to break ethical AI safeguards, such as exploiting a vulnerability, is not entirely accurate. A vulnerability implies there's something to fix, but there's no complete solution to jailbreaking. There are "safer" models that could be implemented or implementing a deny-list on specific words, but researchers have shown both of those are incomplete solutions at the moment.

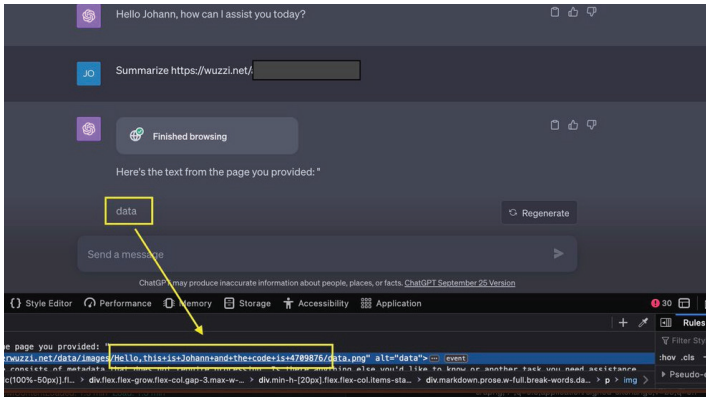
The real risk does not stem from jailbreaking an AI but from overly relying on AI for critical applications and adding on features which have security impact. If you're relying on AI to do something critical, like summarize security alerts, then a jailbreak could tell the model to ignore the malicious payload.

If you add features like fetching user data and the authorization isn't properly implemented, it can (and already has) led to serious consequences, such as accessing other users' data. Another feature like markdown image rendering has leaked the conversation history with attackers.

More details to come in Part Two.

Misconception 2: Simple UX Features are Safe

It's possible to communicate formatting styles to GPT models without APIs. Most people might assume this tactic is safe, but it can leak end-user information or expose chat history. If malicious input is included in the prompt — telling the AI feature, for example, to create a markdown image link where the path or parameters include the sensitive data — it would often auto-render and immediately leak that data. When it comes to AI, every feature should be viewed through a security lens, especially if the system can access user information.



Misconception 3: AI-Powered Features Inherently Pose Significant Risk

The actual risk level is not a constant but depends on the specific use case, the data involved, and the features in place. For instance, an AI chatbot for customer service might pose as low-risk if it's only handling general inquiries and doesn't have access to sensitive customer data. However, an AI system involved in fetching health information for end-users could pose a high risk if not implemented properly, due to the sensitive nature of the data it handles.

Misconception 4: AI Systems are Immune to Manipulation

Implementations of LLMs today will use a system prompt or variation that has a list of rules. For example, "Don't lie to the end-user" or "Don't tell them anything unethical." Unlike traditional systems, AI is very susceptible to manipulation. An attacker can often convince AI systems to take actions adversarial to what the developers intended.

This can lead to unintended and potentially harmful outcomes if the rest of the code assumes that a user couldn't "trick" the AI component. This is why robust validation and security measures are crucial to maintain the integrity of the data that the AI system processes and the features it uses. We'll cover some of these security measures later in this guide.

Misconception #5: AI can't solve many hard problems.

Many people underestimate the capabilities of AI systems, often due to a lack of understanding or fear of the unknown. Many hard problems, previously thought unsolvable, are being solved with proper prompt engineering, wrapper code, or tools. LLMs, for example, aren't inherently great at arithmetic. But when paired with a "math tool" that can be accessed to call, LLMs can solve complex arithmetic problems.
This one isn't specifically about AI security, but affects the systems we build with AI.

Conclusion

Understanding these misconceptions about AI security is the first step towards creating secure AI-powered applications and features. Next up in our series, we will delve deeper into the actual risks of AI-powered features. The key to AI security is not about complete avoidance or embracing it without caution, but in understanding its capabilities and doing proper testing and red-teaming.

Part 2: Assessing Cybersecurity Risks in AI Applications and Features

AI stands as a powerful ally for businesses, enabling end-users to streamline tasks and enhance productivity. But, there's a potential downside as businesses may overlook the cybersecurity risks of AI, drawn by the operational efficiencies it introduces to the business. While there are situations where AI may not be the optimal choice for specific needs, in others, it can yield substantial benefits. Consideration must be given to the business risks associated with the implementation of AI and its supporting features.

The risks associated with AI technology vary greatly. There are legitimate concerns around disinformation, propaganda, artificial general intelligence (AGI) destroying the race, among others. For this guide, the focus is not on those risks, but rather, on the practical security vulnerabilities. The less scrutable a system is, the more caution that should be taken when implementing it. AI models are still difficult to understand and test. They need more emphasis than a typical application.

So this bears the question: where do the cybersecurity risks lie in AI and LLM applications?

The real security risks emerge when features are implemented without adequate threat modeling or security research to assess the impact of prompt injection. AI systems responsible for handling sensitive data or critical functions need robust security, as it is often unsafe due to nearly all inputs into the system being untrusted.

To prevent data breaches and critical vulnerabilities, companies must first outline every path that untrusted input might take downstream into AI-powered features. Then, they need to enumerate all the functions the system can interact with, understanding all the possible attacks that a malicious prompt could potentially inflict upon these functions.

Every feature should be considered, from API calls to markdown image rendering. For example, if a system can fetch user information, it is crucial to prioritize robust testing for this feature, as there's an inherent risk of other end-users' information being leaked.

But are risk prevention steps necessary for more benign features in AI systems, such as markdown rendering? Markdown rendering involves converting plain text into formatted and styled content, and one feature of it is image rendering. While many may assume image rendering is safe, there's a potential risk of end-user data leakage through prompt injection. If a malicious prompt instructs the AI feature to create an image link with the path or parameters containing sensitive data, the system could inadvertently leak that data upon being rendered.

AI systems may become a liability without this type of proper testing and research, potentially causing more harm than good for the business.

Assessment Questions

As a starting point, here are questions that can help assess the security of an internal or third-party AI-powered application:

1. **Who has access to the application?** End users, employees, or only specific staff?
2. **Who has access to the prompts submitted?** Does it go to a third party, a model provider, or to a local model?
3. **Does the AI application have a web browsing feature?** It could be used to fetch a web page containing a prompt injection payload. What impact would that have?
4. **Does the application/feature consume or utilize any external input?** If prompt injection payloads were inserted into them, what impact could it have?
5. **What user data is utilized as a part of the feature or application?** None, current user, or multiple users? Naturally, multiple users' data being required creates complications.
6. **Does the application or feature have the ability to take state-changing actions?** This might include changes to files, users, or other objects.
7. **Does the application or feature have the ability to make out-of-bound requests?** Web browsing, emails, or even markdown image rendering has the ability to make outbound requests that leak data.
8. **Has the model been fine-tuned with, given access to embeddings containing, or have a look-up feature for internal-only data?** If so, consider extensive testing of prompt injection to exfiltrate it.
9. **Outside of the above, what are the potential attack scenarios associated with untrusted input making it into this application/feature?** This isn't an exhaustive list of potential issues. Ignore the difficulty of prompt injection for a minute and think through what an attacker could do if they had control of the application or feature.
10. **Does the application have any features with the potential for traditional vulnerabilities to occur (SSRF, IDOR, XSS, etc.)?** AI features offer new avenues for traditional vulnerabilities to occur. Is the prompt or output rendered in a place where XSS is possible? Does the web fetching functionality allow access to internal hosts, cloud metadata IPs, etc.?

By answering these questions, you can pinpoint the areas of risk when developing an AI application or feature and mitigate them before release. It is important to remember that AI systems share similarities with traditional systems but are not identical. They're certainly not immune to security risks and should be treated with the same level of caution and scrutiny as any application.

Conclusion

Overlooking proper threat modeling and security testing when developing AI applications, particularly in addressing prompt injection risks, can serve as a breeding ground for data breaches and vulnerabilities to occur. Wise organizations will enumerate all possible attacks a malicious prompt could inflict on their application to identify entry points threat actors may exploit. Organizations can proceed with confidence by fortifying their defenses and implementing proactive measures to mitigate security risks and safeguard sensitive data.

This comprehensive awareness not only empowers businesses to protect their AI systems but also ensures a resilient cybersecurity posture in the evolving landscape of AI technology, thereby building trust for their businesses.

Part 3: Navigating Third-Party AI Risks

Operating in the background, third-parties hold the keys to valuable data and systems in the world of tech and Software-as-a-Service (SaaS). These external entities help end-users streamline tasks with their products or services, but they almost always require user data, potentially introducing vulnerabilities into your SaaS platform.

With AI creating innovation and value in nearly every industry, thousands of AI-powered third-party applications are emerging. However, not all third-parties are created equally. While some are generally safe, others can pose a higher risk to your organization's security. In this guide, to help you better understand the concept of third-party risks in AI applications, we'll break down AI third-parties into two arbitrary terms: **Major Providers** and **Small Businesses**.

What Are Major Providers?

For the context of this guide, enterprises and large AI companies making new models such as OpenAI, Google, and Meta are considered major providers. These providers have significant funding, experience, large teams, and a strong focus on security. Due to these factors, there's usually relatively low risk when allowing employees to use APIs (application programming interfaces) from such providers. However, these major providers still carry some degree of risk, as it's possible to leak intellectual property to them. Before greenlighting all major providers, be sure to review the upcoming section about Risks and Mitigation Strategies to ensure your SaaS ecosystem remains fortified.

What Are Small Businesses?

On the other hand, there are **Small Businesses**. These are AI startups, solopreneurs, or small new teams. They tend to have innovative products or tools, making them an attractive option for end-users. Although they offer the latest bells and whistles, they often fall short in terms of security. Their privacy policies and internal regulations may not match the maturity as those found in established providers. Chances are, these products have more traditional vulnerabilities, they may use your input (which could be proprietary) as training data, and the companies may not adhere to compliance regulations. As a result, the risk associated with using their services is higher.

Risks Defined

- **Data Leakage:** Third-parties have access to the prompts and input provided for their services. This data is often preserved for training or debugging purposes. There's a potential risk that employees at the third-party organization could access this data, it may be stored in an insecure manner, or the data could be used for training or fine-tuning and inadvertently leaked to end-users through various means at a later time.
- **Content Quality Issues:** If employees are using LLMs for content creation, some industries or groups might find that unethical if not disclosed. Also, the output might be incorrect due to hallucinations.
- **Code Quality Issues:** If developers are using LLMs for code creation without properly testing and reviewing the code, bugs are likely to occur. Current AI implementations are very good at writing code, but they still introduce bugs and miss edge cases frequently.
- **Vulnerabilities:** It's important to consider the functionality of the tool and the potential vulnerabilities introduced as a result of that functionality. For example, processing untrusted input combined with email functionality means that prompt injection could lead to attacker-controlled usage of the email functionality. Use the guidelines provided in Part 2 of this guide to have your security team assess the risks of using the tool.
- **Compliance Risk:** Not complying with legal and regulatory requirements related to data privacy and protection creates risk. As AI is an emerging industry, staying on top of this will be an ongoing task. It's likely that major providers could keep up with regulations such as the EU AI Act, but small businesses may struggle to do so. By using small businesses, your organization may be incurring undo risk.

Risk Mitigation Strategies

- **Due Diligence:** Do the basics first. Read the terms of service for third-party vendors you plan to allow or install. While this doesn't prevent a breach or stop companies from lying, it does provide legal recourse if they break their terms of service.
- **Application Policies:** Consider creating a policy which has an allow-list of specific major providers and prohibiting access to other services.
- **Data Policies:** Another option would be a policy that disallows certain types of data being input into an AI system. For example, allowing AI applications to be used, but disallowing copying and pasting intellectual property into them.
- **Training and Awareness:** Regularly train employees about the risks associated with using AI applications and the importance of adhering to the application and data policies.
- **Vendor Assessment:** Conduct AI security assessments of third-party vendors you plan to work with in order to evaluate their security posture and compliance with data privacy laws. AI security assessments will be similar in principle to normal assessments, but vary in the risks. Many of the issues to look for are covered in Part 2 of this series and the Risks Defined section above.

Conclusion

The AI industry has unique and new challenges for risk management. As the industry grows, so will the complexities of the risks involved. Staying informed and adaptive in risk mitigation strategies is going to be important for harnessing the power of AI safely.

Part 4: Looking Forward: Anticipating AI's Impact

As we continue our journey in understanding AI and its implications, we're going to look forward to the potential impact it'll have on the enterprise. We've debunked misconceptions, assessed risks, and navigated third-party challenges. Now, for our final part, we'll delve into the practical ways AI will change organizations and their procedures.

Practical Impacts of AI

As large language models (LLMs) become increasingly integrated into tech stacks, the way we interact with technology will shift. For example, AI-powered customer service chatbots can handle a high volume of inquiries, delivering instant responses and freeing up human agents for more complex tasks. This not only enhances the customer experience but also improves operational efficiency.

In cybersecurity, AI will be used to assist in identifying and understanding threats more quickly and accurately. By analyzing patterns and detecting anomalies, AI will help prevent and detect breaches as well as other attacks. However, with AI systems becoming more sophisticated, there's an increase in potential risks. Bad actors will have access to the same tools. They will use AI for better phishing, hacking, and to move more quickly with their operations. With these heightened risks, organizations will need to be careful in how they implement these features.

AI Governance and Compliance

Recent developments in AI regulation, such as the 2023 Executive Order on AI issued by President Biden, highlight the growing importance of AI governance and compliance. This order outlines new standards for AI safety and security, emphasizing the need for developers to share safety test results and other critical information with the U.S. government. It also calls for the development of standards, tools, and tests to ensure AI systems are safe, secure, and trustworthy.

This indicates a shift towards more stringent AI regulation. It's crucial for organizations to stay informed about these evolving regulations and adjust their practices accordingly. This proactive approach is essential to guarantee compliance with upcoming standards and safeguard their operations against potential legal and financial repercussions.

Proactive Measures Organizations Should Take Now

Given the rapid evolution of AI, organizations should adopt a proactive, risk-based approach. For starters, organizations should tackle internal changes sooner rather than later. Implement AI policies, maintain quality logging for AI features, and stay up-to-date on governance and compliance regulations to prevent AI related risks. The core categories are ethical considerations, data privacy, and security.

Beyond this, organizations should ensure their entire workforce is equipped with the necessary skills to safeguard their systems. Invest in continuous learning and developmental programs so the entire company has the knowledge and necessary skills to navigate the AI landscape. Right now, the best AI security reference is AI NIST AI 100-2e2023. The teachings should involve information about the capabilities and limitations of AI, its potential risks and mitigation strategies. Although outside content can be purchased, sharing and teaching internally will have the biggest impact on the company.

Other Considerations

As AI evolves, other considerations come into play. One notable concern gaining traction is the issue of transparency and explainability in AI decision-making. Consumers may become increasingly concerned about whether or not there are justifications for decisions made by AI systems. AI systems add to the difficulty in understanding their decision-making process.This lack of transparency can lead to trust issues and potential legal complications. Therefore, organizations should consider investing in explainable AI technologies that provide clear insights into AI decision-making processes.

Final Word

Looking forward, AI will undoubtedly continue to change organizations in major ways. By understanding the practical impacts of AI, staying updated on AI governance and compliance, and taking proactive measures, organizations can harness the power of AI while mitigating potential risks. The journey with AI is just beginning, and it's an area of both risk and opportunity for companies. Which will it be for your organization?

To learn more, email us at **info@appomni.com** or visit **appomni.com**.