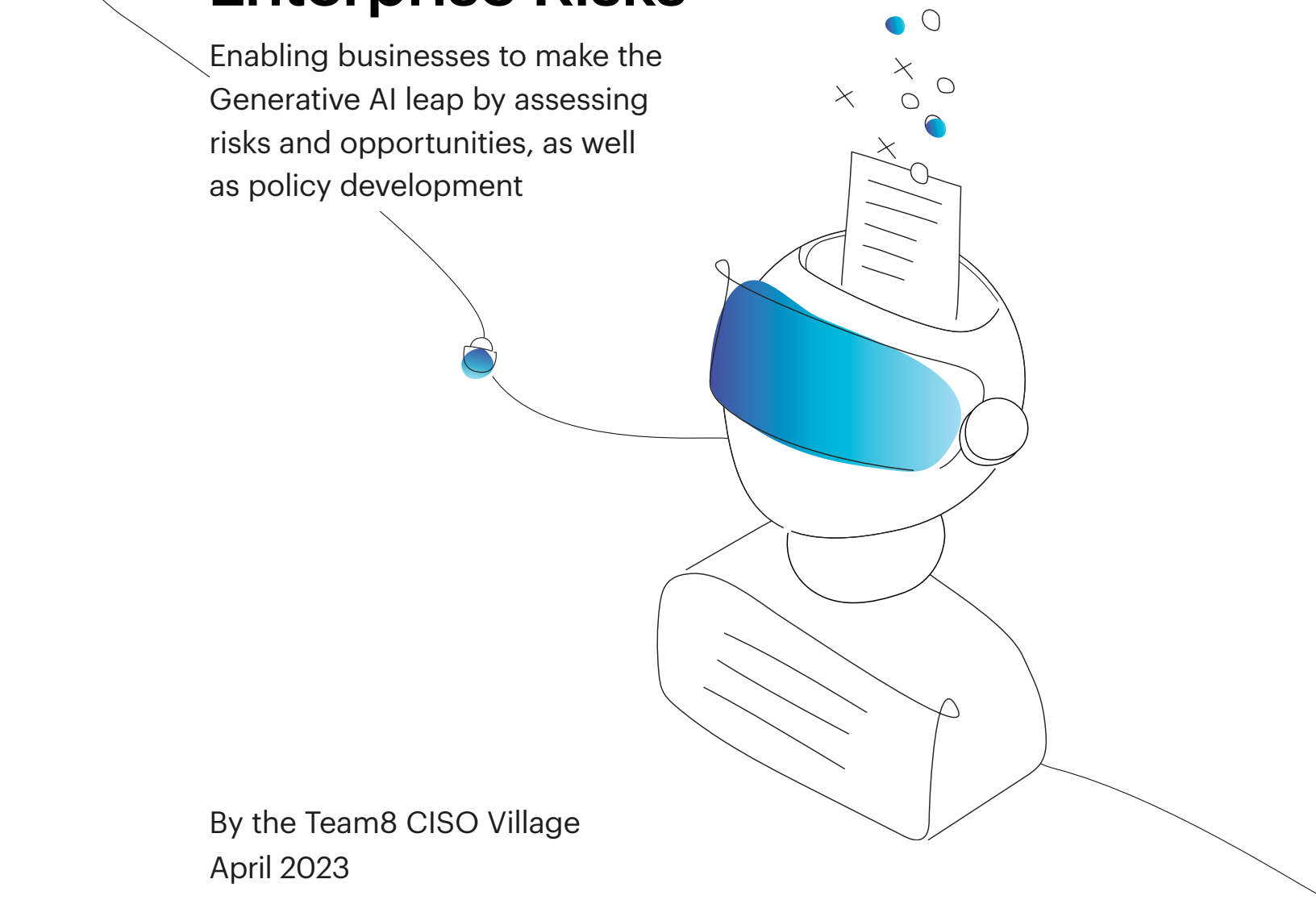
TEAM8™

A CISO'S GUIDE

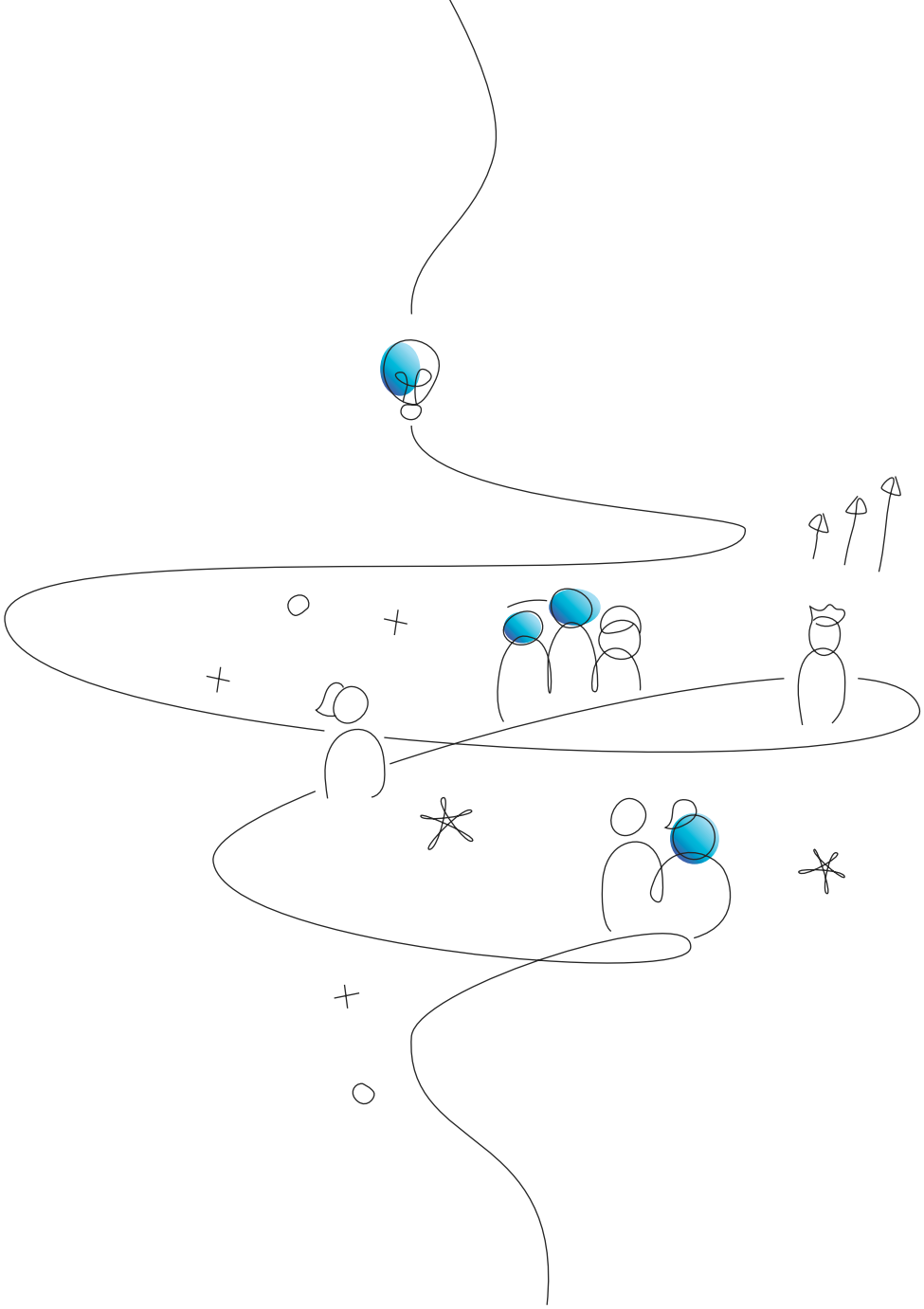
Generative AI and ChatGPT Enterprise Risks

Enabling businesses to make the Generative AI leap by assessing risks and opportunities, as well as policy development



By the Team8 CISO Village

April 2023



The Team8 CISO Village is a community of CISOs from the world's leading enterprises. The primary focus of the Village is to facilitate collaboration among the world's most prominent companies with the goal of sharing information and ideas, conducting intimate discussions on industry and technology trends and needs, and generating value and business opportunities for all parties.
By helping Team8 to identify real pain points and understand the requirements of large organizations, members of the Village are first in line to leverage solutions that are purpose-built by Team8's portfolio companies to support their needs.

To contact the Team8 CISO Village, please email **cisovillage@team8.vc**

WRITTEN BY

Gadi Evron

CISO-in-Residence
Team8

Bobi Gilburd

Chief Innovation Officer
Team8

CONTRIBUTORS

Amit Ashkenazi
Former Legal Advisor of the Israel National Cyber Directorate, and before that Head of the Legal Department at Israel's Privacy Protection Authority

Cassio Goldschmidt
Chief Information Security Officer

David B. Cross
Security and Engineering Executive

Gal Tal-Hochberg
CTO, Team8

Jonathan Braverman
Security and Privacy Executive

Maxim (Max) Kovalsky
Security and Privacy Executive

Richard Barretto
Chief Information Security Officer, Progress

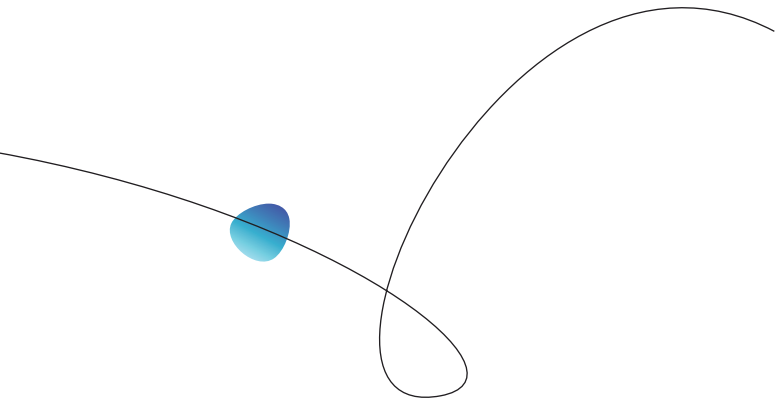
Sounil Yu
Chief Information Security Officer, JupiterOne

Many Team8 CISO Village members, and others from the wider community, assisted in the writing, reviewing, and editing of this document. These are the ones who could share their names publicly: **Aaron Dubin, Adam Shostack, Alyssa Miller, Amir Zilberstein, Ann Johnson, Aryeh Goretsky, Avner Langut, Avi Ben-Menahem, Brian Barrios, Chenxi Wang, Dave Ruedger, Dikla Saad Ramot, Doron Shikmoni, Gidi Farkash, Imri Goldberg, Jeffrey DiMuro, Larry Seltzer, Liran Grinberg, ADM Michael S. Rogers USN (ret), Michal Kamensky, Nadav Zafrir, Nate Lee, Oren Gur, Reet Kaur, Roy Heldshtein, Sara Lazarus, Ric Longenecker, Susanne Senoff, Tomer Gershoni**

Executive Summary and Key Takeaways

Executives, and among them CISOs, are finding themselves behind the technology adoption curve, while employees and business units are rapidly adopting Generative AI (GenAI) technologies.

- Key questions CISOs are asking: Who is using the technology in my organization, and for what purpose? How can I protect enterprise information (data) when employees are interacting with GenAI? How can I manage the security risks of the underlying technology? How do I balance the security tradeoffs with the value the technology offers?
- Risks associated with GenAI are manageable by developing a bespoke policy for the organization, in collaboration with relevant stakeholders.
- While many of the risks inherent to GenAI exist in any cloud or AI/ML-based technologies, new policies that assume wide adoption of GenAI, including by non-technical personnel, are required.
- The most acute risks associated with GenAI at the enterprise stem from four underlying issues:
  - The effects of GenAI on internal operations and processes.
  - The necessity to trust third-party security.
  - Legal and regulatory risks resulting from adoption.
  - Data leak risks, while they do exist, have been unnecessarily hyped in the media.
- These risks can be managed via the application of following strategies:
  - Identify relevant risks and their impacts to your organization;
  - Set organizational policies on how and who can use these tools in a manner that mitigates the above risks to acceptable levels;
  - Choose appropriate GenAI providers based on their security and policy customizability afforded to customers, e.g. opt-out and data retention, and;
  - Examine potential on-premise alternatives under enterprise control.



How to Read This Document

This document aims to provide CISOs, security practitioners, and others with actionable tools to understand and communicate the security implications of GenAI for organizations, and associated legal and compliance risks they should discuss with other enterprise stakeholders.

It supports decision making regarding the implementation, integration, and use of GenAI, in order to write organizational policies that allow the safe and secure use of this novel technology.

It can be read in written order, to gain an understanding of GenAI threats and policy considerations, as a framework to discuss that understanding with organizational stakeholders, and finally apply them in organizational GenAI security policies.

Alternatively, to achieve specific objectives, the sections can be read out of order in the way that is most effective for the reader to achieve their objectives.

To understand generative AI security implications and build threat matrices: Read The Changing Threat Landscape, New Enterprise Risks Introduced by GenAI, and Engaging engineering teams and Threat Modeling.

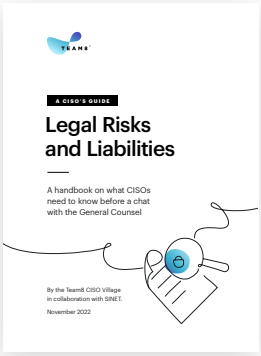
To communicate with organizational stakeholders (board, management, practitioners and product organizations): Read Enterprise GenAI and ChatGPT Policy Considerations and Engaging engineering teams and Threat Modeling.

To write a Generative AI / ChatGPT security policy and make risk decisions: Read Enterprise GenAI and ChatGPT Policy Considerations, Making risk decisions, and Sample GenAI and ChatGPT Policy Template.

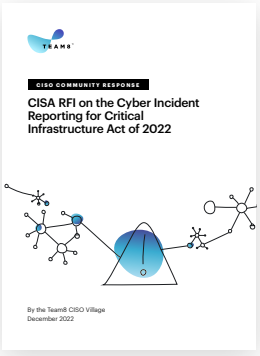
Let's Talk

You're invited to reach out and discuss this and other topics with Team8. Feedback, open discussion, and briefing requests are welcome.
You are also invited to work with us on future CISO Village collaborative projects.

Some of our recent publications:



"A CISO's Guide: Legal Risks and Liabilities"



"CISO Community Response: CISA RFI on the Cyber Incident Reporting for Critical Infrastructure Act of 2022"

Future collaborative projects being considered on this topic include:

- GenAI security controls that can be introduced by enterprises
- Developing GenAI best practices and frameworks for enterprise (security/safety/risk mitigation).

To contact the Team8 CISO Village, please email **cisovillage@team8.vc**

CONTENTS

Introduction

Many enterprises and their employees are already using ChatGPT, Bard, PaLM, Copilot, and other Generative AI (GenAI) or Large Language Models (LLMs), and CISOs are now working with their teams to identify and evaluate relevant enterprise risks. We will refer to the various technologies and models as "GenAI" in this text for simplicity.

GenAI, and ChatGPT specifically, have been gaining popularity among enterprises as powerful tools for streamlining communication, writing code, generating images, audio, and video, enhancing customer service, composing documents and producing initial research, and improving any number of other day-to-day activities.

However, as with any technology, the use of GenAI also poses a range of security risks,threats, and impacts that organizations must consider carefully.

GenAI has been one of the top concerns among security executives over the first few months of 2023, and has been prioritized by the members of our Team8 CISO Village community for the production of a joint document on this topic.

Key questions CISOs are asking: Who is using the technology in my organization, and for what purpose? How can I protect enterprise information (data) when employees are interacting with GenAI? How can I manage the security risks of the underlying technology? How do I balance the security tradeoffs with the value the technology offers?

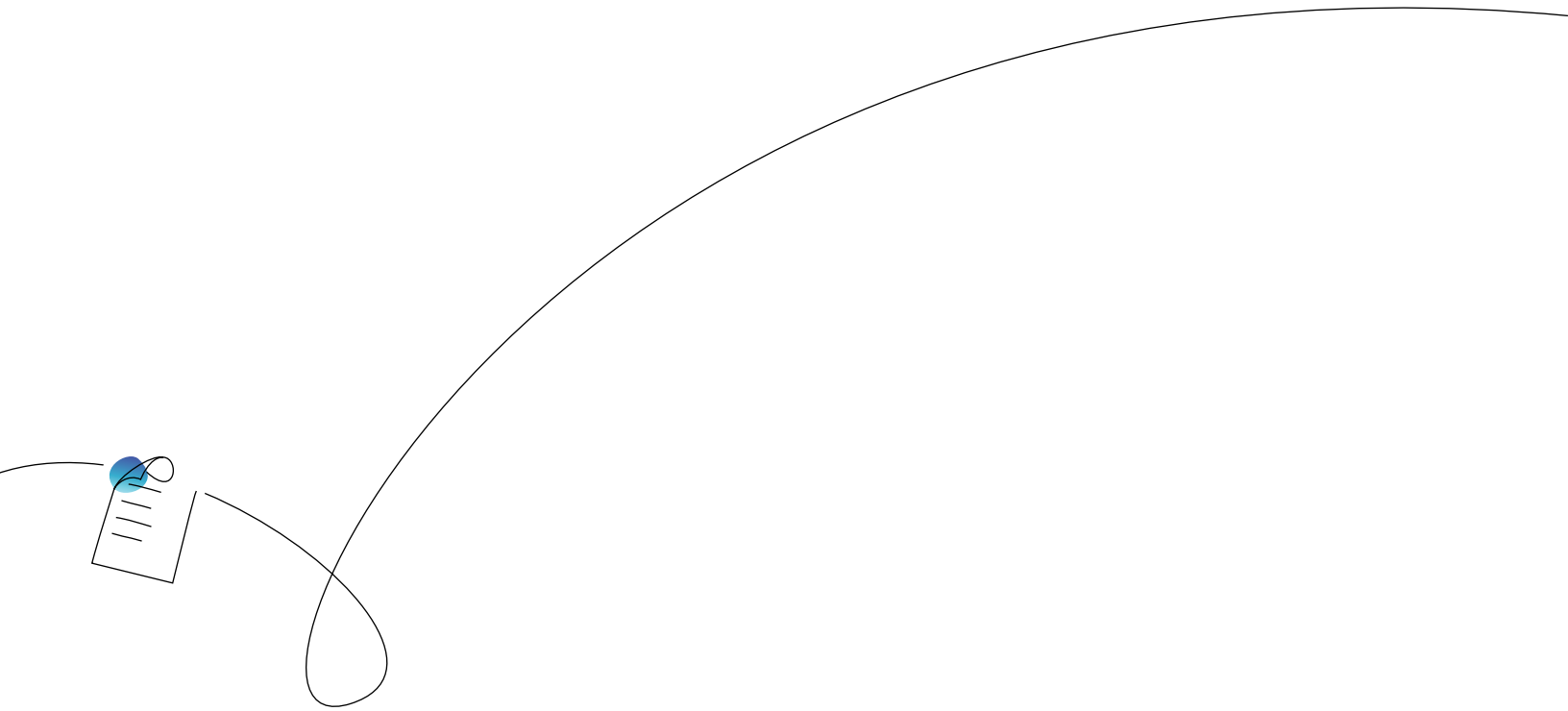
This document provides information on risks and suggested best practices that security teams and CISOs can leverage within their own organizations, and serves as a call to action for the community to evangelize and further engage on the topic.

- **We will explore the potential risks and threats** associated with using GenAI in an enterprise setting, focusing on technical aspects, and some of the legal and regulatory risks that stem from these in turn . We will further discuss threat modeling, engaging engineering teams, and on-premise or open source alternatives.
- **We will then provide actionable recommendations** for how enterprises can take a comprehensive approach that includes developing an organizational policy and action plans, along with a sample policy, so that they can ensure that they are using GenAI in a secure and safe way.

Many CISOs have unfortunately found themselves behind the GenAI adoption curve and risk being seen as business blockers rather than business enablers. Therefore CISOs may feel pressure to allow GenAI broadly, but doing so indiscriminately could create unreasonable risk.

Aside from bringing the enterprise up to standard on GenAI usage, developing written policies, and implementing security controls, we have a unique opportunity as security leaders to promote and enable the business through safe and secure technology innovation.

Given the productivity boost that GenAI provides for all enterprises, organizations need a measured business-enabling alternative to rejecting the use of the technology altogether. Organizations that have initially implemented such restrictions, have more recently been reconsidering their positions and lifting wholesale bans.



[1] Legal and regulatory content is meant to flag issues in preparation for discussions with other enterprise stakeholders, and is not intended to replace legal advice where necessary.

The evolving CISO role

The CISO role needs to evolve to adjust to new enterprise risks, especially considering the development of regulation in this area.

In the context of the developing EU AI act, some have described CISOs as "Ambassadors of Trust." The CISO organization's potential is now recognized as a catalyst for developing a wider approach to risk that includes additional emphasis on data collection, usage, governance, and infrastructures.

These are areas of responsibility where CISOs, who often do not hold the full—or any—authority, do have technical risk literacy, and can provide substantive support to the enterprise mission. They may also be asked to take some level of responsibility.

As senior executives with an understanding of both the business and technology, CISOs are well positioned to help organizations navigate the risks and opportunities engendered by this new technology

Facing a historic opportunity

The CISO community faces a historic opportunity to affect the security and privacy of individuals worldwide, beyond our organizations.

When new technologies are introduced to the world, security and privacy are often an afterthought, whether due to business incentives, budgetary and resource concerns, or simply the sheer force of innovation.

Indeed, this has happened before with social media, the cloud, smartphones, and many other technologies where security and privacy have been introduced very late in the adoption cycle, while uncontrolled usage skyrocketed.

As a collective, the CISO community has the power to shape the best practices and influence the future development of these tools—today, with the safety and privacy of individuals, as well as our organizations', in mind.

Lastly, at the Team8 CISO Village, we are planning future collaborative projects on this topic.

The Changing Threat Landscape

Everyone, including GenAI developers such as OpenAI, are still learning about the full potential of the technology. While some threats seem apparent and are similar to those we have seen before the full roster of risks is still unknown.

GenAI has rapidly introduced novel security risks due to its prevalence, ease of use, high value proposition, and immense potential business enablement value.

As an example of its rapid adoption curve, ChatGPT reached 100 million users in two months, orders of magnitude faster than any previous technology.

Compared to traditional AI/ML technologies, GenAI is much easier to use:

- **Intuitive interaction and novel content production model** – A new approach in AI combined with an interactive chat system, delivering "polished" results. End users have the ability to live-edit the results in a chat interface to enhance future accuracy much more easily than in earlier platforms.
- **Accessible to all** – Many of the GenAI technologies are free or low-cost, open to the public, and accessible to anyone with an Internet connection.
- **Ease of Use** – Many of the GenAI technologies are designed to be easy for people of all positions and roles, using natural language as though conversing with another person.
- **Speed and Agility** – The system can produce information, source code, and data faster than previously possible through manual searching, queries, and indexing. It can synthesize millions of pages of information into a single paragraph.
- **Integration with Third Party Applications** – Many current everyday applications, such as the Microsoft Office 365 suite of tools and browser plugins, point to a reality that GenAI will become ubiquitous in our everyday lives.

New Enterprise Risks Introduced by GenAI

Understanding the range of potential GenAI risks, threats, and impacts in an enterprise setting has become a priority, and must be carefully considered. These can be broken down to technological or process extensions of existing risks, legal and regulatory risks, and some risks that are completely new.

To provide an overview of these risks, we created the reference table below, arranged and prioritized by risk level, current as of April 2023 (Disclaimer: Every organization is unique, these are offered as guidelines).

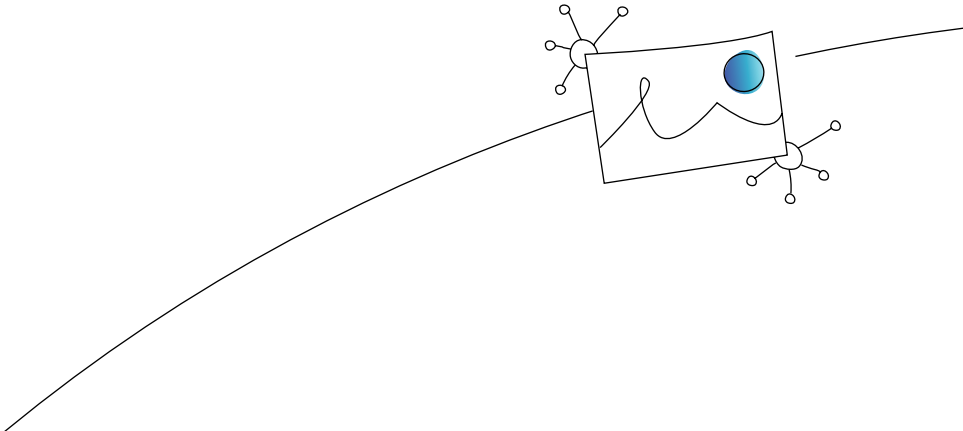
Enterprises can take proactive measures to minimize the potential negative impact of GenAI usage and ensure that they are leveraging this powerful tool in a secure, compliant, and responsible manner.

For example, organizations **could** decide that while data is being sent to a third-party GenAI SaaS platform, opting out of the user prompt information being used to train future models, and accepting the data retention policy of 30 days (as is in the case of OpenAI, for both), is secure enough for their needs and meets their risk appetite.

Others may decide to declare a Risk Exception and accept the risk, or decide to explore an on-premise alternative.

Due to hype in the media, the authors and contributors want to specifically address data exposure risks associated with GenAI, and ChatGPT specifically, on making user input and proprietary data available to unintended audiences, such as competitors. **As of this writing, Large Language Models (LLMs) cannot update themselves in real-time and therefore cannot return one's inputs to another's response, effectively debunking this concern.**

However, this is not necessarily true for the training of future versions of these models. GenAI technologies may use submitted content to improve their algorithms and models in the future. We touch on this risk and its likelihood, as well as specifically OpenAI's position on leveraging user input for model training (opt-out from such usage, data retention limitations, challenges with biased models on user input, etc.) in the table below, and later in this document.



Risk

Data Privacy and Confidentiality

Estimated Risk Level[2]

High

Threat to: Non-public enterprise and private data

Enterprise use of GenAI may result in access and processing of sensitive information, intellectual property, source code, trade secrets, and other data, through direct user input or the API, including customer or private information and confidential information. This has already been reported to be an issue, here.

LEGAL CONSIDERATION consult an attorney

> **Sending confidential and private data outside of the organization's own servers, much the same as with the cloud, could trigger legal and compliance exposure,** as well as risks of information exposure. Such exposure can result from contractual (e.g. with customers) or regulatory obligations (e.g. CCPA, GDPR, HIPAA). The discussion below highlights these types of exposure to support informed risk management and mitigation measures.

- While data sent to GenAI technologies such as ChatGPT has been effectively entrusted to a third-party SaaS, i.e. OpenAI (see also the "Third-Party Risk and Data Security" section of this table), **it is not currently incorporated into the LLM in real-time, and thus won't be seen by other users.** GenAI platforms may choose to use user input to train future models, but that doesn't seem to be the case now. See more on this and why we make this differentiation here.
- Specifically in the case of OpenAI, according to the company documentation, content submitted to their API is not retained for more than 30 days, and it is opt-out by default, while ChatGPT is opt-in by default and opt-out with fee-based accounts. Nevertheless, information submitted is always subject to storage and processing risk. But, this may affect the risk appetite for different organizations.

[2] These estimated risk levels are offered as guidelines based on active threats as seen in open source intelligence, and consensus among the CISOs who wrote and reviewed this document, as of April 2023. We recommend our readers to conduct their own estimates and risk analysis to account for their unique enterprise risks and the specific technologies used. Doubly so when attorneys need to be consulted

Risk

Enterprise, SaaS, and Third-party Security

Estimated Risk Level[2]

High

Threat to: Non-public enterprise data, third- and fourth-party software

Due to GenAI's wide adoption and proliferation of integrations in third-party applications, there are concerns among CISOs that data would be shared with third parties at a much higher frequency than in the past, and potentially following much less predictable patterns.

- With supply chain, increased outsourcing risks can fall into three broad categories: Reliance on third-party security while processing enterprise information, reliance on third-party quality assurance when producing content and code (see Insecure Code Generation below in this table), and integration with GenAI technologies.
- If the GenAI platform's own systems and infrastructure are not secure, potential data breaches, such as the recent one with OpenAI (creators of ChatGPT), may occur and lead to the exposure of sensitive information such as customer data, financial information, and proprietary business information. Further, these platforms are relatively new, and their security experience and posture may increase concerns in security due diligence.
- Third-party applications: GenAI and specifically ChatGPT are quickly being embedded into many third-party platforms like the Microsoft Azure OpenAPI, as well as applications, from the Microsoft Office 365 suite of tools to browser plugins. Thus, the risk is not focused on one service only, and in the case of Office 365, is stated to be a closed-off instance of ChatGPT.
- Considering the limited number of available GenAI platforms, and how they are effectively becoming an infrastructure used for multiple purposes by many organizations, they represent a high-value target for threat actors. Thus, concentration is introduced and overall risk is increased.

Risk

AI Behavioral Vulnerabilities (e.g. Prompt Injection)

Estimated Risk Level[2]

High

Threat to: Non-public enterprise data, model operator

Actors may use models or cause models to be used in ways which will expose confidential information about the model or cause the model to take actions which are against its design objectives.

- For example, using maliciously crafted inputs, attackers can bypass expected AI behavior or make AI systems perform unexpected jobs. This is sometimes known as "jailbreaking" and might be possible to perform in GenAI systems to adversely impact other organizations and stakeholders to encounter and receive maliciously crafted results without their knowledge.
- On the user side, for example, third party applications leveraging a GenAI API, if compromised, could potentially provide access to email and the web browser, and allow an attacker to take actions on behalf of a user
- One common attack currently seen in the wild is where a customer support chatbot is targeted with injection attacks, and unauthorized access to enterprise systems could potentially be achieved by an attacker.

Risk

Legal and Regulatory[3]

Estimated Risk Level[2]

High

Threat to: Regulatory compliance

- **Regulatory Consideration (consult an attorney):** Using GenAI as part of enterprise processing of PII must be compliant with data privacy regulations such as GDPR (Europe), PIPEDA (Canada), or CCPA (California). In fact, Italy's data protection agency has now temporarily banned the use of ChatGPT specifically (not affecting other GenAI technologies, nor private instances of ChatGPT such as with Microsoft Office 365), due to similar concerns, and Germany is now considering the matter.
- **Regulatory Consideration (consult an attorney):** When GenAI is used as part of a regulated use case in consumer-facing communications, whether for direct consumer interactions or to produce consumer-facing materials (such as consumer information notices), regulatory or private law may include requirements, and create liability.

 **LEGAL CONSIDERATION** consult an attorney

> The use of ChatGPT in conjunction with a chatbot service, could create legal or regulatory exposure. For example, a company could potentially face regulatory action or even a lawsuit for failing to disclose the fact consumers are interacting with a chatbot service to customers.

Risk

Threat Actor Evolution

Estimated Risk Level[2]

Medium

Threat to: Enterprise readiness, third parties

Threat actors use GenAI for malicious purposes, increasing the frequency of their attacks and the complexity level some are currently capable of, e.g. phishing attacks, fraud, social engineering, and other possible malicious use such as with writing malware, although that remains a limited capability at this stage.

This would require a re-assessment of security awareness training and other social engineering controls, as well as mitigating controls where these previous ones are not up to par (or have already proven ineffective previously).

[3] The legal and regulatory landscape is as of yet in the early stages of discovery, and the topics mentioned are far from decided, and our writing can not be taken as legal advice. You should conduct your own research and consult with your own attorney

Risk

Copyright and Ownership

Estimated Risk Level[2]

Medium

Threat to: Organization's legal exposure

GenAI models are trained on diverse data, which may include an unknown quantity of copyrighted and proprietary material, raising ownership and licensing questions between the enterprise, and other parties whose information was used to train the model (as explored in this Forbes article).

LEGAL CONSIDERATION consult an attorney

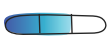
- Some GenAI models have been reported to have used content created by others, which they were trained on originally (in the referenced case, code, although it could be text, an image, etc.) instead of content uniquely generated by the model, raising risks of intellectual property infringement or even plagiarism. In addition, the same content could potentially be generated for multiple parties.
- Using the output of GenAI could risk potential claims of copyright infringement, due to the training of some GenAI models on copyrighted content, without sufficient permission from dataset owners.
- The US Copyright Office has published guidance refusing copyright protection for works produced by GenAI. That could potentially be interpreted to permit all software output from GenAI to be copied and used by anyone freely.
- GenAI may return snippets or fragments of content or code that could include proprietary content or content covered by Open Source licensing models, e.g. GPL 3. Such content could create legal exposure and in some cases require organizations to release the code produced using these tools as Open Source. Any detected violations, even if unintended, could create legal exposure. Consulting with an attorney is highly recommended.
- Currently, there is a lack of definitive case law in this area to provide clear guidance for legal policies. Policies should be developed according to current intellectual property principles.

Risk

Insecure Code Generation

Estimated Risk Level[2]

Medium



Threat to: Software development projects and developers

Code generated by GenAI could potentially be used and deployed without a proper security audit or code review to find vulnerable or malicious components. Further, this can cause widespread deployment of vulnerable code as it is used in other organization systems and as "ground truth" in future model learning.

Risk

Bias and Discrimination

Estimated Risk Level[2]

Medium

Threat to: Organization's brand and reputation

LEGAL CONSIDERATION consult an attorney

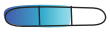
Training on biased data may lead to illegal discrimination, to potential damage to reputation, andpossible legal repercussions for the enterprise. When displaying such output GenAI may not be aware of potentially defamatory, discriminatory, or illegal content. In addition, there is a clear risk of tainting, abuse, and contamination of data models by malicious users (recommended background reading), see more under Data abuse or Taint higher in this table. In use cases when such risks arise, an attorney should be consulted.

Risk

Trust and Reputation

Estimated Risk Level[2]

Medium



Threat to: Organization's brand and reputation

There are substantial reputational risks stemming from GenAI producing erroneous, harmful, biased, or embarrassing output, as well as resulting safety considerations such as in doxxing, or hate speech.

- The current generation of GenAI models have been observed to output incorrect, inaccurate, wrong and misleading information.
- Incorporation of AI outputs in organizational work products, communication or research without vetting for accuracy may lead to publication of incorrect statements and information.

LEGAL CONSIDERATION consult an attorney

There is a potential risk that publishing or sharing content that is inaccurate, inciting, or defamatory, even when produced without malicious intention, may trigger legal liability, such as libel.

Risk

Software Security Vulnerabilities

Estimated Risk Level[2]

Low

Threat to: Non-public enterprise data and system integritys

- Internal and third-party applications using GenAI must be up-to-date and protected by proper controls against classic software vulnerabilities and ways they interact with evolving AI vulnerabilities.
- Any GenAI system is exposed to the same risks as any traditional software systems. Additionally, software vulnerabilities may interact with AI vulnerabilities to create additional risk.
- For example, a vulnerability in a front-end may allow prompt injection on a back-end model. Alternatively in situations where model output is used programmatically, attackers may try to affect the model output to attack downstream systems (e.g., causing a model to output text which causes an SQL injection when added to an SQL query).

Risk

Availability, Performance, and Costs

Estimated Risk Level[2]

Low

Threat to: Enterprise systems' resilience

This risk isn't necessarily novel, as any new service or software being considered could pose a potential operational risk. The use of GenAI may introduce new availability and infrastructure risks, such as system downtime, performance, or user errors, needs to be included in threat modeling and architecture planning, and requires robust backup and disaster recovery procedures to minimize impact, especially where has been integrated into critical enterprise processes.

Operating a LLM can be tremendously expensive. OpenAI specifically has stated that each response from ChatGPT costs "single digit cents. Across potentially large populations of enterprise and end customer users generating a vast number of daily queries this can add up to significant costs".

Risk

AI Ethics and Regulatory Developments

Estimated Risk Level[2]

Low

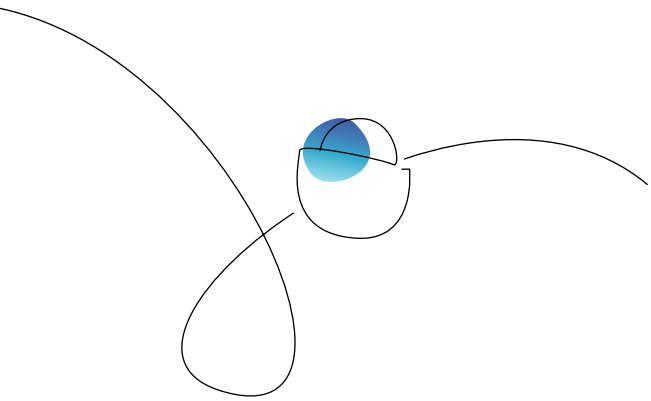
Threat to: Regulatory compliance

The use of AI is increasingly drawing attention in the context of dedicated AI ethical principles, which include safety, security, fairness, transparency, explainability, and general accountability requirements.

Although not yet binding, some of these principles are already being integrated within existing legal and regulatory frameworks. Thus it is a best practice to conduct an initial screening on whether there is a risk that these principles are impacted and how and whether specific mitigation measures should be considered. CISOs have an opportunity to influence the regulations and educate the regulators in advance.

Similarly, Environmental, Social, and Governance (ESG) risks should be considered. For example:

- Research Process Harms – what are the labor practices around data labeling? Is the model trained by "data labeling sweatshops"?
- Environmental impact - What is the environmental impact due to significant compute / energy consumption required to train LLMs?



Enterprise GenAI and ChatGPT Policy Considerations

This section discusses tools for effective enterprise policy development on GenAI broadly, and ChatGPT specifically. (See also the appendix with a list of requirements the enterprise and users may put forward, to inform the policy development process.)

GenAI risks can be managed via the application of following strategies:

- Identify relevant risks and their impacts to your organization;
- Set organizational policies on how and who can use these tools in a manner that mitigates the above risks to acceptable levels;
- Choose appropriate GenAI providers based on their security and policy customizability afforded to customers, e.g. opt-out and data retention; and
- Examine potential on-premise alternatives.

Also included as an appendix with a sample policy for GenAI usage in the enterprise, and another with a list of requirements the enterprise and users may put forward to inform the policy development process.

Do we need a GenAI or ChatGPT-specific policy?

Yes! The authors of this document are in agreement on the need for new GenAI-specific enterprise policies, but only where existing policies do not suffice.

While there is not yet a full consensus on the topic, recommendations below can serve both purposes. Enterprises across all industries are proceeding with dedicated policies, and it seems that approach is becoming an industry best practice.

The scope of a policy for AI/ML can cover several types of technologies, and we need to understand which ones are being addressed when writing our policies for a more accurate risk mitigation model:

1. ChatGPT as a specific service, due to its current popularity;
2. GenAI broadly, including systems other than ChatGPT such as Bard, PaLM, and Copilot;
3. Third party products and services using GenAI technologies;
4. Enterprise applications using on-premise or cloud-based GenAI technologies in internal development;
5. Third-party products and services making use of AI/ML technologies, or;
6. Internal development of AI/ML models and applications.

In this document we address the first, second, and third topics. As for the fourth, due to the limited amount of information available as of this writing we are not prepared to provide substantive guidance.

The area of GenAI-integrated third-party products is still nascent, and currently the risks which apply to GenAI in general apply to all third-party software products. Internal development of GenAI models is currently limited and we have insufficient information at this time.

Revise Existing Policies

Most AI/ML-related risks should be covered by existing enterprise policies. While every enterprise has unique requirements, these should be incorporated appropriately into its existing internal regulatory and policy body of work.

For example, policies for writing business emails, sharing data with third parties, or using third-party code projects, should already be well-established. These governing documents should be reviewed and updated for GenAI-specific risks.

It is incumbent on the CISO organization to educate our user base on the risks and how they are already governed by existing policies. Awareness campaigns around the subject should be considered.

Develop New Policies

The advent of GenAI and ChatGPT also introduces the need for new policies and controls, especially in the contexts where GenAI technology has had a significant impact on user and system behavior.

To further illustrate the last point: Users may have used cloud services for spell-checking in the past, which could have exposed sensitive data to a third party. But, they have not changed their entire content production workflow or provided raw data to be used for automated text generation and analysis work such as a presentation created with GenAI from uploaded documents.

Further, the incorporation of ChatGPT and other GenAI systems into third-party applications, from the Microsoft Office 365 suite of tools to browser plugins, is fast becoming ubiquitous and contributes to the rapid expansion of the risk surface.

Another area where a well-formulated policy can provide direction is in making the choice to use a large platform, one that supports many use cases, in contrast to selecting smaller use-case specific tools (for example platforms like copy.ai or jasper.ai used for copyrighting).

Implications for advocating one direction over another are similar to our current monoculture tradeoffs, e.g. using the same operating system everywhere for ease-of-use, maintenance, reduced attack surface, and patching, where the same operating system presents a homogeneous and potentially vulnerable infrastructure across-the-board.

A joint (and shared) responsibility

The CISO organization should be aligned with relevant stakeholders across the organization, and with any existing AI/ML team(s) in particular.

The 2020 Gartner State of AI Cyber Risk Management Study showed a sizable disconnect between Chief Information Security Officers' (CISOs') view of AI//ML risk and their AI/ML team's outlook. The October 2022 Gartner article Quick Answer: How Can Executive Leaders Manage AI Trust, Risk and Security? argues that misalignment on what risks may exist or occur demands the creation and agreement on policies across all roles and usage.

Further, not all GenAI risks are limited to the realm of cyber security, and thus mitigating potential risks requires a comprehensive enterprise approach. As we discussed above, risks such as in privacy and data protection, intellectual property exposure, sector-specific regulation, and AI ethics should be considered as part of a holistic risk management strategy.

Relevant enterprise functions, such as, but not limited to, the Chief Data Officer, Chief Information Officer, Data Protection Officer, Chief Risk Officer, and General Counsel, should be involved in the formulation of the strategy. The CISO organization can play an important role in coordinating a streamlined approach, and together, the organization can develop policy recommendations regarding risk level and recommended measures for board-level approval.

Considerations workflow

GenAI policies should be designed to provide clarity and direction on the the following topics, at a minimum:

- What are the requirements for the enterprise's use of GenAI, and how do these correlate with the risks, threats, and impacts described above? Are there any gaps?
- What are the risks and controls specific to the enterprise that have been identified and are applicable to the usage of GenAI?
- How do the GenAI security, privacy, data retention, and other policies and terms of service affect the enterprise and its choices around usage of GenAI?
- Are these customizable for users or customers?
- What enterprise business, application, and infrastructure dependencies could be impacted by the use of GenAI?
- Who can use GenAI, for what purposes, and under what circumstances?
- What integrations does the application utilizing GenAI provide access to, when considering the GenAI implementation?
  - › For example, does a customer support chatbot have access to all user data and is able to offer compensation on missed deliveries, service outages, and the like?
  - › Alternatively, how could the implementation affect enterprise systems on the back-end?
- Does the organization prefer to use specific tools per use case, or one multi-purpose platform?
- When a technology is chosen, does it access a larger platform for its operation? Is it a private instance of that platform (such as is reported to be the case with Office 365 and ChatGPT), or does it operate independently?

- Who can review and approve any proposed usage or experimentation?
- What devices or environments would be allowed for the usage of GenAI? (Workstations, VDI, Sandbox, Secure Enterprise Browser, etc.)
- Where and how should individuals in the organization report potential violations of policies or the discovery of sensitive data that might be stored or exposed in the GenAI data repository?
- Similarly, establish an elevation and an escalation path for individuals in the organizations to make contact, as well as report when the results of GenAI are suspect, and could affect the business.
- How can individuals in the organization, or the organization as a whole, opt-out to ensure that GenAI technologies don't use their prompts, or content, for further training?
- What are the security and governance measures designed to deal with applicable enterprise risk that must be in place before and while using GenAI to protect sensitive data?
- Does the GenAI technology take into account, and supports compliance, with ESG and corporate responsibility risks?

Some of the above questions and guardrails stem from the lack of ability to fully understand how AI/ML is constructed or operated, at times even by the developers themselves. The below are lower-level considerations that can, as the need arises depending on the organization's risk appetite, also be addressed in policy:

- **Capability Overhang:** Do we have a complete understanding of a model's capabilities?
  - › How do we put guardrails around models whose capabilities we do not completely understand?
  - › When there are hidden capabilities that were not planned for and are generally unknown even to the model developers – how do we account for these?
- **Model Drift:** How is Model drift being accounted for and managed by the GenAI, or by the enterprise?
  - › Model drift is not directly an enterprise risk, however, the enterprise needs to monitor for such "drift" over time, to avoid reduced accuracy and ability to effectively support the capability it was chosen for. These can lead to false or inaccurate results.
  - › Organizations will also need to monitor model drift for risk-level changes, and continually assess that it is staying within expected capability and use case scope. A low-risk model supporting enterprise use cases can become a high-risk model over time.
- **Edge Use Cases and Model Purpose:** Similar to the above, when selecting a platform, clarity on its intended purpose and what it was designed to do is paramount. If models are used for "edge use cases" that they were not designed for, this could lead to results which are inaccurate, incomplete, or false.

How should we interpret OpenAI's explicit statement on using prompt information for training?

OpenAI stated that the prompt information entered by users is used to train ChatGPT, after they remove PII. They add: "We also only use a small sampling of data per customer for our efforts to improve model performance." While it is possible that OpenAI employees could read through the mountains of prompt data

and take selective bits and use these to train the core model, this is unlikely to be done at scale. Instead, the training that is referred to by OpenAI is likely for the purposes of fine-tuning ChatGPT against a third aspect of the threat model – breaking guardrails (User Input Manipulation).

ChatGPT 3.5 was released publicly for the purposes of testing these guardrails. Over time, these have improved significantly as the general public found various adversarial examples (also known as "jailbreaking") that broke through the guardrails. The core ChatGPT model was trained on data dating back to September 2021.

That core model has not been retrained, and as such, we should be able to quickly dispel claims that ChatGPT is regurgitating user prompt data. As noted elsewhere in the document, this does not preclude the risk of such training happening in future versions of the model, opt-out policy aside.

Making Risk Decisions

Informed risk decisions can be reached by providing clarity and direction to the enterprise on the considerations listed above. Depending on its risk appetite, an enterprise could allow employees to share some enterprise data with a third party GenAI platform.

For example, assuming the enterprise considers allowing using tools developed and hosted by OpenAI, it should consider enforcing certain administrative guardrails, or principles:

- ☑ Selecting to opt-out of user prompt information being used to train future models, as it currently an option under OpenAI's policy;
- ☑ Accepting the data retention policy of 30 days, which is OpenAI's current policy;
- ☑ Requiring users to follow the Acceptable Use Policy, and to undergo risk-awareness training;

Then, other guardrails could be introduced as overlaying technical controls. It isn't clear to what degree these, and others, are possible at this stage:

- ☑ Restricting the use of GenAI generated code to be limited to open source and software packages with permitting licenses;
- ☑ Reviewing all images and graphics that were generated from GenAI queries for copyright and trademark infringement.

Another risk treatment approach may be to declare a permanent Risk Exception and allow employees to use a GenAI service as-is, or a temporary Risk Acceptance with a view to re-evaluate the decision at some later time.

Yet another way to manage the GenAI risks is to host the technology on-premises, where the enterprise security team has full control over hardening and privacy configurations. This solution is further discussed in this appendix.

Engaging Engineering Teams and Threat Modeling

So far we have focused our discussion on the enterprise consumers of GenAI. This section is intended for developers of GenAI systems, and more importantly for the purpose of this document, engineering teams that incorporate external GenAI capabilities into their own software. We also explore data leak concerns discussed earlier.

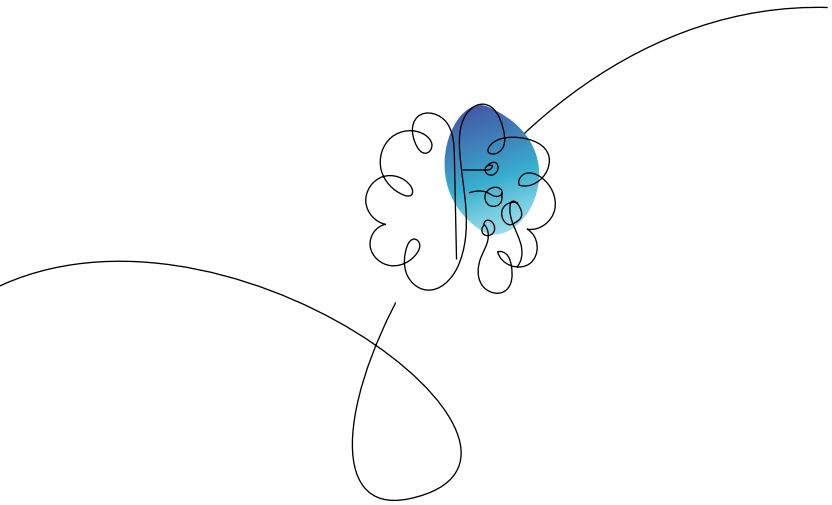
By fostering a threat-centric mindset, we can proactively address key security, privacy, and safety concerns, while still harnessing the capabilities offered by these tools. We explore the risks discussed from a technical standpoint, and introduce a taxonomy for further discussion.

In the Threat Modeling Manifesto, a set of authors use the Four Question Framework to frame threat modeling work:

1. What are we working on?
2. What can go wrong?
3. What are we going to do about it?
4. Did we do a good enough job?

A small sketch of the system can illustrate the focus of an analysis. For example, there are threats to the AI/ML tool, and threats to the users of the tool (or further parties). One example is prompt injection threats which are overall threats to the GenA system(s).

As can be seen from the above, there are many ways to address the question of what can go wrong. One additional example to understand the range of potential attacks is the Berryville Institute for Machine Learning's simple taxonomy used below. Here is that taxonomy with some minor adaptations and examples for clarity.



Object of Attack	Threats Against Confidentiality (a.k.a. Extraction)	Threats Against Integrity (a.k.a. Manipulation)
User Input	**IP Theft:** Losing intellectual property through the prompts. **Input Derivation:** Deriving input data from the outputs (a.k.a. "Model inversion")	**Prompt Injection:** Getting GenAI to act in a way that compromises wider system integrity - e.g. a document that when summarized by GenAI results in GenAI providing misleading information, or user input that causes GenAI to take actions that it should not from user input. **Jailbreaks:** Getting GenAI to be racist, criminal, etc. (a.k.a. "Adversarial examples"). Some consider these to be a sub-type of prompt injection.
Training Data	**Training Data Derivation:** Deriving training data from the outputs (a.k.a. "Model inversion" again).	**Data Poisoning:** Introducing bias, false information.
AI Model	**Model Theft:** Opening the black box.	**Backdoor Models:** Adding covert functionality which attackers can activate to cause the model to behave in ways it has not been designed to.

Examining GenAI through the lens of this threat model helps to dispel some of the concerns about the leakage of an organization's confidential data through the usage of GenAI. Some of these concerns are suspect and others are clear, so the threat model helps us to see the true nature and differences among the issues. There are three different aspects of the threat model in consideration.

The primary concern that most organizations have is the loss of intellectual property through User Input Extraction. As we expanded on above, the concern that the prompt inputs are used to train future models can be managed through the provider's opt-out process. Specifically for ChatGPT, OpenAI does not retain API-based inputs, but does reserve the right to use non-API-based inputs unless one chooses to opt-out.

A second concern arises from the belief that the user-provided prompt inputs will be incorporated into the core model such that one person's inputs (which may include intellectual property or other non-public information) might become part of another person's outputs, thereby resulting in the loss of intellectual property to not just the GenAI provider, but also potentially all the users of that GenAI provider.

GenAI providers that take this approach risk two major problems. First being concerns of the users who do not wish for their intellectual property to propagate further than needed. More importantly, the second problem of incorporating prompt inputs directly into the training of the core model is that it gives attackers a direct path to conduct Data Poisoning attacks (Training Data Manipulation).

Given well-established security failures stemming from improper handling of user-supplied inputs (e.g., buffer overflows, cross-site scripting, SQL injection), security teams should strongly advise engineering teams that are building their own GenAI models to avoid consuming user-supplied input as training data without proper safety controls.

The third concern arises from the jailbreaks that violate guardrails established by the GenAI provider. For ChatGPT specifically, the user-provided inputs are being actively used to train and fine-tune this part of the product to improve the guardrails and the overall user experience. It is computationally expensive to recreate the core model, and as such, the actual core model does not change with every new release of ChatGPT and has not changed since September 2021.

CONCLUSIONS

To enable GenAI technology adoption, organizations need to quickly and diligently identify and understand all the potential risks, threats, and impacts that may occur in their businesses. Reducing the pitfalls and unacceptable use of GenAI will require formulating risk mitigations expressed in clear policy statements, which can then be feasibly implemented within the context of the enterprise.

This document's recommendation is for enterprises to implement cross-organizational working groups to evaluate the risks, perform threat modeling exercises, and implement guardrails specific to their environment based on the risks highlighted.

The security community continues to research the subject, and our themes, findings, and recommendations will be updated and expanded as additional collaborative projects are formed and industry best practices documented.

APPENDIX 1

Open Source / On-Premise Alternatives

Although a threat model analysis may alleviate and dispel some concerns around the purported instances of intellectual property leakage or theft, there are still legitimate reasons to be concerned about the loss of User Input Confidentiality through the use of commercial services that operate as a SaaS offering. To address this concern, organizations may be interested in examining an on-premise alternative to the AI tools that are delivered through the cloud, and thus subject to potential interception by other parties.

The table below lists several open source alternatives that enable organizations to examine and instantiate an on-premise version that can be considered and used by organizations to address intellectual property concerns. Note that open source options have the potential for Backdoor Models (AI Model Manipulation) in accordance with the threat model described above.

Generative AI Goal	Commercial / SaaS-Based Examples	Open Source / On-Premise Options
Audio Transcription and Analysis	• Airgram, Descript, Otter • Chorus, Gong, Revenue.io	• Whisper
Image and Video Creation	• Hugging Face, Midjourney • Runway	• Stable Diffusion • ModelScope
Written Words and Code Generation	• ChatGPT, Jasper, Writesonic • Amazon CodeWhisperer, Ghostwriter, GitHub Co-Pilot	• Alpaca, LLaMA, Vicuna • CodeGeeX, GPT-Code-Clippy

APPENDIX 2

Sample GenAI and ChatGPT Policy Template

Purpose

[Organization Name] recognizes the potential benefits and risks associated with the use of Generative AI (we will refer to various Generative AI technologies and Large Language Models, or LLMs, collectively as "GenAI"). This policy outlines our commitment to responsible implementation of this technology to ensure that its use is consistent with our values and mission, business standards, security policies, and that the associated risks are appropriately managed.

Background

Recent GenAI innovations offer a multitude of business benefits which enterprises are actively exploring, and which the industry is reporting many employees are already leveraging.

Specifically, ChatGPT has caused a surge of interest on the topic. It is a chatbot developed by OpenAI that uses a machine-learning technique with natural spoken language for informational inquiries and responding back with human-like generated responses.

GenAI technologies introduce risks which the enterprise should be aware of, and prepare for.

For example, potential intellectual property exposure, the third party risks associated with using a GenAI platform, the data set leveraged to train the machine learning model, which has the potential to produce flawed or inaccurate responses, and other risks as described below.

Further, GenAI can be used to help attackers become more sophisticated, which affects our security program building. From a broader perspective, there are additional legal, regulatory, and privacy risks that should be considered.

This policy will provide guidance on practices the organization must or should adhere to, from writing an acceptable use policy, to developing user education and awareness campaigns.

Enterprise Risks

For the purposes of this section, review the risks table above, and identify which elements are relevant to your own organization, risk, and technology choices.

Corporate Policy

This policy provides guidelines for the use of GenAI at [Organization Name] and is intended to promote responsible and ethical use of this technology.

Acceptable Use Policy

The following is a list of guidelines for employees to follow when making use of GenAI generically, including ChatGPT. Employees should be trained on the appropriate use of the GenAI system and the relevant policies and regulations governing its use.

Violations of GenAI usage policies may result in disciplinary action, up to and including termination of employment.

- Employees must not disclose confidential or proprietary information to a GenAI technology, directly or through a third party application, unless through following the guidelines of the policy.
- Employees must use GenAI in a respectful and professional manner, refraining from using profanity, discriminatory language, or any other form of communication that could be perceived as offensive.
- Employees must comply with all relevant laws and regulations, including those related to data privacy and information security, according to our internal policy [policy name, link to the policy].
- Employees should report any concerns or incidents related to the use of GenAI to their supervisor or the appropriate department.

GenAI and ChatGPT implementation and integration guidelines

The following list are examples are potential guidelines enterprises may consider using for their

1. Use GenAI technology in a responsible manner that aligns with our mission and values:
   a. Ensure that any use of GenAI technology complies with applicable laws and regulations
   b. Conduct appropriate risk assessments to identify and manage potential risks associated with the use of GenAI technology
   c. Consider the potential impact of GenAI on stakeholders, including customers, employees, and partners
   d. Prepare awareness campaigns for employees and others leveraging the technology.
   e. Beyond awareness of the risks, and guardrails for engagement withGenAI technology, employees should also be reminded that:
      i. It is easy to forget the answers from the other side are not coming from a human.
      ii. What they input could potentially be reused by the GenAI technology in the future, when interacting with someone else.
   f. Prepare an elevation and escalation path for employees to make contact, or report in, both violations of the policy, as well as in case suspect results are returned from the GenAI technology

2. Identify any technology, infrastructure, or business processes and systems reliant on, making use of, or that have dependencies to or with GenAI technologies, that need to be evaluated and validated for all GenAI integrations with the enterprise:

a. Explicitly map the attack surface available to attackers reaching GenAI input

b. Implement the appropriate safeguards or controls to mitigate the risks associated with the use of GenAI technology, including measures to protect sensitive information, prevent unauthorized access, and ensure business continuity

c. All GenAI usage must be logged and archived according to applicable laws and regulatory requirements, to the degree possible with current technology, and properly detail where that is not possible

d. All access and actions taken by GenAI in general and on behalf of users should also be logged and easily searched for. Flag sensitive actions and anomalous values and amounts for further analysis

e. Research explainability in all aspects of technology you choose, to increase confidence and adoption.

3. Identify all data, intellectual property, integrations, internal and external applications and services a GenAI application or integration might access to and control:

a. Implement proper security and access controls

b. Provide the minimal access, data, and permissions necessary for the GenAI application or integration to perform its tasks

c. Minimal in this case will be defined both in the overall context for the application, or integration, but also per a specific request:

i. A customer support application should not have access to company source code

ii. A customer support application, when assisting a particular user, should not have access to other users' data

d. Always assume your GenAI may be compromised through Prompt Injection, and plan your mitigations and controls accordingly.

4. When building GenAI integrations:

a. Consider regulatory context and requirements for audits and compliance

b. Identify risks to intellectual property, terms and conditions, opt-out mechanisms, data retention policy, end-user license, or click-through agreements

c. Output is validated for accuracy and free from fabricated answers or citations.

As part of an RFI or risk assessment process, examine the privacy-by-design and security architecture considerations built into the system or model. There are significant differences between the various systems that require specific analysis for each.

APPENDIX 3

Future Potential Security Control Options

A draft list of security controls and mitigation strategies has been initially drafted, but it is a topic that is still being investigated and will be the subject of an upcoming Team8 CISO Village collaborative project, planned to be released as a paper in August 2023. It is included as a starting point for the community to research and consider. The August 2023 publication will be shared from the working group to provide recommended controls to consider for the growing list of risks being discovered across the spectrum of areas related to GenAI, and as new controls are introduced into the industry, also taking into account new ISO, NIST, and IEEE initiatives. Once again, user education and awareness training would be a good fit for most of the below, so will be left out of the table.

Risk	Control
Privacy and Confidentiality	• Legal disclaimer in privacy policies that mention AI is used in products or processes • Interactive and explicit end user opt-out when using services that have embedded GenAI
Enterprise, SaaS, and Third-party Security	• Filters, masks or scrubs sensitive content between organization APIs and chatbot AI services • Secure Enterprise browser
AI Behavioral Vulnerabilities (eg. Prompt Injection)	• Models should have input validation to catch malicious prompts • Model should have output validation to catch problematic behavior
Legal and Regulatory	• Review and negotiate, whenever possible, the third party policies and terms of use. • Licensing of content for use produced by GenAI technologies
Threat Actor Evolution	• Adjustment of social engineering training to consider targeted and high quality phishing and similar attacks
Copyright and Ownership	• Favor solutions trained on curated or licensed content, including the use of internally trained systems using the OpenAI API • Detection of intellectual property misuse, or plagiarism (GenAI for cases where content has been copied instead of generated) • Trademark detection
Insecure Code Generation	• Create a GenAI DMZ/staging ground to observe applications using AI/ML-generated code • Code review should include AI/ML-generated code, possibly marked as such
Bias and Discrimination	• Currently out of scope of this document, as it is a more generic AI/ML issue
Trust and Reputation	• Consider GenAI data use in enterprise system dependencies • Add AI content to review processes • Prompt filtering • Inclusion of a safety system on top of the AI app to filter and monitor responses.
Software Security Vulnerabilities	• Model interactions with other systems should be analyzed to identify potential interactions • Use model output filtering to identify problematic outputs
Availability, Performance, and Costs	• Map out infrastructure dependencies on systems using GenAI • Backup and redundancy • Recovery preparedness plan includes GenAI dependencies

APPENDIX 4

Enterprise Use Cases

While CISOs are researching GenAI, and ChatGPT specifically, enterprises and employees are already using them. The table below lists some of the primary use cases for GenAI in the enterprise that CISOs should consider when developing policies on the subject and to assist with ideation on other possible uses in your organization.

Experimenting with training models	Enterprise analysts can use GenAI to experiment with training models and improve the accuracy of language processing. This can help develop more sophisticated conversational agents and chatbots to handle complex customer queries and interactions.
Research and development to enhance current product offerings or internally developed software	Enterprise product and development teams can use GenAI for research and development to enhance current product offerings or software. GenAI can help develop and improve natural language processing algorithms, which can be later used in various applications such as voice recognition, sentiment analysis, and chatbots.
Use cases to enhance the quality of code	GenAI can be used by programmers in the software development lifecycle to improve code quality. By training the model on large datasets of code and programming languages, GenAI can identify potential known bugs (including security defects), provide suggestions for code optimization, and improve the overall efficiency of the development process.
Data analysis	Analysis of large amounts of data, such as customer feedback, to identify trends and insights.
Customer service and support	Automation of customer service and support tasks, such as answering common questions, assisting with product or service inquiries, and solving common problems.
Content creation	Generating content for marketing and advertising purposes, such as blog posts, social media posts, and email marketing campaigns, at least as a first draft. Generating Intellectual Property content, such as graphics and dialogs for relevant companies (e.g. game companies)
Personalization	Personalized customer interactions and experiences, such as recommending products or services based on their preferences or past behavior.
Automation	Automating tasks and processes, such as scheduling appointments, managing inventory, and handling routine administrative tasks.
Innovation	Generating ideas and insights to assist in the development of new products and services, as well as the improvement of existing ones.

APPENDIX 5

Curated Further Reading

Incidents

- ChatGPT Suffers First Data Breach, Exposes Personal Information
- Samsung Engineers Feed Sensitive Data to ChatGPT, Sparking Workplace AI Warnings
- Whoops, Samsung workers accidentally leaked trade secrets via ChatGPT
- GitHub Copilot under fire as dev claims it emits 'large chunks of my copyrighted code'
- Sydney, We Barely Knew You: Microsoft Kills Bing AI's Bizarre Alter Ego

Reactions and controversy

- Italy temporarily blocks ChatGPT over privacy concerns
- Artificial intelligence: stop to ChatGPT by the Italian SA Personal data is collected unlawfully, no age verification system is in place for children [Italian]
- Amazon warns employees not to share confidential information with ChatGPT after seeing cases where its answer 'closely matches existing material' from inside the company
- Germany could block ChatGPT if needed, says data protection chief - The Economic Times
- Walmart OKs ChatGPT for workers

Frameworks and papers

- NIST AI Risk Framework
- White House AI Bill of Rights - Agency Rule making and Legislation
- AI Trust Risk & Security Management; Gartner Market Guide - Avivah Litan
- Trustworthy AI | Deloitte
- MITRE | ATLAS™
- A Taxonomy of ML Attacks – Berryville Institute of Machine Learning

Law and regulation

- National AI Strategy [GOV.UK]
- The Artificial Intelligence Act [EU]
- FPF Report: Automated Decision-Making Under the GDPR - A Comprehensive Case-Law Analysis [Future of Privacy Forum]
- "Large Libel Models" Lawsuits, the Aggregate Costs of Liability, and Possibilities for Changing Existing Law
- Who Ultimately Owns Content Generated By ChatGPT And Other AI Platforms?
- US begins study of possible rules to regulate AI like ChatGPT
- The lawsuit against Microsoft, GitHub and OpenAI that could change the rules of AI copyright - The Verge
- EU Proposes Rules for Artificial Intelligence to Limit Risks - SecurityWeek
- Explained: What is the European Union AI Act, and it may mean for ChatGPT
- New EU AI Regulations Are Turning CISOs into Ambassadors of Trust - DATAVERSITY

Policy

- ChatGPT Risks and the Need for Corporate Policies [National Law Review]
- Gartner: Quick Answer: How Can Executive Leaders Manage AI Trust, Risk and Security?
- ChatGPT Risks and the Need for Corporate Policies
- AI Policy Observatory [OECD]
- Advancing accountability in AI [OECD]

ESG and corporate responsibility

- OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive
- The Power Requirements to Train Modern Large Language Models

Misc. articles and blogs

- ChatGPT and the security risks of Generative AI
- Beyond The PowerPoint: How To Implement Generative AI In Your Business
- How hackers can abuse ChatGPT to create malware
- What ChatGPT Reveals About the Urgent Need for Responsible AI

Books

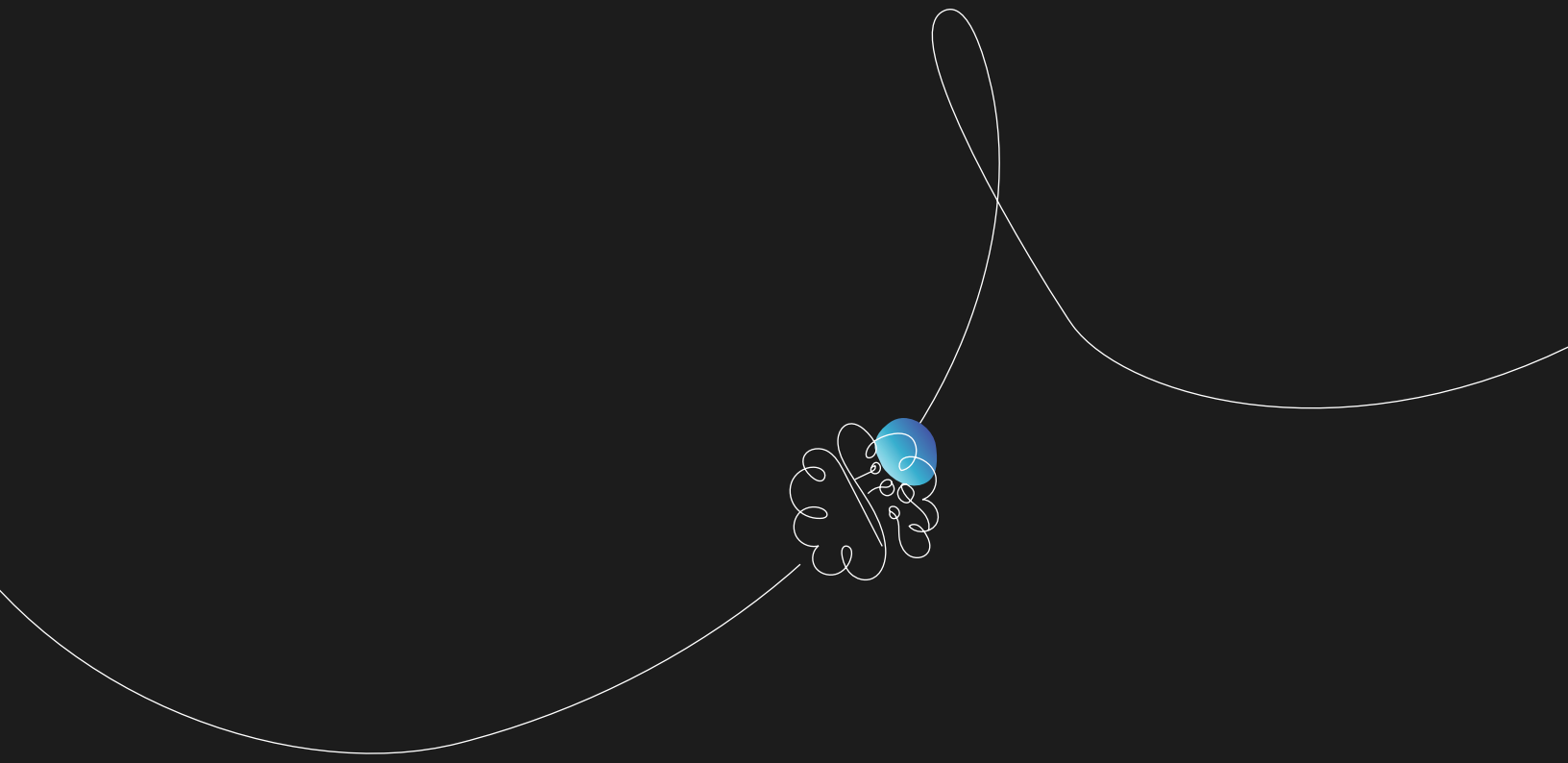
- Not with a Bug, But with a Sticker: Attacks on Machine Learning Systems and What To Do About Them

Prompt injections (a.k.a. AI Injection and AI Jailbreaking)

- AI Injection part 2: OpenAI's APIs are broken by design
- Reverse Prompt Engineering for Fun and (no) Profit
- The Hacking of ChatGPT Is Just Getting Started
- LLM Hacker's Handbook
- GPT Prompt Attack (CTF)
- Attacking LLM - Prompt Injectionw

OpenAI documentation

- OpenAI security page
- OpenAI security portal
- OpenAI's Bug Bounty Program
- API data usage policies
- How your data is used to improve model performance
- Data usage for consumer services FAQ



For more information

Contact us at: cisovillage@team8.vc | www.team8.vc

TEAM8™