CHARLI AI

Trust, Risk and Security Management in AI Systems

Charli AI Position Paper

January 2024

As AI is gaining traction in the industry, enterprises are faced with existential questions: **"If we trust an AI system with our business, who will be the adult in the room?"** In this document we go over the trust, risks, and security management (TRiSM) aspects in detail as it relates to AI solutions, specifically in the following three areas: **Data protection**, **content integrity** and AI **application security**.

Most AI vendors on the market are focusing on inherent AI functionality and woo customers by delivering the most impressive and value-add results, often at the expense of the integrity of the solution, including age old threats that every business is facing.

TRiSM specific offerings are entering the market to augment existing AI solutions, however it's often difficult to integrate due to compatibility challenges.

To navigate through this tumultuous evolution in the world of IT, it is important to understand the **threats** and **take appropriate steps** to take advantage of AI as an opportunity to disrupt entire market segments without jeopardizing existing assets and IP.

Generative AI has transformed the public opinion on Artificial Intelligence, driving both excitement and fear at the same time. Companies have needed to adapt extremely quickly to the initial challenges of almost overnight availability of extremely viral capabilities made available to their workforce. These new products offered no checks and balances to avoid catastrophic data leakage, copyright infringements, and sophisticated attack vectors threatening their existing infrastructure.

This was a very rude awakening for IT professionals after the already challenging "bring your own device" era with smartphones and tablets. But now, in addition to the initial wave of generative AI applications, there is a brand-new actor that had slipped through the conventional HR sensitivity and security training. Say hi to your new team member, **The AI Worker**.

AI Workers, also referred to as AI agents, are embedded in existing infrastructure and are performing high volume and repetitive tasks that cannot be carried out by traditional software applications. These types of AI deployments are poised to change the way businesses run and require them to be trusted and secure. AI workers have access to customer data and have far more capabilities in terms of processing and storing capacity than humans. This makes any unsafe action much riskier and more damaging. Selecting the right vendor that can deploy AI workers responsibly and ethically is critical. With the right AI solution, the security threat can be turned into an opportunity to improve a company's security posture through consistent and repeatable security processes.

Hackers and industrial espionage groups count on humans to make mistakes to exploit vulnerabilities. Reducing the risk of human error with AI can greatly reduce that exposure. In the next section, we will dive deeper to better understand the threats.

> "If we trust an AI system with our business, who will be the adult in the room?"

PART 1: THREATS

According to a Gartner survey, 42% of respondents were most concerned about **Data Protection** when it comes to their level of comfort with AI solutions. For many businesses, both internal and customer data is the foundation that fuels financial growth.

There are three areas to consider when assessing data protection with AI systems.

DATA PROTECTION

1. Initial Model Training

This is often the most daunting phase as AI vendors require access to example data to ensure adequate configuration, training, and fine tuning of AI models. It often requires effort and resources to ensure the vendor can be trusted with the data and ensure that even in an event of leakage, the data provided is not critical. The risk here comes from teams of data scientists and data labelers becoming a weak link in the effectiveness of data protection measures. Tight processes and guidelines are required on the vendor's side to ensure the training data is handled appropriately. An important aspect of training data is to ensure it is only going to benefit the company that provides the data unless explicit permission is granted to use the data to train generic AI models.

DATA PROTECTION

2. Continuous Model Training in Production

Scalable AI solutions will provide ways to have continuous learning in production, meaning that customer data and changes provided through human supervisions will be automatically used to optimize the AI models. It is important to understand how these models are being trained, where they reside and ensure that the necessary measures are taken to avoid models injecting answers that are proprietary into other customer use cases. It's important to consider that many companies require confidential handling of data not only to protect against competitors, but also to protect unauthorized access within the company.

Other phenomenon to consider are data drift and concept drift: This is the challenge of decreasing accuracy as the data evolves. This requires a continuous monitoring framework to measure the drift and remedy through dynamic learning techniques instead of static learning only.

3. Data processing and storage

Beyond issues with models, there are two other aspects of data protection; the location where the content being processed by AI and the resulting metadata created by the AI, or "AI Data" are stored. Both the static and dynamic threats need to be considered, the static aspects being: is the data protected, how is it retained, and what happens if the data is lost? The dynamic aspects being where does the data transit to and from, and is it encrypted? Traditional security and data protection practices can be applied here; it does however require the AI solution to be compatible with these practices and not open new attack vectors.

CONTENT INTEGRITY

The next area we are considering is **Content Integrity**. This means managing the risk to the business of generating inaccurate results. As with traditional content, quality control needs to be in place, but since AI demonstrates a high degree of confidence, it may be hard at first to detect integrity problems such as hallucination, inaccuracies, and biases.

1. Hallucinations

One of the main challenges in using large language models (LLMs) is the phenomenon of hallucination. This happens when the models are given full autonomy in deciding what consists of an acceptable answer or decision to a question or problem presented.

There are advanced techniques to avoid hallucination, including context window size chunking and retrieval augmented generation techniques. This allows for limiting of the scope of possible answers for the LLM and drastically lowers the occurrence of hallucinations.

CONTENT INTEGRITY

2. Bias and Inaccuracies

AI Models, like humans, always have biases. This is caused by the bias that exists in training data since it is not feasible to cover all possible permutations for a model that has billions of parameters. This creates output that lean toward the scope covered by the training data. In the same vein, accuracy can vary depending on how rich the training data is and how much deviation there is between the data used for training and the data processed in production. Having the appropriate mechanism in place to provide adequate oversight is critical to avoid uncaught errors generated by the AI. This will prevent incidents like sending an invoice to the wrong customer or generating a financial report with numbers taken from the wrong quarter.

APPLICATION SECURITY

3. Application security

Applications are facing new security challenges with AI where data processing, compute and third-party components are used to connect AI functions to existing infrastructure. Since AI interfaces handle requests and determine the answers, they increase the possibility of exploitation by attackers. An example is a database with payment information that may be completely secure, but asking the right question to an AI interface may return a payload with unauthorized access to the payment information. This is because most AI systems do not provide fine-grained security controls when it comes to portions or even single values needing protection.

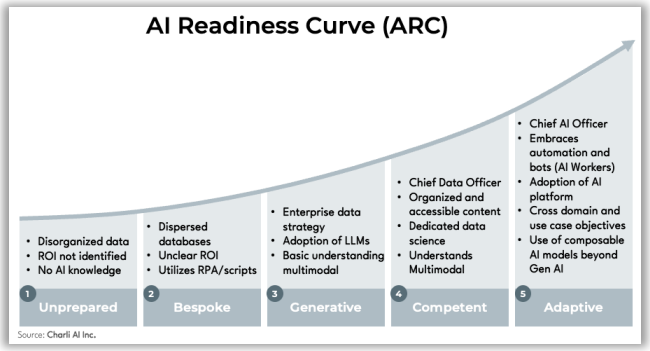
Additionally, existing application security teams have security policies implemented across the enterprise, and these policies must be propagated to the AI system to ensure a coherent security strategy and implementation across the board.

> "Say hi to your new team member, the AI Worker"

Now that we have identified the threats, let's look at how to get started. Preparation steps apply not only to TRiSM, but are fundamental to any AI initiative. A detailed readiness assessment needs to be conducted to avoid a costly false start. From a high level, Gartner suggests these three points to be addressed at a minimum:

1. AI Data Readiness
2. Clearly Defined AI Principles and Objectives
3. Sufficient Resources and Budget Allocated

Discussing with the AI vendor early about access to data and potential jurisdiction constraints is extremely important. This includes compliance requirements to regulations such as GDPR, CCPA and PIPEDA.



Charli AI's AI Readiness Curve

PART 2: TRISM

What should we be looking for as we select AI vendors and solutions?

Trust, Risk and Security Management are too often decoupled from vendor selection and are conducted as a post process once a vendor has been pre-selected. It is important to understand the key characteristics of a mature AI platform to avoid delays in the process. Here are several topics that you should ask your vendors and what to look for in the provided answers.

TRUST

AI Explainability

There are many technical and process factors involved in providing responsible and explainable AI, especially in highly regulated industries such as financial services. Having a digital paperwork audit trail is only one method used to deal with regulations requirements, but by itself is insufficient for transparency and confidence on AI decision making. The methods that must be incorporated in the Platform include traceability on end-to-end processing at every step in a decision flow. This includes full transparency into the input and output operations from each of the steps, expanding into the micro-decisions made within complex decision flows. Furthermore, the AI must provide rich metadata associated with processing including weightings, confidence levels, and outcome arrays. Moreover, the decision flow engine needs to maintain visible/observable metadata on why processes are executed to further enhance the explanations.

RISK

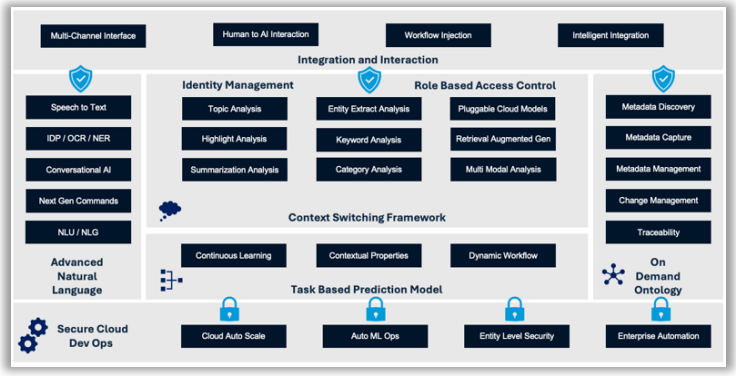
Data protection across groups and organizations

The knowledge capture and the learnings must be captured at the user-level to streamline user operations, processes and analysis. Technically, this is at the account-level and users will have accounts, but systems can also have accounts. This allows for fine-grained control over the knowledge and the learning system. Knowledge capture (in the form of entities in the metadata repository) and learnings at the group or organization level need to be configured such that the fine-grained information from accounts can flow into the training data used for group/organization models. This flow of information is configured to enforce policies within organizations and ensure accurate mapping to the organization structure for groups/departments/etc.

RISK

There are two primary concepts within the system that need to be applied to ensure security in this case including Identity Management and Entity Management. Identity Management concepts are similar to what might be defined in commonly used IAM/RBAC systems and allow for fine-grained control over account/user/group/organization definitions and scopes. Entity Management is a very fine-grained policy control over how entities are tracked, traced and attributed from a data lineage perspective. This is similar in concept to ABAC controls – with much more traceability on data as it flows through the system.

Ensuring that entities are used in knowledge and training at the account-level, group-level and organization-level is done through configuration of the permissions/mappings between identity definitions and entities.



Charli AI's Ancaeus is a secure by design AI Platform

SECURITY

Security Model and Data Exchange

The security model must be designed to adapt to the customers model, which requires having the ability to apply security policies in the most fine-grained manner to ensure the full support of the customer model is enabled through the AI Platform. The security framework must support many methods for authentication/authorization as well as layered and fine-grained access to data, even down to the element level. These policies are defined by our customers and followed by the platform for read/write/update access where necessary within the use cases and workflows.

SECURITY

Data exchange should be fully supported and allowing for secure reading of data and exchanging data with systems/data mesh as part of its workflow. The AI tracks a rich set of metadata, and this metadata is also available for use by customers. Even the data "wrangling" results can be shared/exchanged securely with customers.

AI Model Training Principles

We recommend that organizations only select vendors that can guarantee that no customer data will be used to train generic models. This requires model separation. It is also important to have the ability to forget a set of training data, as some customers need full control of the proprietary data used for model training, ensuring that sensitive data can be retracted from a deployment when needed.

In addition, the vendor must have a specific sensitive data handling process in place for the model training phase and a contractual obligation that customer-specific models are segregated and only used for this customer use case.

> "We recommend that organizations only select vendors that can guarantee that no customer data will be used to train generic models."

Typical AI platform built in security features

Industry best in class protection must be included.

- ☐ SOC2 compliance to have all the necessary procedural requirements to secure data.
  - 2FA on all systems (oAuth)
  - Compartmentalized access across multiple users
  - Kubernetes Automated monitoring
  - Disaster recovery with regular dry runs on it.
- ☐ Ability to deploy in private cloud environment for very sensitive or regulated use cases. Central knowledge repository tied directly to customer instance, storing only metadata.

AI in a federated environment

There must be a wide variety of configurations and capability to discover data across a federated environment and leverage that data as needed. The dependency has more to do with access control and discovery mechanisms (semantic search, crawling, etc.) than it does with limitations in the AI framework.

SECURITY

There are also many methods of deployment, even hybrid deployments in cases where data might be remote. Look for Platforms that operate in Public Cloud, Private Cloud and On-Premises deployments and provide an infrastructure for AIOps.

In most cases, access/discovery can be granted directly to the AIOps infrastructure, and this allows the AIOps environment to tap into cloud and enterprise data sources directly.

This greatly simplifies the overall infrastructure and reduces risk as per-node deployment is not required. Data movement between processes and location must be encrypted to avoid unauthorized access.

Recommendations

As you embark on a new AI initiative, we recommend that the following steps are taken.

1. Review your data readiness and discuss with vendors what processes they have in place to protect your data. Ensure you have anonymized or readied to-be-shared data samples to avoid delays in your engagement.
2. Clearly define your objectives, not just technical but business objectives, and share these with your vendors. This will avoid surprises during the Pilot and production deployment.
3. Plan for budget and resources to ensure the vendor has adequate support, information and feedback as the use case is being configured.

References

1. Gartner: Research paper **Innovation Guide for Generative AI in Trust, Risk and Security Management** Published 5 December 2023 - By Analyst(s): *Avivah Litan, Jeremy D'Hoinne, Gabriele Rigon*