Overview
OWASP AI Exchange

(owaspai.org)

Aruneesh Salhotra

May 9th

About me and Journey with OWASP AI Exchange

- LinkedIn: https://www.linkedin.com/aruneeshsalhotra
- Blogs: https://www.snmconsultinginc.com/blogs/

About Me

- Technologist **Generalist**
- Fractional CISO
- Life time learner
- Avid **Technologist** and **Researcher**
- Passion for **Marketing and Promotion**
- NY Resident
- Passionate about non-profits
- Educating the masses about the connecte world and risk associated

OWASP and Me

- Lead an indepedent GenAI and Quantum Research Group
- Been with AI Exchange for 5 months
- Lead **Marketing** and **Step By Step** Adop Guide
- Humbled to represent on behalf of AI Exchange Leader **Rob Van der Veer** and OWASP AI Exchange here today

OWASP AI Exchange

Mission:
Be the authoritative source for consensus,
foster alignment,
and drive collaboration among initiatives
- NOT to set a standard.

Be the **top bookmark** for people in AI security

By doing so, it provides a safe, open, and independent place to find and share insights for everyone.

KEY FACTS

- Exchange is CC0 1.0 licensed: **free of copyright** and attribution.
- Volume is **130 pages**
- Accessible and Available at **owaspai.org**

A **collaborative working document** providing **comprehensive overview of AI threats, vulnerabilities, and controls** to foster alignment among different standardization initiatives.

Relation to other OWASP or other organization initiatives

- The **OWASP AI security and privacy guide** is the official OWASP project under which the AI Exchange was established. The deliverable consists of the **AI Exchange content plus guidance on AI privacy**.
- The OWASP LLM top 10 provides a list of the **most important LLM security issues**, plus deliverables that focus on LLM security, such as the LLM AI Security & Governance Checklist.
- The OWASP ML top 10 provides a list of the **most important machine learning security issues**.
- OpenCRE.org **holds a catalog of common requirements across various security standards** inside and outside of OWASP.

Target Audience

- Cyber Security Experts
- Privacy/Regulatory Professionals
- Legal professionals
- AI leaders
- Developers
- Data scientists

Scope & Responsibilities

- Develop a **comprehensive framework for AI threats, risks, mitigations, and controls**.
- Create a map integrating **AI regulatory and privacy regulations.**
- Establish a **common taxonomy** and glossary for AI security.
- Provide **guidance on testing tools** with outcome assessments.
- Formulate a **shared responsibility model** for third-party AI model usage.
- Offer **supply chain guidance** and an incident response plan.

Collaboration Efforts and Engagement (#)

- We have regular collaboration with
  - CSA
  - ISO/IEC
- Liaison with CEN/CENELEC
- We have regular meetings with
  - NIST
  - MITRE
  - ITU
- We're part of the AISIC.

Furthermore we provide the content as CC0, so free of copyright and attribution.

Key Achievements

- Our direct flow of content into **ISO/IEC 27090** and standards for the **EU AI Act**.
- The liaison relation with CEN/CENELEC.
- **Recognition by standard makers** across the globe

Copyright

- The AI security community is marked with CC0 1.0 (Creative Common) meaning you can use any part freely, without attribution. If possible, it would be nice if the OWASP AI Exchange is credited and/or linked to, for readers to find more information.

EcoSystem and Standards Bodies

OWASP

Machine Learning Security Top 10

OWASP

OWASP Top 10 for LLM

Free of Copyright

OWASP AI Exchange

Global Security and AI Community

ISO

NIST

MITRE

How to address AI Security?

Imperative to approach AI applications with a **clear understanding of potential threats** and which of those threats to prioritize for each use case.

Standards and governance help guide this process for individual entities leveraging AI capabilities.

- Implement **AI governance**
- Extend **security and development practices**
- Improve regular application and system security through **understanding of AI particularities**
- **Limit the impact of AI** by minimizing privileges and adding oversight
- **Countermeasures in data science** through understanding of model attacks

Threat Model

Threats

We distinguish **three types of threats**:

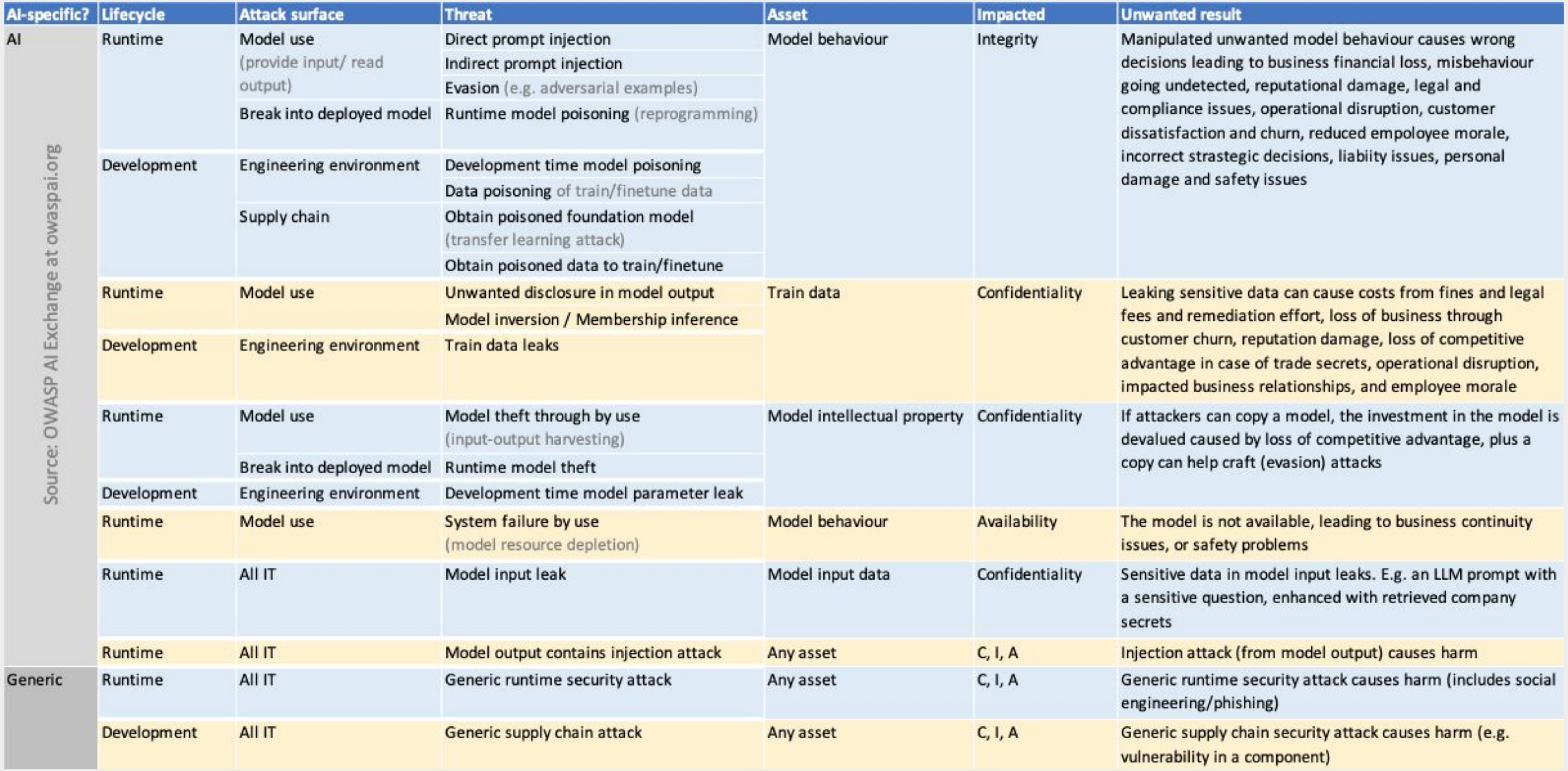
1. During development-time
2. Through using the model
3. By attacking the system during runtime

Impacts

In AI we distinguish 6 types of impacts:

1. Confidentiality of **train/test data**
2. Confidentiality of **model Intellectual property** (the *model parameters* or the process and data that led to them)
3. confidentiality of **input data**
4. integrity of **model behaviour** (the model is not manipulated to behave in an unwanted way)
5. **availability** of the model
6. confidentiality, integrity, and availability of **non AI-specific assets**

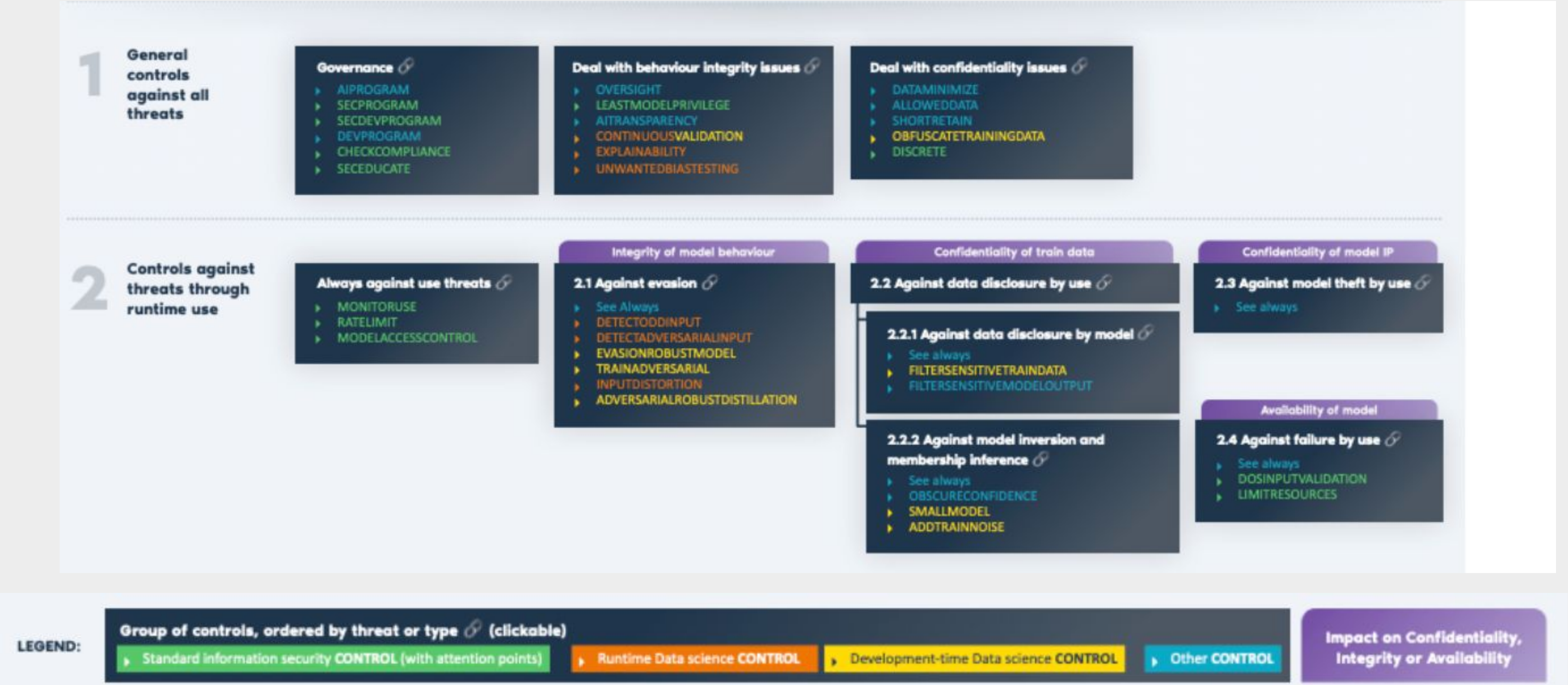
AI Security Matrix



AI-specific?	Lifecycle	Attack surface	Threat	Asset	Impacted	Unwanted result
AI	Runtime	Model use (provide input/ read output)	Direct prompt injection	Model behaviour	Integrity	Manipulated unwanted model behaviour causes wrong decisions leading to business financial loss, misbehaviour going undetected, reputational damage, legal and compliance issues, operational disruption, customer dissatisfaction and churn, reduced empoloyee morale, incorrect strastegic decisions, liabiity issues, personal damage and safety issues
			Indirect prompt injection			
			Evasion (e.g. adversarial examples)			
		Break into deployed model	Runtime model poisoning (reprogramming)			
	Development	Engineering environment	Development time model poisoning			
			Data poisoning of train/finetune data			
		Supply chain	Obtain poisoned foundation model (transfer learning attack)			
			Obtain poisoned data to train/finetune			
	Runtime	Model use	Unwanted disclosure in model output	Train data	Confidentiality	Leaking sensitive data can cause costs from fines and legal fees and remediation effort, loss of business through customer churn, reputation damage, loss of competitive advantage in case of trade secrets, operational disruption, impacted business relationships, and employee morale
			Model inversion / Membership inference			
	Development	Engineering environment	Train data leaks			
	Runtime	Model use	Model theft through by use (input-output harvesting)	Model intellectual property	Confidentiality	If attackers can copy a model, the investment in the model is devalued caused by loss of competitive advantage, plus a copy can help craft (evasion) attacks
		Break into deployed model	Runtime model theft			
	Development	Engineering environment	Development time model parameter leak			
	Runtime	Model use	System failure by use (model resource depletion)	Model behaviour	Availability	The model is not available, leading to business continuity issues, or safety problems
	Runtime	All IT	Model input leak	Model input data	Confidentiality	Sensitive data in model input leaks. E.g. an LLM prompt with a sensitive question, enhanced with retrieved company secrets
	Runtime	All IT	Model output contains injection attack	Any asset	C, I, A	Injection attack (from model output) causes harm
Generic	Runtime	All IT	Generic runtime security attack	Any asset	C, I, A	Generic runtime security attack causes harm (includes social engineering/phishing)
	Development	All IT	Generic supply chain attack	Any asset	C, I, A	Generic supply chain security attack causes harm (e.g. vulnerability in a component)

Source: OWASP AI Exchange at owaspai.org

AI Security Threats and Controls Navigator



1 General controls against all threats

Governance
- AIPROGRAM
- SECPROGRAM
- SECDEVPROGRAM
- DEVPROGRAM
- CHECKCOMPLIANCE
- SECEDUCATE

Deal with behaviour integrity issues
- OVERSIGHT
- LEASTMODELPRIVILEGE
- AITRANSPARENCY
- CONTINUOUSVALIDATION
- EXPLAINABILITY
- UNWANTEDBIASTESTING

Deal with confidentiality issues
- DATAMINIMIZE
- ALLOWEDDATA
- SHORTRETAIN
- OBFUSCATETRAININGDATA
- DISCRETE

2 Controls against threats through runtime use

Always against use threats
- MONITORUSE
- RATELIMIT
- MODELACCESSCONTROL

Integrity of model behaviour

2.1 Against evasion
- See Always
- DETECTODDINPUT
- DETECTADVERSARIALINPUT
- EVASIONROBUSTMODEL
- TRAINADVERSARIAL
- INPUTDISTORTION
- ADVERSARIALROBUSTDISTILLATION

Confidentiality of train data

2.2 Against data disclosure by use

2.2.1 Against data disclosure by model
- See always
- FILTERSENSITIVETRAINDATA
- FILTERSENSITIVEMODELOUTPUT

2.2.2 Against model inversion and membership inference
- See always
- OBSCURECONFIDENCE
- SMALLMODEL
- ADDTRAINNOISE

Confidentiality of model IP

2.3 Against model theft by use
- See always

Availability of model

2.4 Against failure by use
- See always
- DOSINPUTVALIDATION
- LIMITRESOURCES

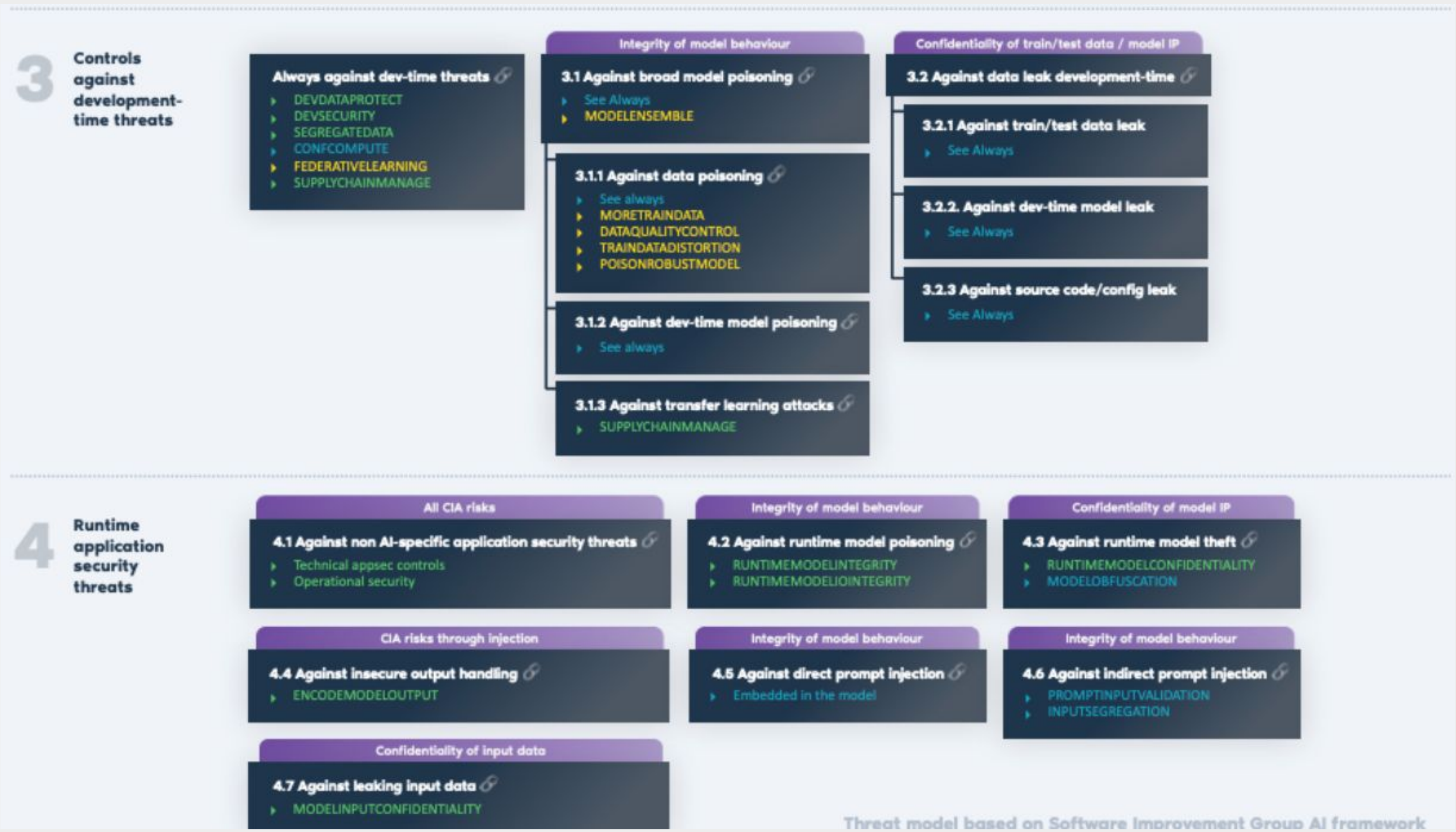
LEGEND:

Group of controls, ordered by threat or type (clickable)

- Standard information security CONTROL (with attention points)
- Runtime Data science CONTROL
- Development-time Data science CONTROL
- Other CONTROL

Impact on Confidentiality, Integrity or Availability

AI Security Threats and Controls Navigator



3 Controls against development-time threats

Always against dev-time threats
- DEVDATAPROTECT
- DEVSECURITY
- SEGREGATEDATA
- CONFCOMPUTE
- FEDERATIVELEARNING
- SUPPLYCHAINMANAGE

Integrity of model behaviour

3.1 Against broad model poisoning
- See Always
- MODELENSEMBLE

3.1.1 Against data poisoning
- See always
- MORETRAINDATA
- DATAQUALITYCONTROL
- TRAINDATADISTORTION
- POISONROBUSTMODEL

3.1.2 Against dev-time model poisoning
- See always

3.1.3 Against transfer learning attacks
- SUPPLYCHAINMANAGE

Confidentiality of train/test data / model IP

3.2 Against data leak development-time

3.2.1 Against train/test data leak
- See Always

3.2.2. Against dev-time model leak
- See Always

3.2.3 Against source code/config leak
- See Always

4 Runtime application security threats

All CIA risks

4.1 Against non AI-specific application security threats
- Technical appsec controls
- Operational security

Integrity of model behaviour

4.2 Against runtime model poisoning
- RUNTIMEMODELINTEGRITY
- RUNTIMEMODELIOINTEGRITY

Confidentiality of model IP

4.3 Against runtime model theft
- RUNTIMEMODELCONFIDENTIALITY
- MODELOBFUSCATION

CIA risks through injection

4.4 Against insecure output handling
- ENCODEMODELOUTPUT

Integrity of model behaviour

4.5 Against direct prompt injection
- Embedded in the model

Integrity of model behaviour

4.6 Against indirect prompt injection
- PROMPTINPUTVALIDATION
- INPUTSEGREGATION

Confidentiality of input data

4.7 Against leaking input data
- MODELINPUTCONFIDENTIALITY

Threat model based on Software Improvement Group AI framework

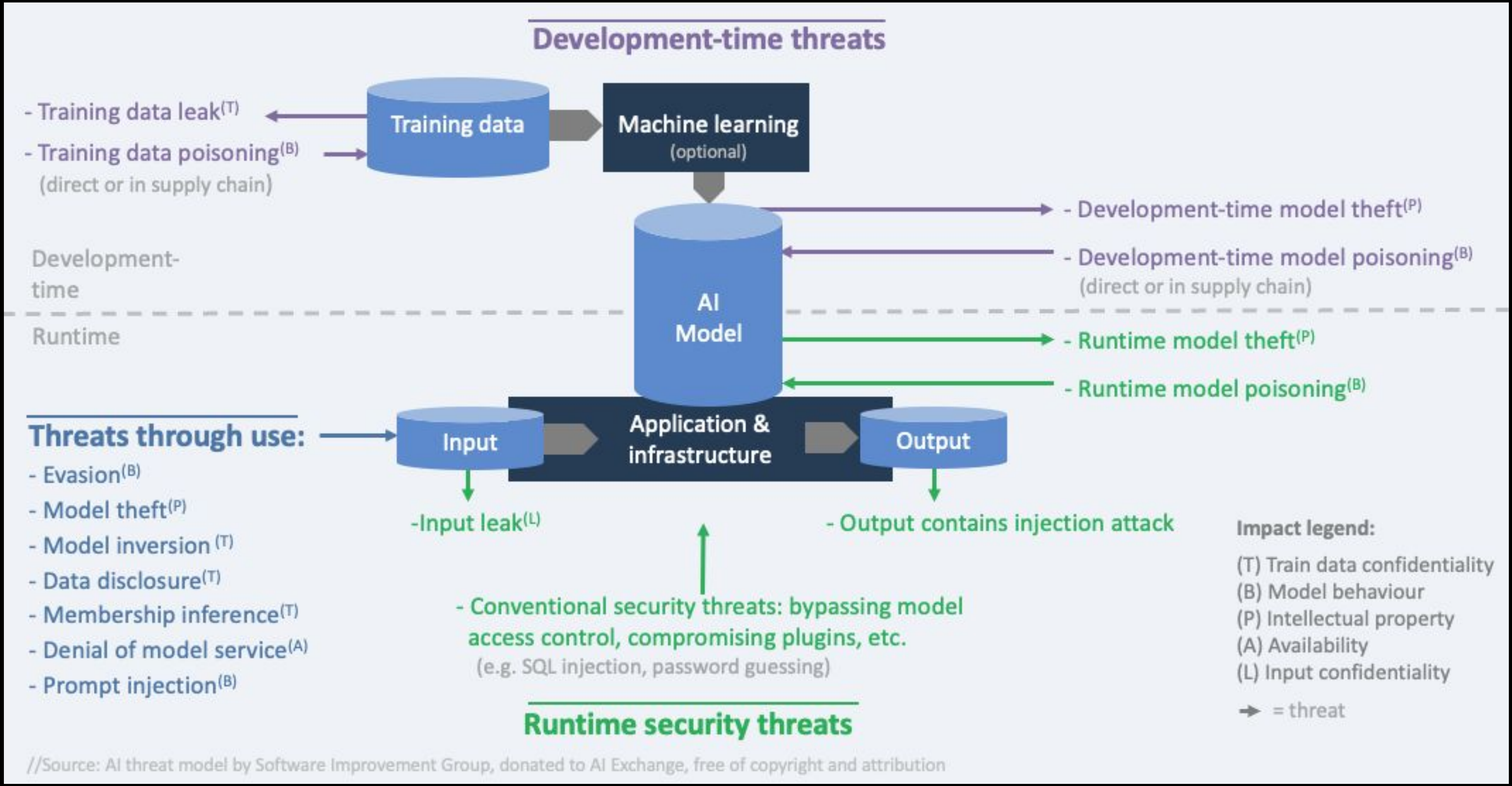
LEGEND:

Group of controls, ordered by threat or type (clickable)

- Standard information security CONTROL (with attention points)
- Runtime Data science CONTROL
- Development-time Data science CONTROL
- Other CONTROL

Impact on Confidentiality, Integrity or Availability

Threat and Impact



Development-time threats

- Training data leak(T)
- Training data poisoning(B) (direct or in supply chain)

Training data

Machine learning (optional)

AI Model

- Development-time model theft(P)
- Development-time model poisoning(B) (direct or in supply chain)

Development-time

Runtime

- Runtime model theft(P)
- Runtime model poisoning(B)

Threats through use:

- Evasion(B)
- Model theft(P)
- Model inversion (T)
- Data disclosure(T)
- Membership inference(T)
- Denial of model service(A)
- Prompt injection(B)

Input

Application & infrastructure

Output

-Input leak(L)

- Output contains injection attack

- Conventional security threats: bypassing model access control, compromising plugins, etc. (e.g. SQL injection, password guessing)

Runtime security threats

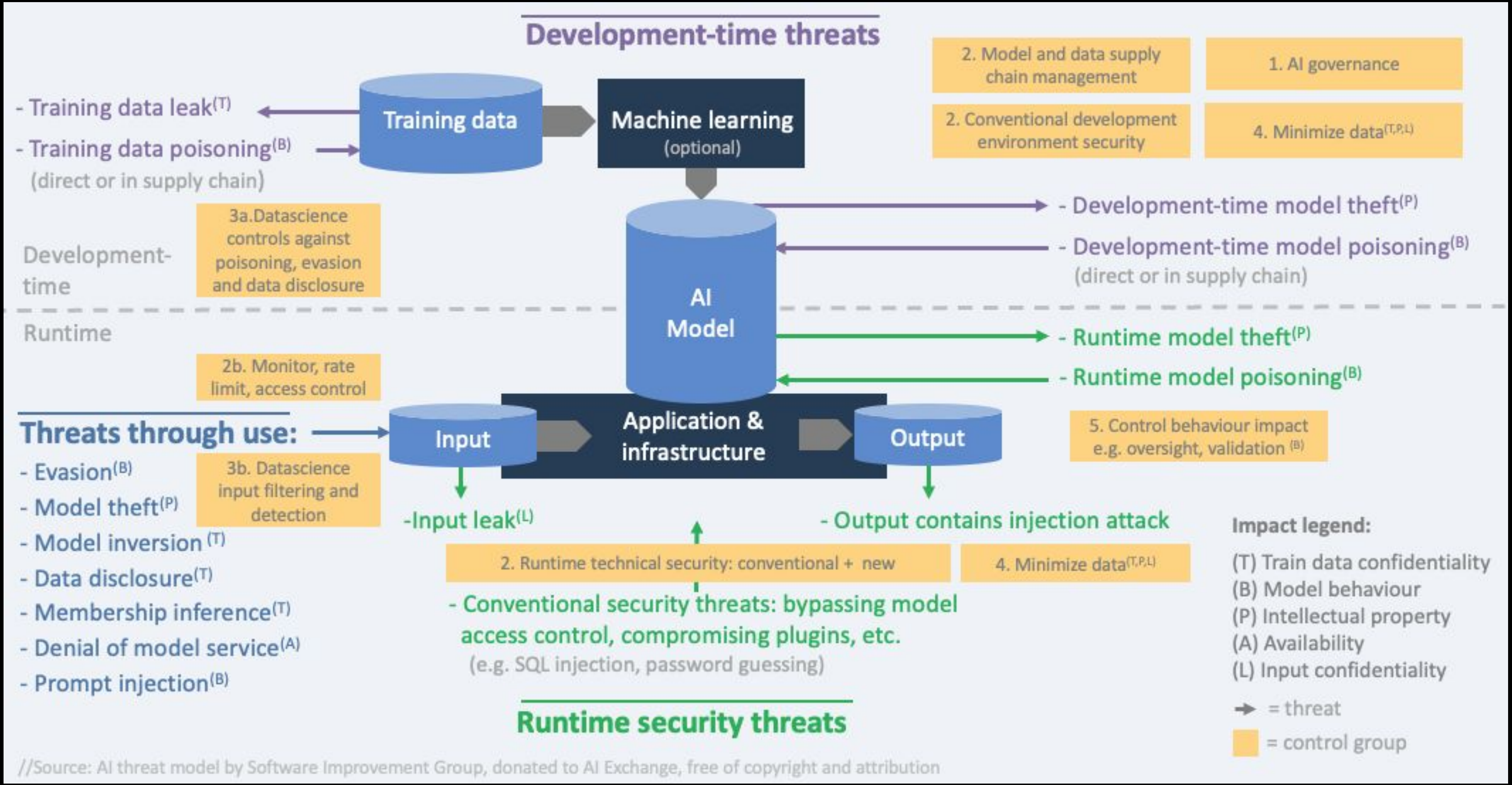
Impact legend:

(T) Train data confidentiality
(B) Model behaviour
(P) Intellectual property
(A) Availability
(L) Input confidentiality

→ = threat

//Source: AI threat model by Software Improvement Group, donated to AI Exchange, free of copyright and attribution

Threat and Impact with Controls



Development-time threats

- Training data leak(T)
- Training data poisoning(B) (direct or in supply chain)

Training data

Machine learning (optional)

AI Model

2. Model and data supply chain management

1. AI governance

2. Conventional development environment security

4. Minimize data(T,P,L)

- Development-time model theft(P)
- Development-time model poisoning(B) (direct or in supply chain)

Development-time

3a.Datascience controls against poisoning, evasion and data disclosure

Runtime

2b. Monitor, rate limit, access control

- Runtime model theft(P)
- Runtime model poisoning(B)

5. Control behaviour impact e.g. oversight, validation (B)

Threats through use:

- Evasion(B)
- Model theft(P)
- Model inversion (T)
- Data disclosure(T)
- Membership inference(T)
- Denial of model service(A)
- Prompt injection(B)

3b. Datascience input filtering and detection

Input

Application & infrastructure

Output

-Input leak(L)

- Output contains injection attack

2. Runtime technical security: conventional + new

4. Minimize data(T,P,L)

- Conventional security threats: bypassing model access control, compromising plugins, etc. (e.g. SQL injection, password guessing)

Runtime security threats

Impact legend:

(T) Train data confidentiality
(B) Model behaviour
(P) Intellectual property
(A) Availability
(L) Input confidentiality

→ = threat
🟨 = control group

//Source: AI threat model by Software Improvement Group, donated to AI Exchange, free of copyright and attribution

General Controls

The following threats and controls are highlights from the AI Exchange of which most are not in the llm top 10.

Governance

Controls
AI Program
Security Program
Secure Development Program
Development Program
Check Compliance
Security Education

Limit the effects of unwanted behavior

Control
Oversight
Least Privilege
AI Transparency
Continuous Validation
Explainability
Unwanted Bias Testing

Sensitive Data Limitation

Control
Data Minimization
Allowed Data
Short Retain
Discrete

Threats Through Use

Model Behavior Manipulation

Threat and Impact
Evasion
Closed-box evasion
Open-box evasion
Evasion After Data Poisoning

Development-Time Threats

Introduction

Control	Description
Development Security	Sufficient security of the AI development infrastructure, also taking into account the sensitive information that is typical to AI: training data, test data, model parameters and technical documentation
Segregate Data	Store sensitive development data (training or test data, model parameters, technical documentation) in a separated areas with restricted access.
Confidential Compute	If available and possible, use features of the data science execution environment to hide training data and model parameters from model engineers - even while it is in use.
Federated Learning	Federated learning can be applied when a training set is distributed over different organizations, preventing that the data needs to be collected in a central place - increasing the risk of leaking.
Supply Chain Management	Managing the supply chain to minimize the security risk from externally obtained elements. In regular software engineering these elements are source code or software components (e.g. open source).

Sensitive Data Leak During Development

Threat and Impact	Description
Development-time data leak	Unauthorized access to train or test data through a data leak of the development environment. This has an Impact on the confidentiality breach of sensitive train/test data.
Model theft through development-time model parameter leak	Unauthorized access to model parameters through a data leak of the development environment.This has an impact on the confidentiality breach of model intellectual property.
Source code/configuration leak	Unauthorized access to code or configuration that leads to the model, through a data leak of the development environment. SUch code or configuration is used to preprocess the training/test data and train the model. This has a direct impact on confidentiality breach of model intellectual property.

Runtime Application Security Threats

Leak Sensitive Input Data (*)

Control	Description
Leak Sensitive Input Data	Input data can be sensitive (e.g. GenAI prompts) and can either leak through a failure or through an attack, such as a man-in-the-middle attack. Impact: Confidentiality breach of sensitive input data.

Roadmap

Key Deliverables

- Prep 1.0: Review by community and by ourselves -> release 1.0
- Feed the Exchange 1.0 into at least the AI Act and ISO 27090
- Make it easier for readers to recognize their deployment model and select only what is relevant to them
- More illustration of threat models and attack vectors
- Further alignment with Mitre Atlas, NIST, the LLM Top 10, ENISA’s work, and the AIAPP International Privacy Group

Get Involved and Contribute

Engage with the OWASP AI team through various platforms.

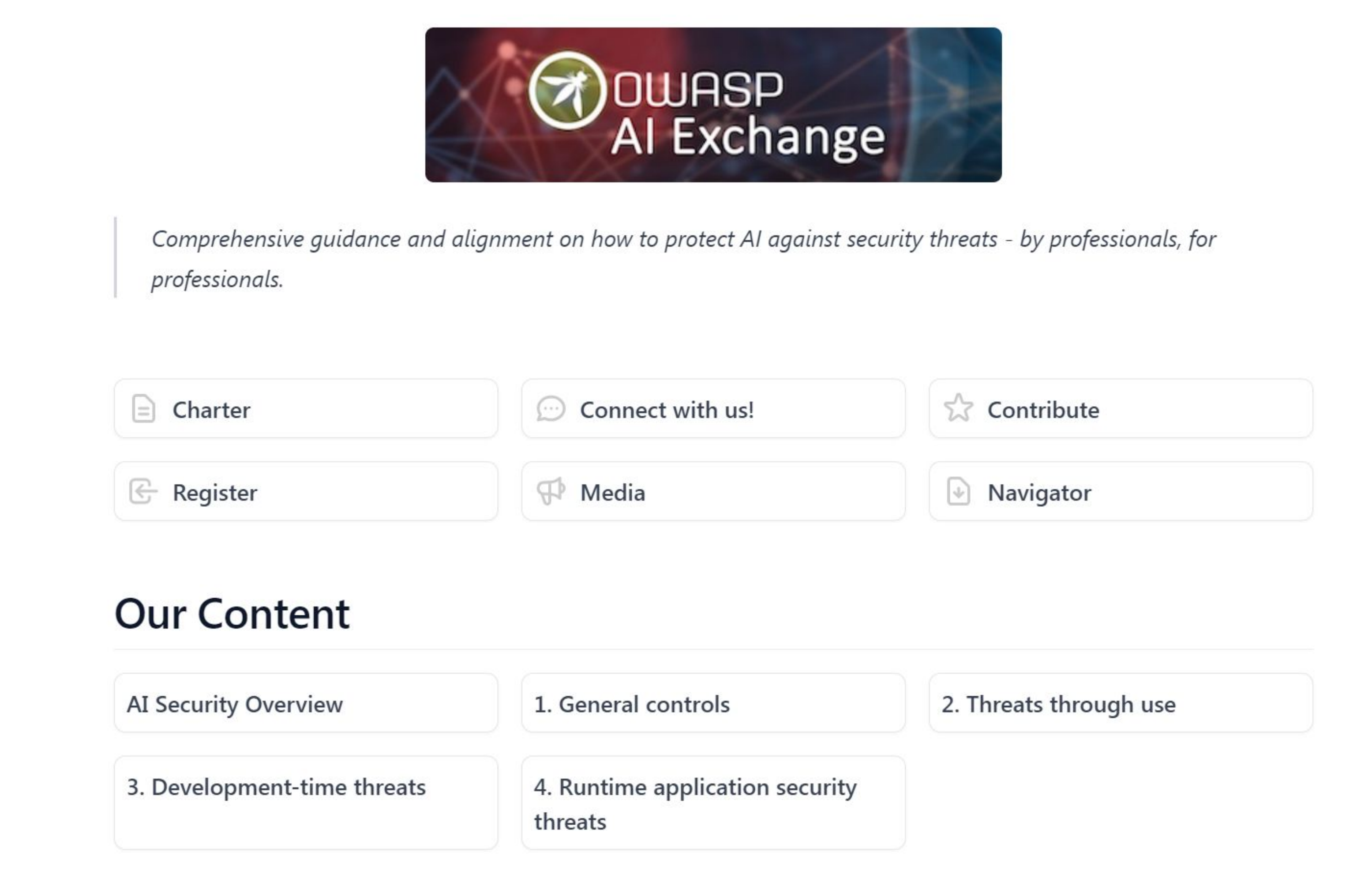
- **Connect** with us on the OWASP Slack workspace in the `#project-ai-community` channel. Authors are in the closed `#project-ai-authors` channel.
- Keep up with the latest **updates** by following us on Twitter and LinkedIn.
- For technical inquiries and suggestions, **participate** in our GitHub Discussions, or report and track issues on GitHub Issues.

If contributing interests you, check out our Contribution Guidelines or get in touch with our project leaders.

The Exchange is built on expertise from contributors around the world and across all disciplines.

Where can I find more information?

OWASPAI.ORG



> *Comprehensive guidance and alignment on how to protect AI against security threats - by professionals, for professionals.*

Our Content

AI Security Overview
1. General controls
2. Threats through use
3. Development-time threats
4. Runtime application security threats

Participate in Content Development

- 📥 Send your suggestion to the project leader.
- 👋 Join `#project-ai-community` in our Slack workspace.
- 🗣️ Discuss with the project leader how to become part of the writing group.
- 💡 Propose your concepts, or submit an issue.
- 📄 Fork our repo and submit a Pull Request for concrete fixes (e.g. grammar/typos) or content already approved by the core team.
- 🙌 Showcase your contributions.
- 🐞 Identify an issue or fix it on a Pull Request.
- 💬 Provide your insights in GitHub Discussions.
- 🙏 Pose your questions.

OWASPAI.ORG

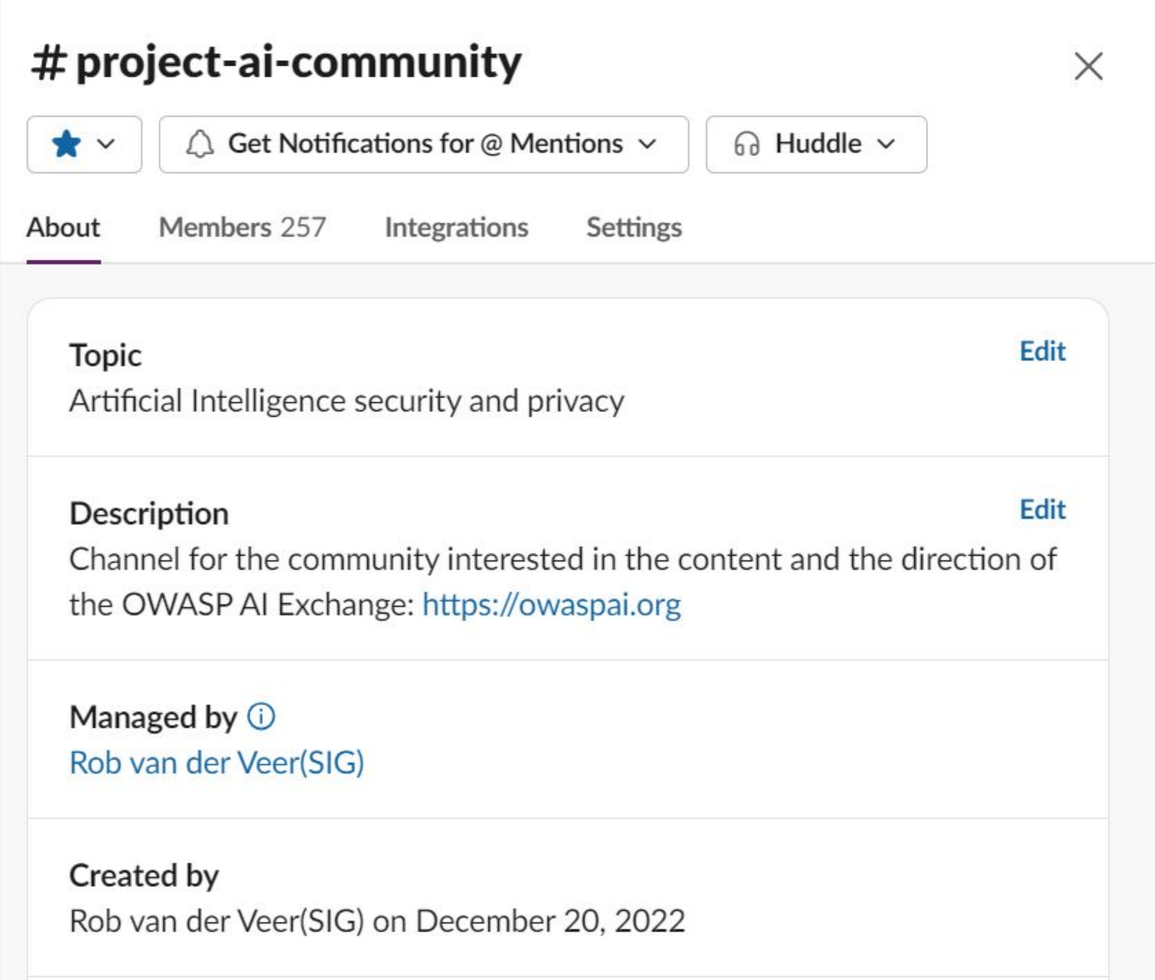
Follow us on LinkedIn

Stream the bi-weekly meetings (8)

OWASP AI Exchange

Get Involved as part of the Slack communication channel

(over 250 members)



project-ai-community

Topic
Artificial Intelligence security and privacy

Description
Channel for the community interested in the content and the direction of the OWASP AI Exchange: https://owaspai.org

Managed by
Rob van der Veer(SIG)

Created by
Rob van der Veer(SIG) on December 20, 2022

FAQ

What is our the stance on privacy?
https://owaspai.org/docs/ai_security_overview/#how-about-privacy

What is our the stance on copyright?
https://owaspai.org/docs/ai_security_overview/#how-about-copyright

How can I get associated?
https://owaspai.org/contribute/

What are all the frameworks that exist around GenAI?
https://owaspai.org/goto/references/

How does it complement OWASP LLM Top Ten?
The LLM is about the top 10 issues. **The Exchange is about all issues in all of AI**

Key Reference Links

- Bi-Weekly Meeting
- Contribute
- OWASP Slack Invite
- OWASP LLM top 10
- ENISA ML threats and countermeasures 2021
- MITRE ATLAS framework for AI threats
- NIST threat taxonomy
- ETSI SAI Problem statement Section 6
- Microsoft AI failure modes
- NIST
- NISTIR 8269 - A Taxonomy and Terminology of Adversarial Machine Learning
- OWASP ML top 10
- PLOT4ai threat library
- AVID AI Vulnerability database
- OECD AI Incidents Monitor (AIM)
- ENISA AI security standard discussion
- ENISA's multilayer AI security framework
- Alan Turing institute's AI standards hub
- Microsoft/MITRE tooling for ML teams
- Google's Secure AI Framework
- NIST AI Risk Management Framework 1.0
- ETSI GR SAI 002 V 1.1.1 Securing Artificial Intelligence (SAI) – Data Supply Chain Security
- ISO/IEC 20547-4 Big data security
- IEEE 2813 Big Data Business Security Risk Assessment
- BIML
- Media
- **OWASPAI.ORG**

OWASP AI Exchange (owaspai.org)

Thank You