

OWASP AI EXCHANGE

@SSCA Forum at MITRE

May 19 2026

Aruneesh Salhotra: Strategist & Technologist



Leadership & Global Authority

Founder & CEO of a Stealth Startup and CISO, Principal at SNM Consulting Inc. provides strategic direction for Security and AI.

- Project Co-Lead: OWASP AI Exchange
- Project Lead: OWASP AIBOM (AI Bill of Material)
- Forbes Technology Council Executive Member
- Angel Investor in Tech and Non-Tech
- Board Member, Advisory Board - GTM

terminal

CISSP, C-CISO, and G-CISO certified.

AWS * 5, Kubernetes * 2

Globally recognized for AI security policy impact.

campaign

RSA, Black Hat, OWASP, SANS, Harvard, ADDO, ISACA, Machines Can See, IAPP etc.
O'Reilly published author.

As an Angel Investor: My Investment Thesis

The AI innovation landscape is shifting from "Chatting" to "Acting," creating new categories of engineering, security complexities and liability.

account_ balance Actionable & Comprehensive

As an Angel Investor, I look for products that bridge the gap between "Boardroom mandates" and "Engineering reality".

Key Investment Pillars

verified

- Solves a defined, real-world problem
- Community support & rapid adoption
- Evolves with the changing landscape
- Inherent flexibility and cohesion

diversity_3 Multi-Persona & Strategic Fit

Resonates with CEOs, CFOs, and CISOs by solving real problems through a flexible, cohesive approach.

My Personal Journey with AI and OWASP AI Exchange

2022 - 2023

AI felt like **magic**—an autonomous brain.

Initially, use cases were limited to emails and summaries.

It wasn't yet seen as a serious enterprise tool.

Late 2023

Research Group

Started: I led 7 professionals to track the evolving AI landscape.

Monitored frontier models, regulations, incidents, investments and advancements.

2024 - Today

Joining OWASP:

Connected with Rob van der Veer via LinkedIn.

Witnessed the initiative grow from a post to a global authority in AI Security.

A Global Initiative for AI Security

The OWASP AI Exchange is a global think tank and open-source initiative that brings together experts in AI and cybersecurity.

OUR MISSION

To bring clarity in AI security by connecting practitioners, researchers, industry, and policymakers.

groups

Practitioners

science

Researchers

business

Industry

gavel

Policymakers

The comprehensive foundational technical framework for AI security

driver of standards and understanding

300 pages in one document at owaspai.org

Co-developed with **SANS**

Constantly updated through global open source model (160+ selected **experts**)

Aligned and linked with other **initiatives** e.g. SANS controls, MITRE

Forms the **basis** of ISO/IEC and AI Act standards

Official **liaison** partnership ISO & CEN/CENELEC

All content **free** of copyright & attribution

Covers **all** of AI, all threats and controls

 OWASP
flagship project

Scope & Responsibilities



- Develop a **comprehensive framework for AI threats, risks, mitigations, and controls.**
- Develop a map integrating **AI regulatory and privacy regulations.**
- Establish a **common taxonomy** and glossary for AI security.
- Provide **practical 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.
- An **integrated and cohesive resource** instead of a stream of isolated documents.
- **Open-source the international standardization** of AI security, fostering global collaboration and transparency

EcoSystem and Standards Bodies

SANS



Free of Copyright



Global Security
and AI Community



NIST

MITRE

RELEVANCE/
RECOGNITION



Growing Community of Industry Leading AI Security Experts and Companies

160+

Contributors

66+

Companies

1000+

Slack Joiners

9200+

LI Followers*

Accenture

AIFT

AIShield

BMO Financial Group

Casco

Dekra

Amazon Web Services

Ancharski Advisory

Capital One

Capitol Technology

University

Carl ZEISS AG

Centil

Cycubix

Preamble

Coalfire

Code4Thought

Cyber Dynamo Pty Ltd

DataBricks

Dave

Tata

Endava Adriatic

Fraunhofer institute

Google

Normalyze

Hiddenlayer

ISACA Silicon Valley

L&T Technology

Services, India

Left Crossword

Cybersecurity

Mindgard & Lancaster

MP Materials

Uni

Nokia

Noma

Noma security, Israel

Oak Ridge National

Laboratory, US

Palo Alto Networks

Paycom

Prediction Guard

Raiffeisen Bank International AG

Riskology

Riskview

Rockcyber

IDEXX

SANS Institute

SC Ventures by Standard Chartered

Schneider Electronic

Sentra

SIG

SNM Consulting Inc.

Splunk

St. Fox Consulting

Straiker

Trail of bits

Universidad Isabel

VE3

Vodafone Group

Workforce tech LLC

Zalando

Recent and upcoming Conferences/Webinars

RECENT

- SANS AI Summit: two workshops. Maryam + Aruneesh, Rob + Shoshana

UPCOMING

- Rob misc. webinars and misc. talks: Infosecurity (Apr 28), University lustrum talk(May 9), AI4SEC4AI conference panel (June 11)
- Aruneesh at NIST session on the Exchange in DC. At MITRE (May 19,20)
- Aruneesh at Developer Week in NYC (Jun 9,10)
- Aruneesh at IAPP EU 2026 (Jun 1-4)
- Aruneesh at IIT (Gandhinagar, India) (Jun 5)
- Behnaz, Yuvi at BSides Denmark (Jun 20) - Training
- Aruneesh at UN Open Source Week (Jun 22-26)
- OWASP Global Appsec Vienna June 25/26
 - <https://owasp.glueup.com/event/owasp-global-appsec-eu-2026-vienna-austria-162243/>
 - Rob will do training there
 - Mike Novack - presents
 - AI Exchange project showcase session!
 - Huddle there

MORE

- <https://www.cognitivesecurityinstitute.org/cognitive-security-conference> (reported by Allen)
- BlackHat 2026
- BSides Chicago
- Ai4 - Vegas
- Andrew Ng (AI-DEV)

COLLABORATIONS

OWASP & CSA: Joint Initiative for Agentic AI Red Teaming

Developing the "Agentic AI Red Teaming Guide"

Addressing unique challenges in systems with planning, reasoning, and orchestration capabilities.

- **Practical Steps:** Uncover vulnerabilities in agent authorization and goal manipulation.
- **Community-Driven:** Insights from 40+ contributors across AI and Cyber Security.
- **Actionable Assessment:** Helping red teams secure complex agentic applications.

Bi-weekly deep-dive sessions starting March 10, 2025.



OWASP AI EXCHANGE

.....
CLOUD SECURITY ALLIANCE



July 2025: SANS and OWASP AI Exchange established a formal partnership



Done and doing:

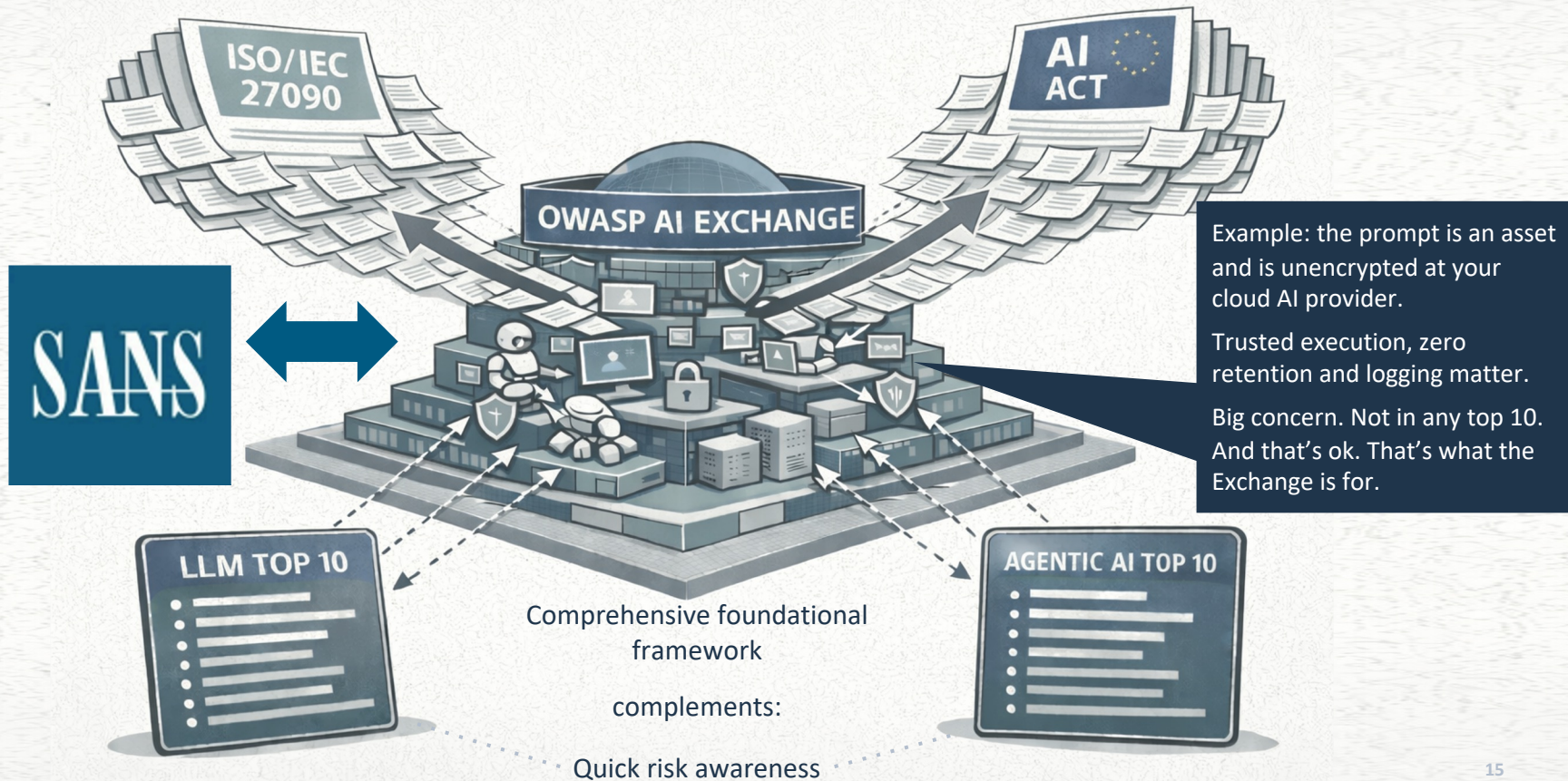
- SANS contributing to the Exchange
(e.g. work for the AI Act)
- Co-develop unified controls, making this content part of SANS' strategic direction
- Organize events
(e.g. Policy forum)

SANS and OWASP Join Forces to Standardize AI Security Controls

With attacks escalating and no clear standards in place, two of cybersecurity's most trusted communities are joining forces to give defenders the tools they've been missing



OWASP AI Exchange – effectively the global standard in AI security

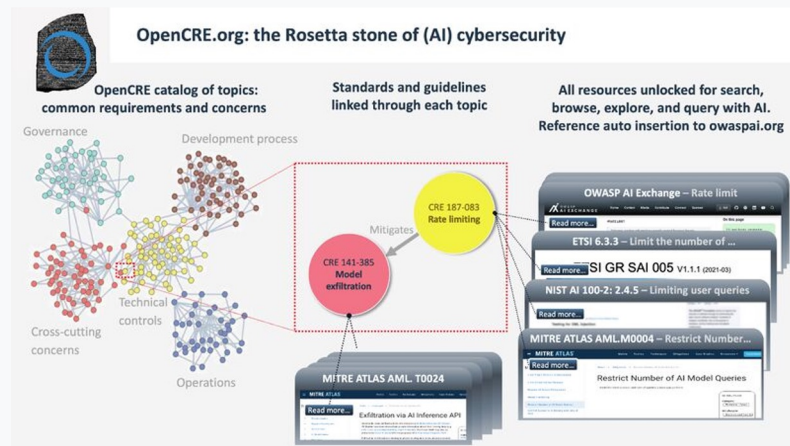


OWASP AI Exchange & OpenCRE: Harmonizing AI Security

Strategic Mapping & Integration

Linking AI-specific threats and controls to the global catalog of security requirements.

- **Unified Framework:** OpenCRE acts as the central mapping hub for AI Exchange content.
- **Automated References:** Real-time integration of standards (NIST, ISO, CSA) via the OpenCRE platform.
- **Harmonization:** Ensuring AI security is not isolated but connected to foundational security principles.



OWASP AI EXCHANGE

...

OPENCRE.ORG

MOSAIC: A Breakthrough in AI Security Coordination

MOSAIC: Multi-Organization Secure AI Coordination

On April 21, 2026, leading global initiatives united to harmonize AI security standards.

- **Collective Action:** Moving from fragmented efforts to a unified, multi-organization coordination model.
- **The Mission:** Establish clear, consistent guidelines to mitigate escalating AI-specific risks.
- **Strategic Goal:** Transform a complex landscape into actionable, global security requirements.



5 STEPS TO AI Security



5 STEPS FOR AI SECURITY



OWASP
AI EXCHANGE

Setup an AI management system (ISO/IEC 42001)

Inventory of AI applications

Responsibilities

Legal counsel

Awareness/expertise

Compliance

Policy

Stakeholders

Impact assessment: legal? ethical? transparency? bias? expertise access?



Think before you start, for each idea

Is the application area **legal**?

Do you really **need AI** to solve the problem?

Do you have access to the right **expertise** (e.g. data scientists)?

Are you allowed to **use the data** for this purpose – given privacy principles?

Can you manage risks of harm by human **oversight**

Can you provide the legally and ethically required **transparency** into how your AI works?

Can you prevent **unwanted bias** sufficiently?

2 - UNDERSTAND – use the AI Exchange

200 pages material at
owaspai.org

AI threat model
donated by Software
Improvement Group

129 selected global experts:
Academia, start ups, data
scientists, legal, vendors

Alignment with LLM top 10, NIST,
MITRE, ENISA, CISA, CSA, Alan Turing
Institute, AISI, SANS – partnership

78 pages
contributed to
ISO/IEC 27090

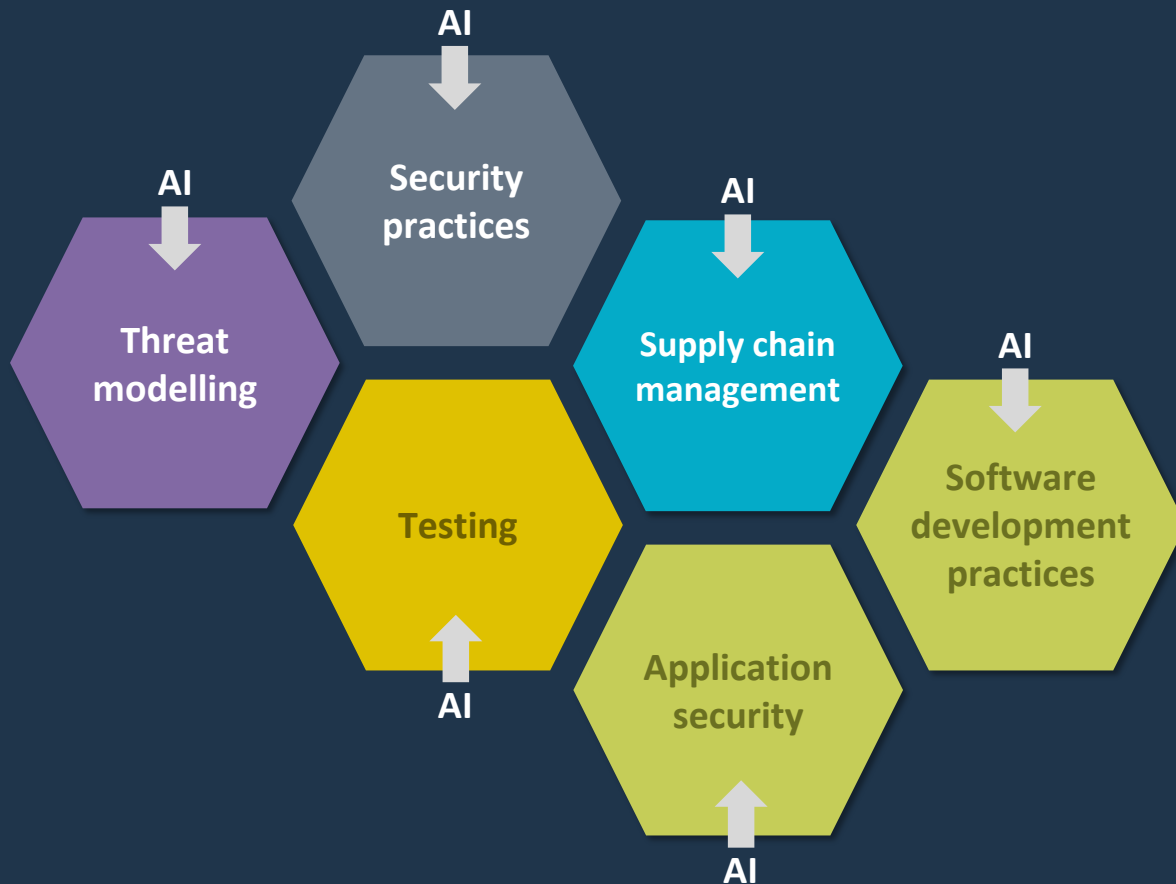
Official liaison
partnership
CEN/CENELEC

Free of copyright
and attribution

Effectively open sourced
AI standardization

70 pages contributed to
AI Act draft – ongoing work

3 – ADAPT your existing practices



**KEEP
CALM**

AI is just
software

4 – REDUCE the impact of manipulated AI

Minimizing or obfuscating sensitive data and limiting the effect of unwanted AI behaviour.



WORKSTREAMS

Strategic, Technical, and Specialized Project Initiatives

Our mission is to be the global go-to resource for security and privacy practitioners working with AI and data-centric systems—bringing alignment and encouraging collaboration across initiatives.

Priority Content Workstreams

- **Agentic AI:** Emerging security challenges of autonomous multi-agent systems.
- **Harmonization:** Mapping to NIST, MITRE ATLAS, and ETSI.
- **Red Teaming:** Evaluating tools for Predictive and Generative AI.
- **EU AI Act Flow Back:** Integrating research into core documentation.
- **Step-by-Step Guide:** Structured path for implementation.

Strategic & Operational

- **Standards Alignment:** Drive global clarity (Rob & Aruneesh).
- **Promotion & Outreach:** Marketing and visibility (Aruneesh).
- **Sponsorship:** Fundraising and sustainability (Aruneesh).

Specialized Project Groups

Risk Analysis: Formal assessment processes.

Education: Training for data scientists.

Legislation: Tracking global laws & AI Act reqs.

Mathematical: Models for AI robustness.



The Navigator as your primary teaching artefact

owaspai.org/go/navigator/



Threat model based on Software Improvement Group AI framework

Interactive PDF with deep links

Organizes threats by lifecycle

Lists all controls

Categorizes controls



AI security threats matrix

owaspai.org/go/aiecuritymatrix/

AI-specific?	Lifecycle	Attack surface	Threat/Risk category	Asset	Impacted	Potential harm
AI	Runtime	Input attack	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 employee morale, incorrect strategic decisions, liability issues, personal damage and safety issues.
			Indirect prompt injection			
			Evasion			
		Break into deployed model	Direct runtime model poisoning			
	Development	Break into augmentation data	Augmentation data manipulation			
		Break into train/finetune data	Direct development-time model poisoning			
			Data poisoning			
		Compromise supply chain	Supply-chain model poisoning			
			Supply chain data poisoning			
	Runtime	Input attack	Disclosure in output	Augmentation data	Confidentiality	Fines, legal fees, remediation effort, reputation damage, operational disruption, etc.
		Break into deployed system	Direct augmentation data leak			
	Runtime	Input attack	Disclosure in output	Training data	Confidentiality	Fines, legal fees, remediation effort, reputation damage, operational disruption, etc.
			Model inversion / Membership inference			
	Development	Break into training data	Direct training data leak			
Conventional	Runtime	Input attack	Model exfiltration	Model parameters	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 input attacks.
		Break into deployed model	Direct runtime model leak			
	Development	Break into model data	Direct development-time model leak			
	Runtime	Input attack	AI resource exhaustion	Model behaviour	Availability	Excessive running costs and/or unavailable AI system, leading to business continuity issues, or safety problems.
	Runtime	All IT	Input data leak	Model input data	Confidentiality	Sensitive data in model input leaks. E.g. a prompt with a sensitive question, enhanced with retrieved company secrets.
	Runtime	All IT	Output contains conventional injection	Any asset	C, I, A	Injection attack (from model output) causes harm.
	Runtime	All IT	Generic runtime security threats	Any asset	C, I, A	Generic runtime attack causes harm (includes social engineering/phishing).
	Development	All IT	Conventional development-environment and supply-chain threats	Any asset	C, I, A	Generic development or supply-chain attack causes harm (e.g. vulnerability in a component).



AI-specific?	Lifecycle	Attack surface	Threat/Risk category	Asset	Impacted	Potential harm
AI	Runtime	Input attack	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 employee morale, incorrect strategic decisions, liability issues, personal damage and safety issues.
			Indirect prompt injection			
			Evasion			
		Break into deployed model	Direct runtime model poisoning			
	Development	Break into augmentation data	Augmentation data manipulation	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 employee morale, incorrect strategic decisions, liability issues, personal damage and safety issues.
		Break into train/finetune data	Direct development-time model poisoning			
		Compromise supply chain	Data poisoning			
			Supply-chain model poisoning			
	Runtime	Input attack	Disclosure in output	Augmentation data	Confidentiality	Fines, legal fees, remediation effort, reputation damage, operational disruption, etc.
		Break into deployed system	Direct augmentation data leak			
	Runtime	Input attack	Disclosure in output	Training data	Confidentiality	Fines, legal fees, remediation effort, reputation damage, operational disruption, etc.
			Model inversion / Membership inference			
	Development	Break into training data	Direct training data leak			
Conventional	Runtime	Input attack	Model exfiltration	Model parameters	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 input attacks.
		Break into deployed model	Direct runtime model leak			
	Development	Break into model data	Direct development-time model leak			
	Runtime	Input attack	AI resource exhaustion	Model behaviour	Availability	Excessive running costs and/or unavailable AI system, leading to business continuity issues, or safety problems.
	Runtime	All IT	Input data leak	Model input data	Confidentiality	Sensitive data in model input leaks. E.g. a prompt with a sensitive question, enhanced with retrieved company secrets.
	Runtime	All IT	Output contains conventional injection	Any asset	C, I, A	Injection attack (from model output) causes harm.
	Development	All IT	Conventional development-environment and supply-chain threats	Any asset	C, I, A	Generic development or supply-chain attack causes harm (e.g. vulnerability in a component).



Linking threats to controls – the periodic table of AI security

owaspai.org/go/periodictable/

Asset & Impact	Attack surface with lifecycle	Threat/Risk category	Controls
Model behaviour integrity	Runtime - Model use (provide input, read output)	Direct prompt injection	Limit unwanted behavior , Input validation , further controls implemented in the model itself
	Runtime - Break into deployed model	Indirect prompt injection	Limit unwanted behavior , Input validation , Input segregation
	Development - Engineering environment	Evasion (e.g. adversarial examples)	Limit unwanted behavior , Monitor , rate limit , model access control plus: Detect odd input , detect adversarial input , evasion robust model , train adversarial , input distortion , adversarial robust distillation
	Development - Supply chain	Model poisoning	Limit unwanted behavior , Runtime model integrity , runtime model input/output integrity
	Training data Confidentiality	Development-environment model poisoning	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
	Runtime - Model use	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
	Development - Engineering environment	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
	Model confidentiality	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
	Model behaviour availability	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
	Model input data Confidentiality	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
Any asset, CIA	Runtime - All IT	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
Any asset, CIA	Runtime - All IT	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
Any asset, CIA	Runtime - All IT	Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble

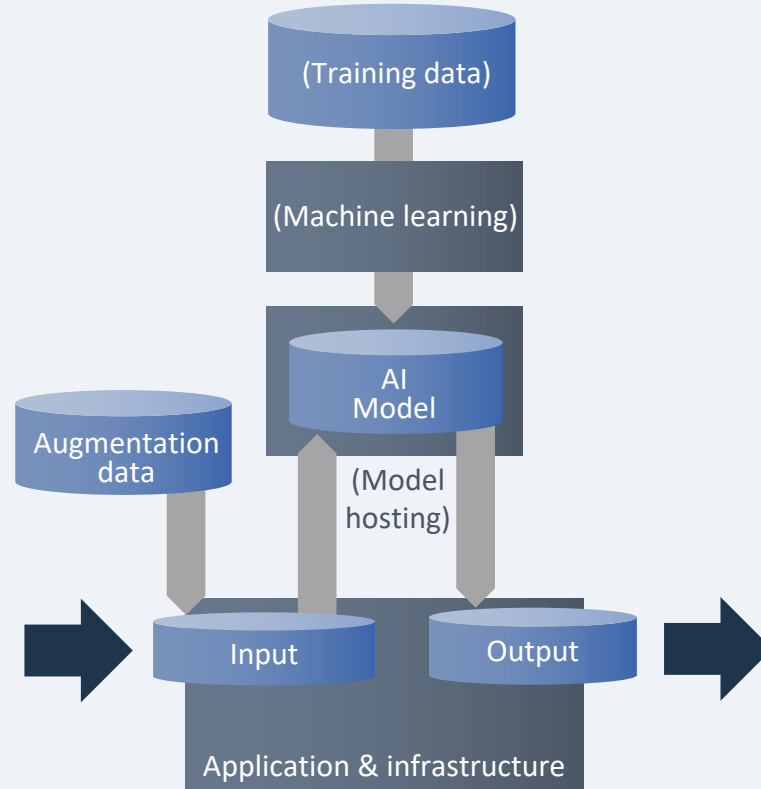
Asset & Impact	Attack surface with lifecycle	Threat/Risk category	Controls
Model behaviour Integrity	Runtime -Model use (provide input/ read output)	Direct prompt injection	Limit unwanted behavior , Input validation , further controls implemented in the model itself
		Indirect prompt injection	Limit unwanted behavior , Input validation , Input segregation
		Evasion (e.g. adversarial examples)	Limit unwanted behavior , Monitor , rate limit , model access control plus: Detect odd input , detect adversarial input , evasion robust model , train adversarial , input distortion , adversarial robust distillation
	Runtime - Break into deployed model	Model poisoning runtime (reprogramming)	Limit unwanted behavior , Runtime model integrity , runtime model input/output integrity
Development - Engineering environment	Development - Engineering environment	Development-environment model poisoning	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble
		Model input data Confidentiality	Limit unwanted behavior , Development environment security , data segregation , federated learning , supply chain management plus: model ensemble



// AI SECURITY THREATS

Basic AI system

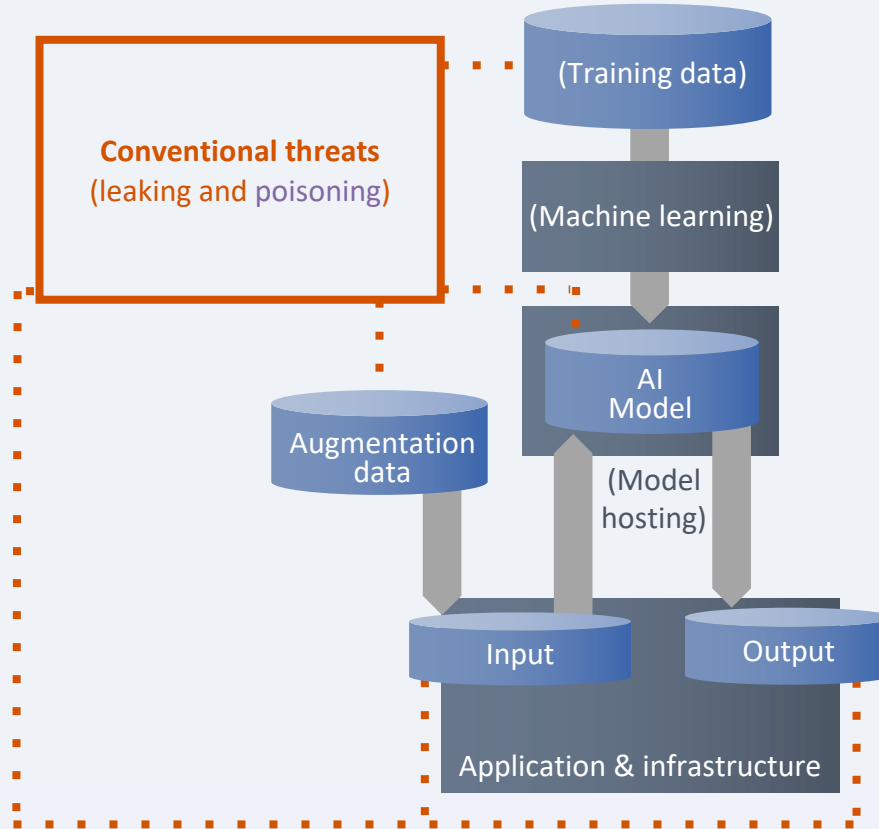
owaspai.org/go/essentials/





Conventional threats

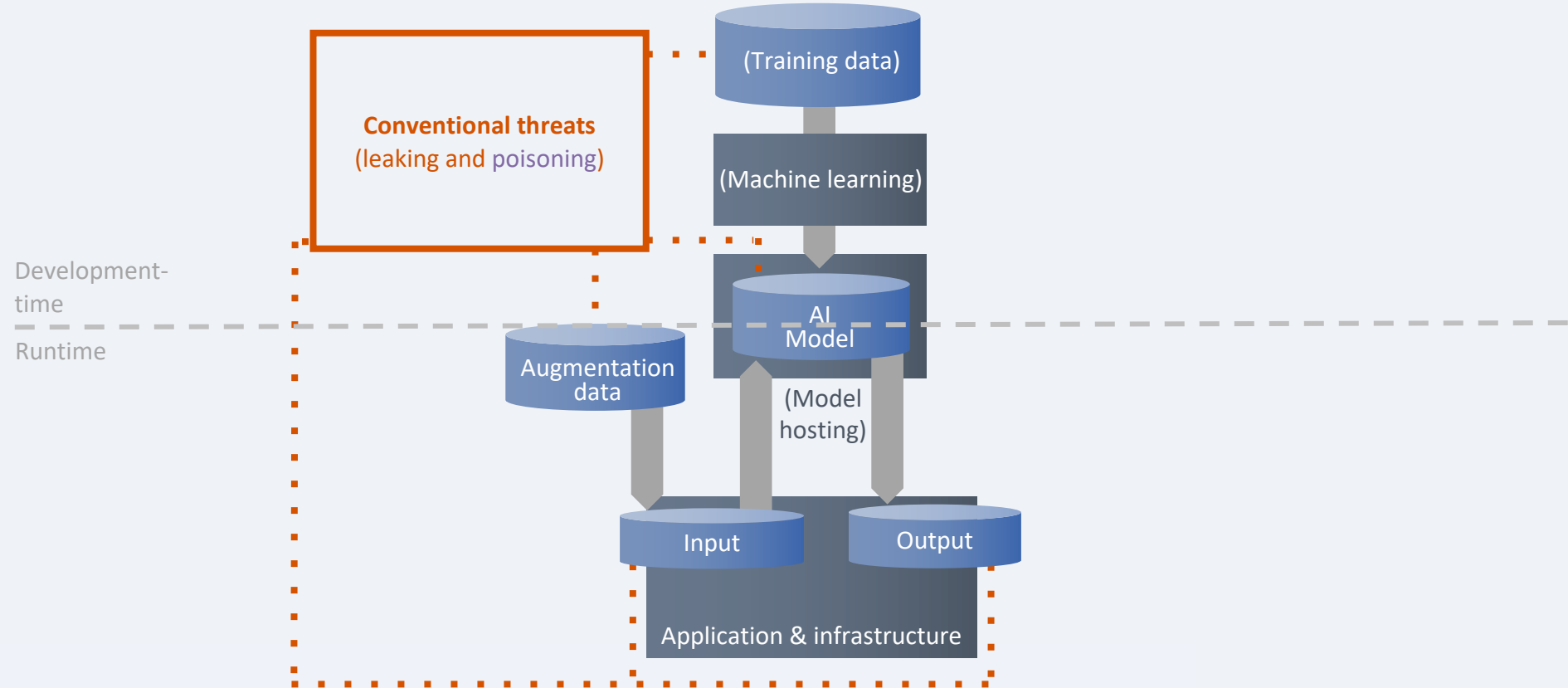
THREATS





Conventional threats – development-time and runtime

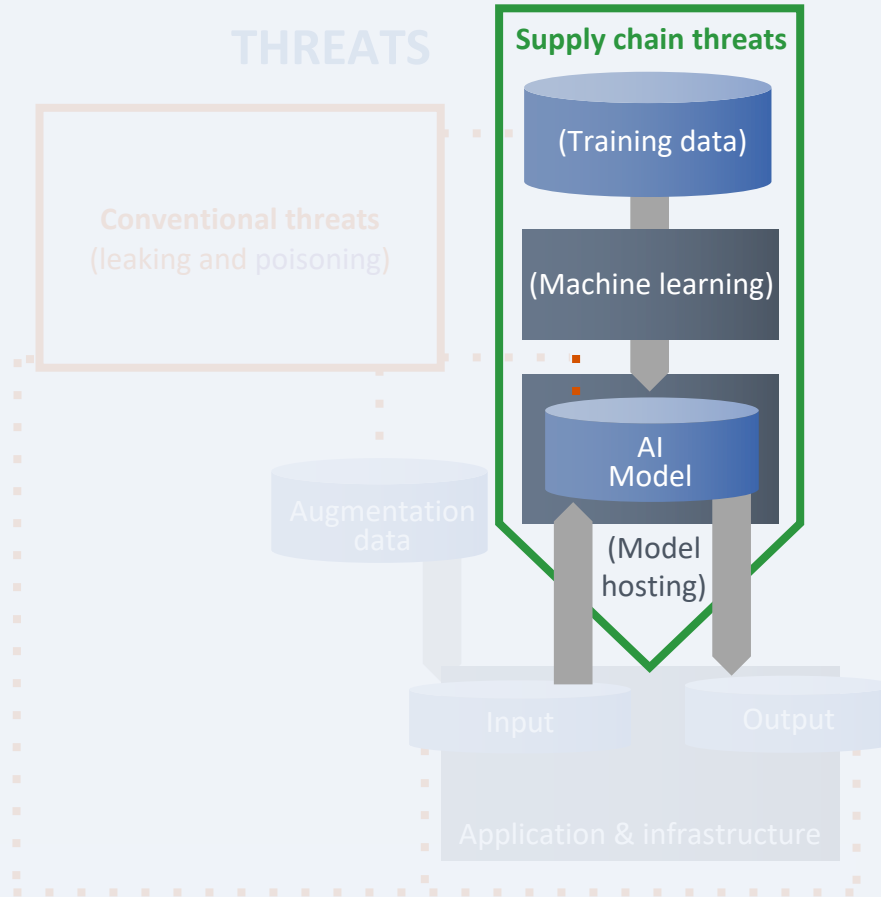
THREATS





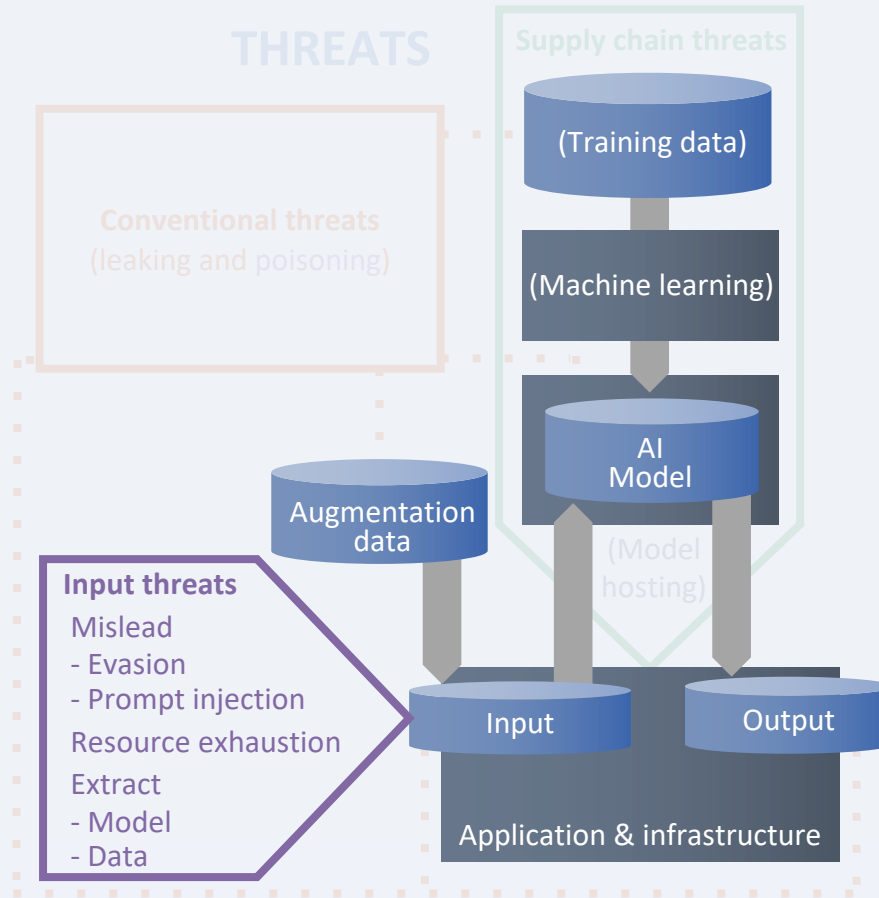
// AI SECURITY THREATS

Supply chain threats



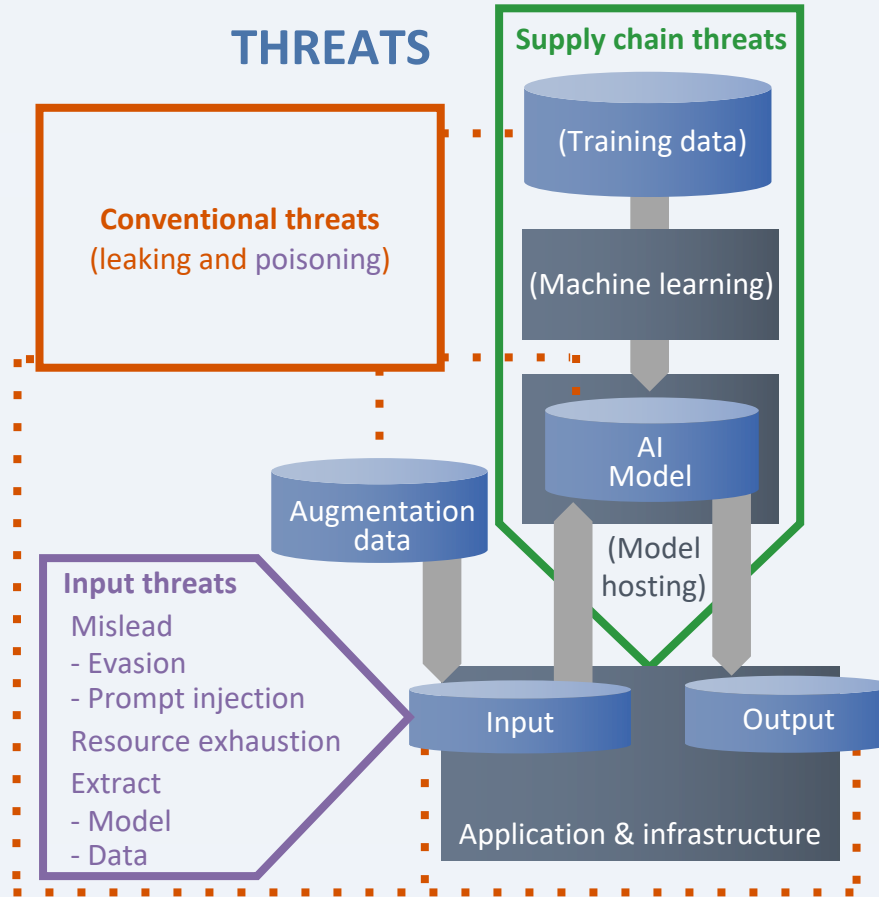


Input threats



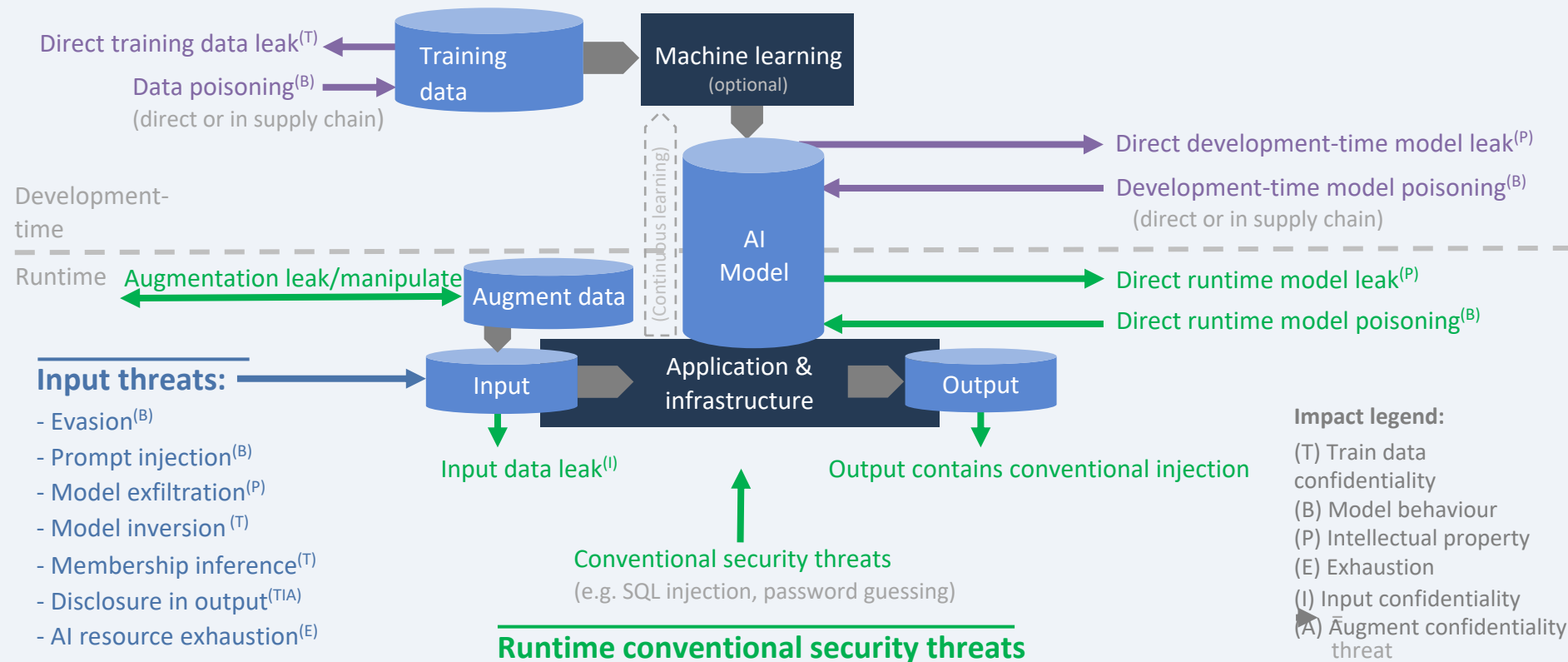


All threats



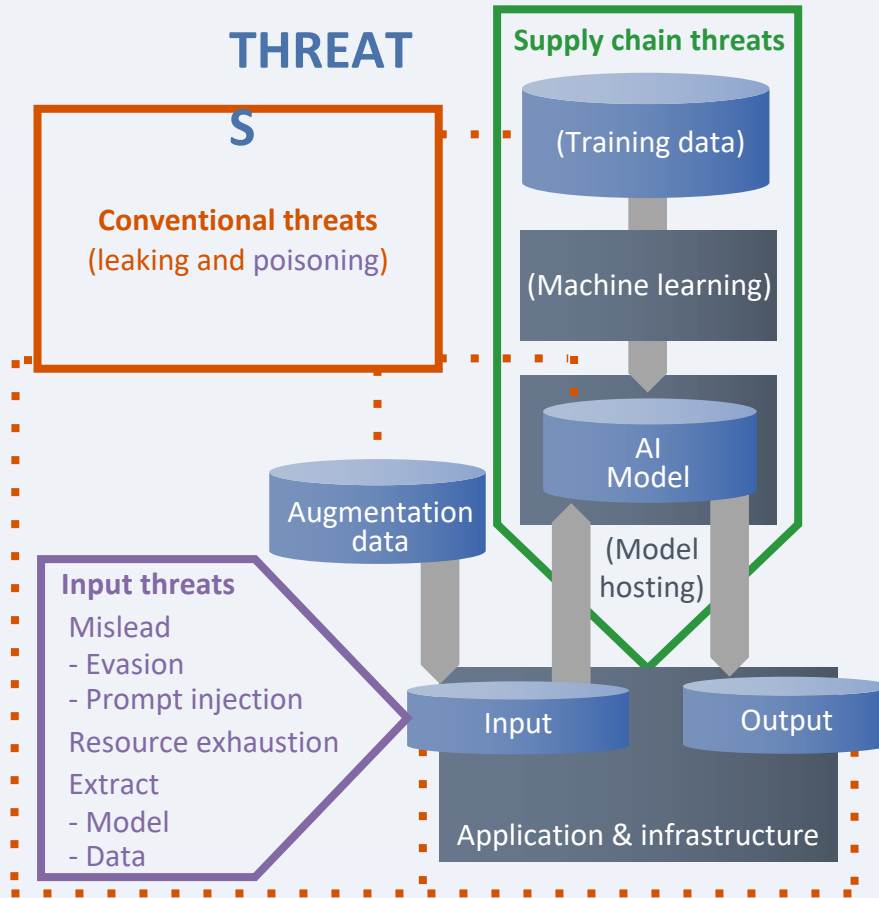


Development-time threats



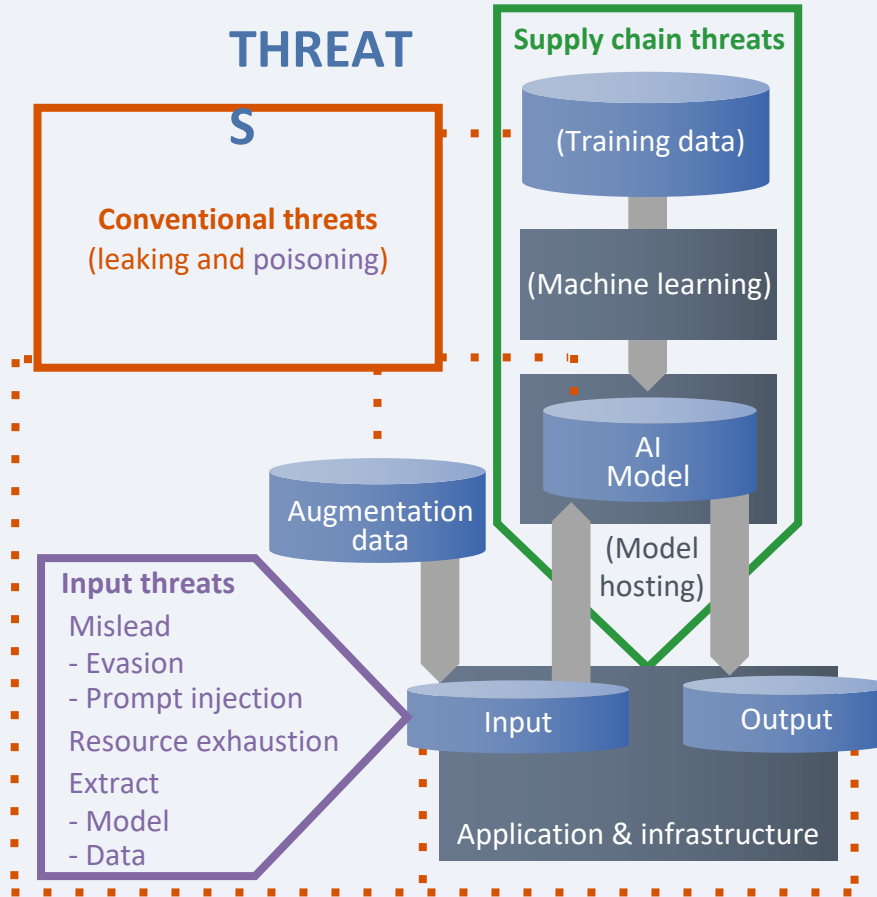


Controls



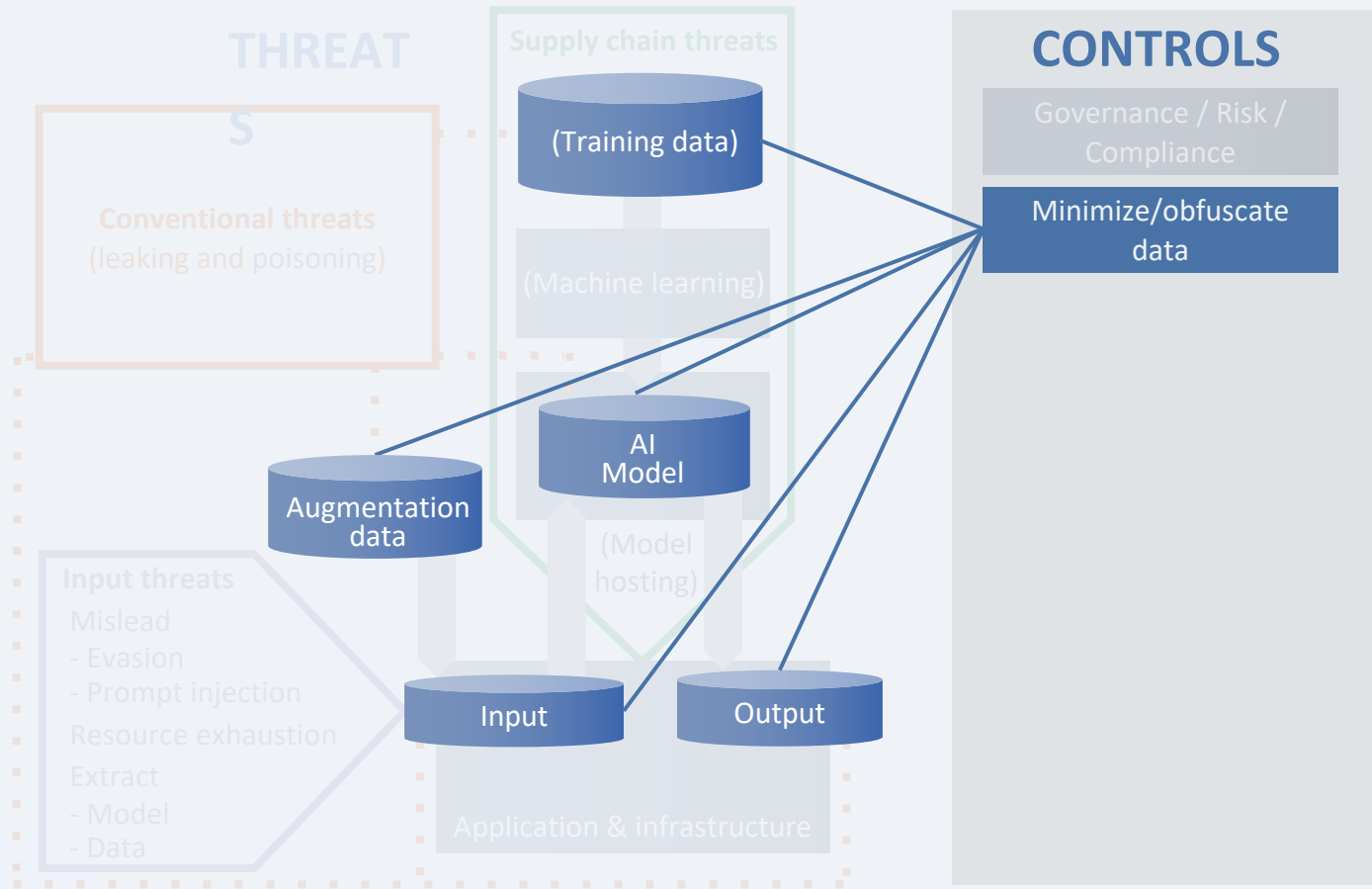


Controls - governance



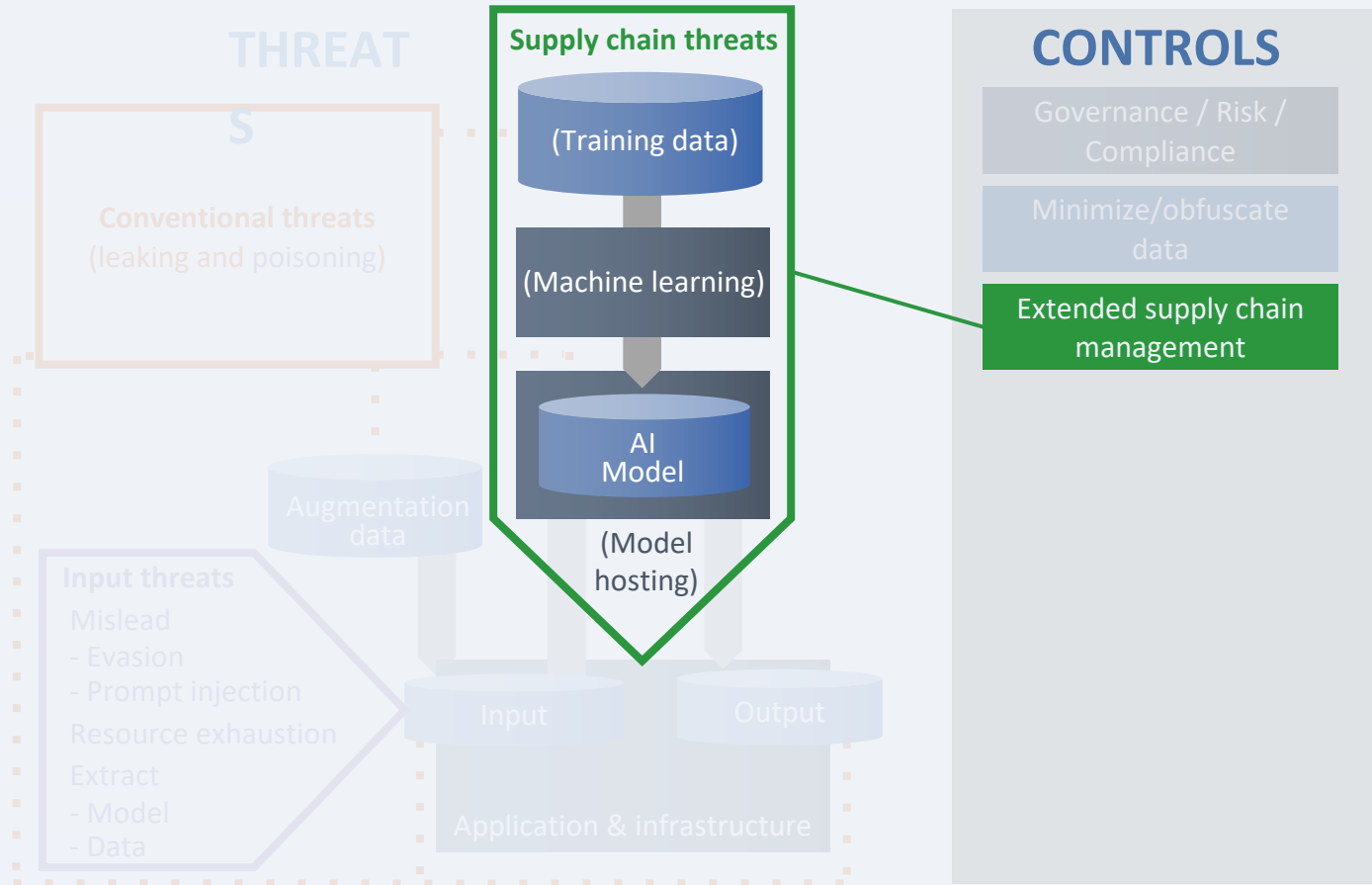


Controls – minimize/obfuscate data



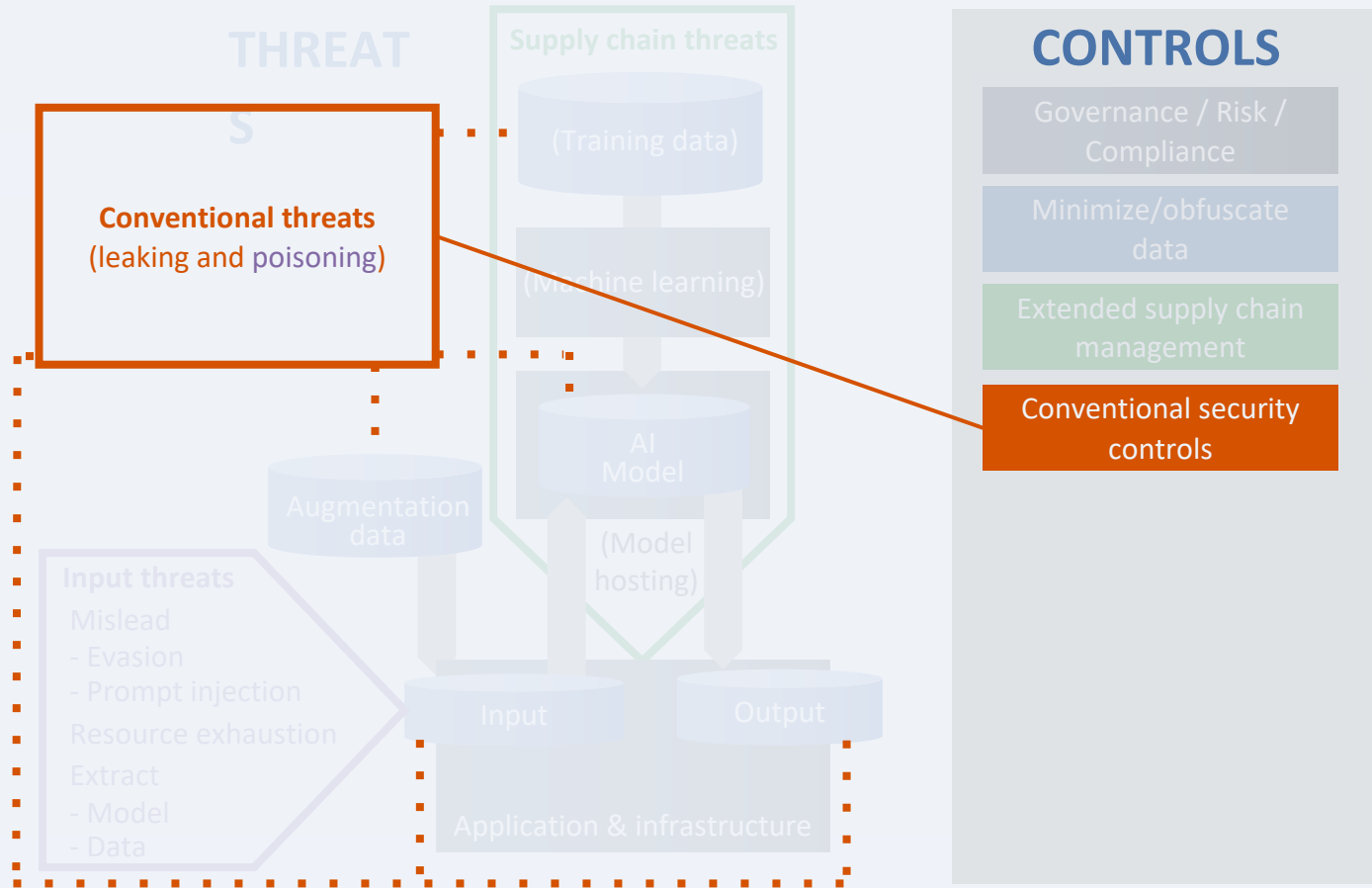


Controls – supply chain management



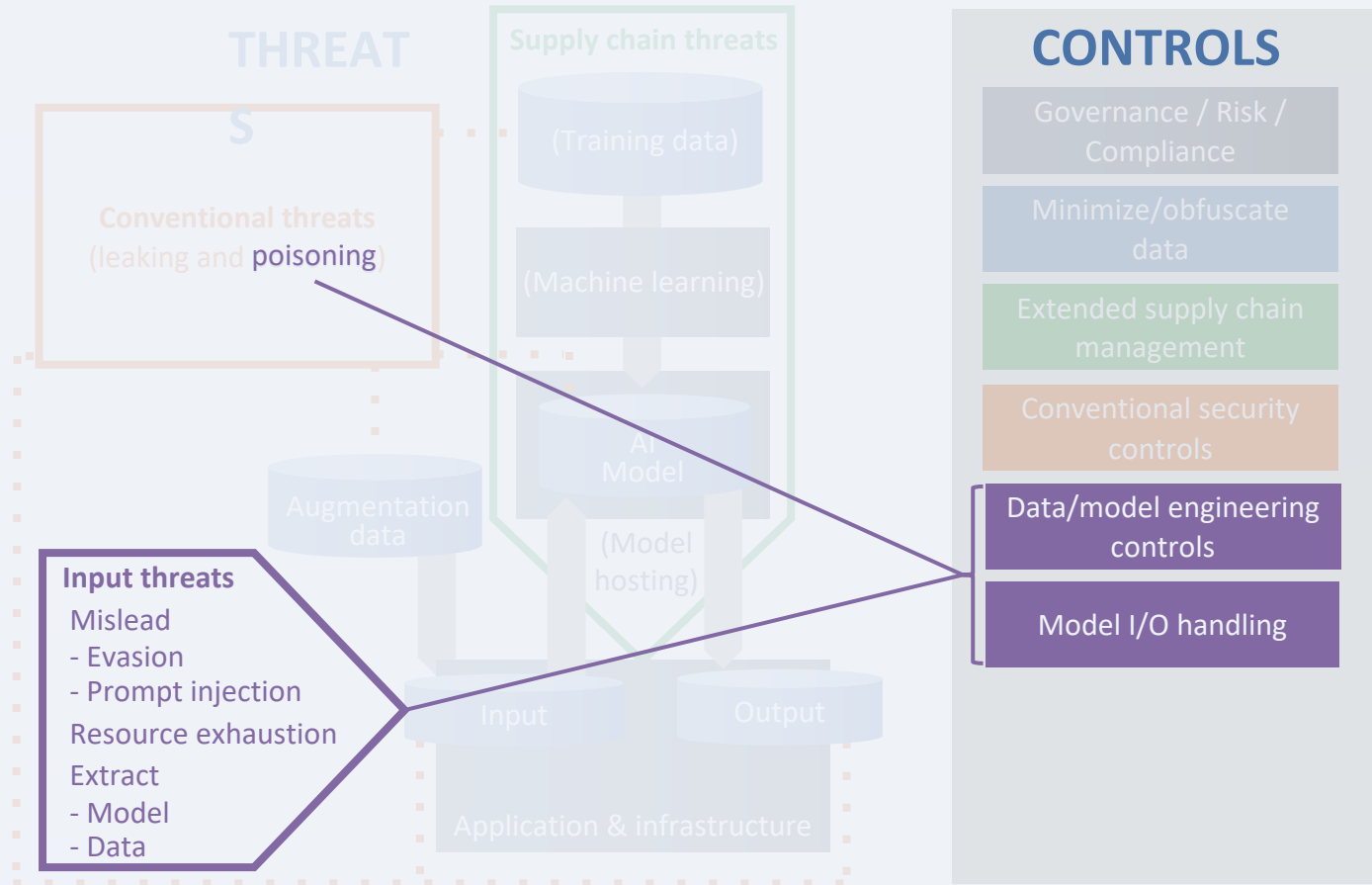


Controls - conventional



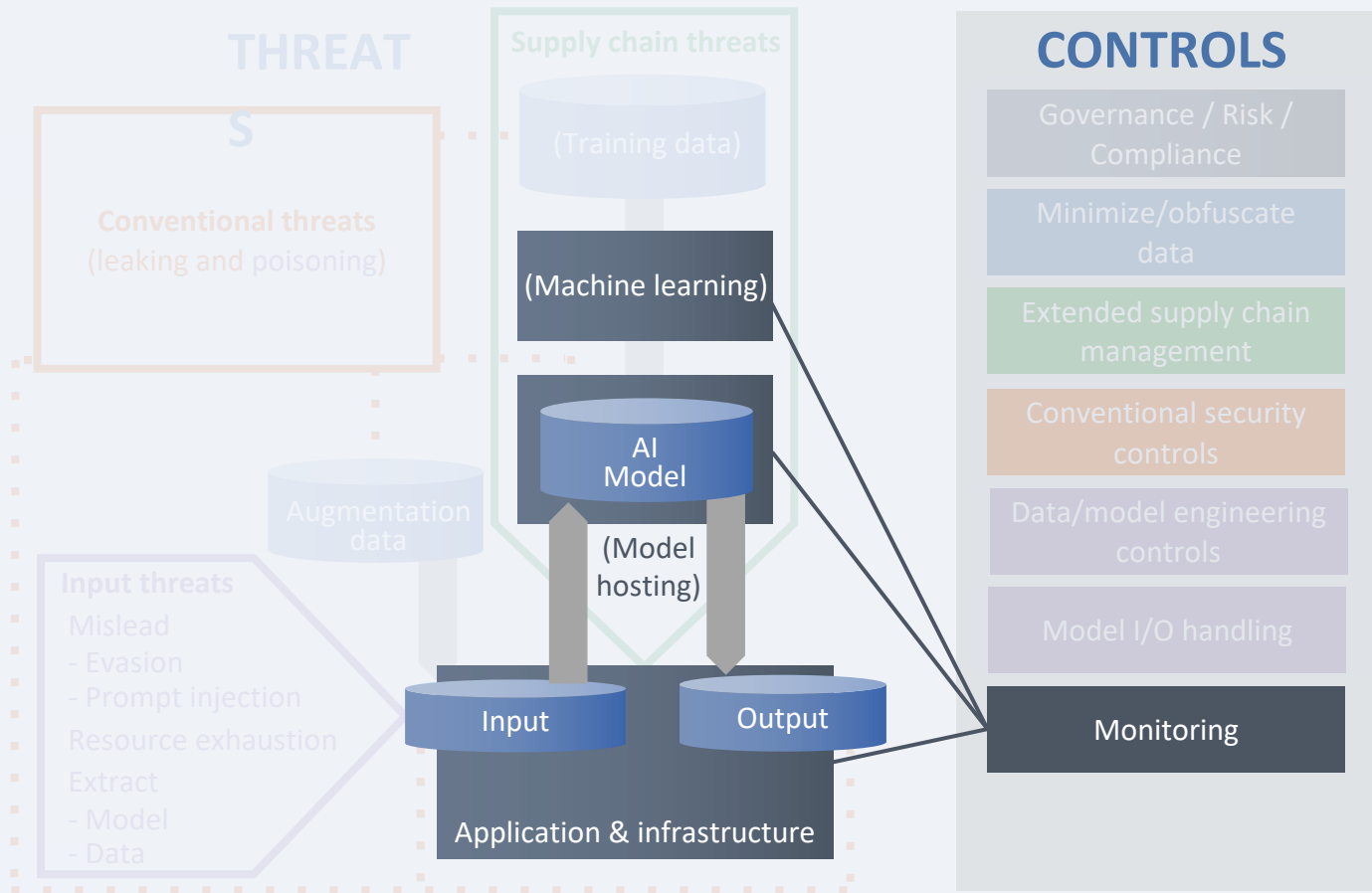


Controls – AI engineer controls



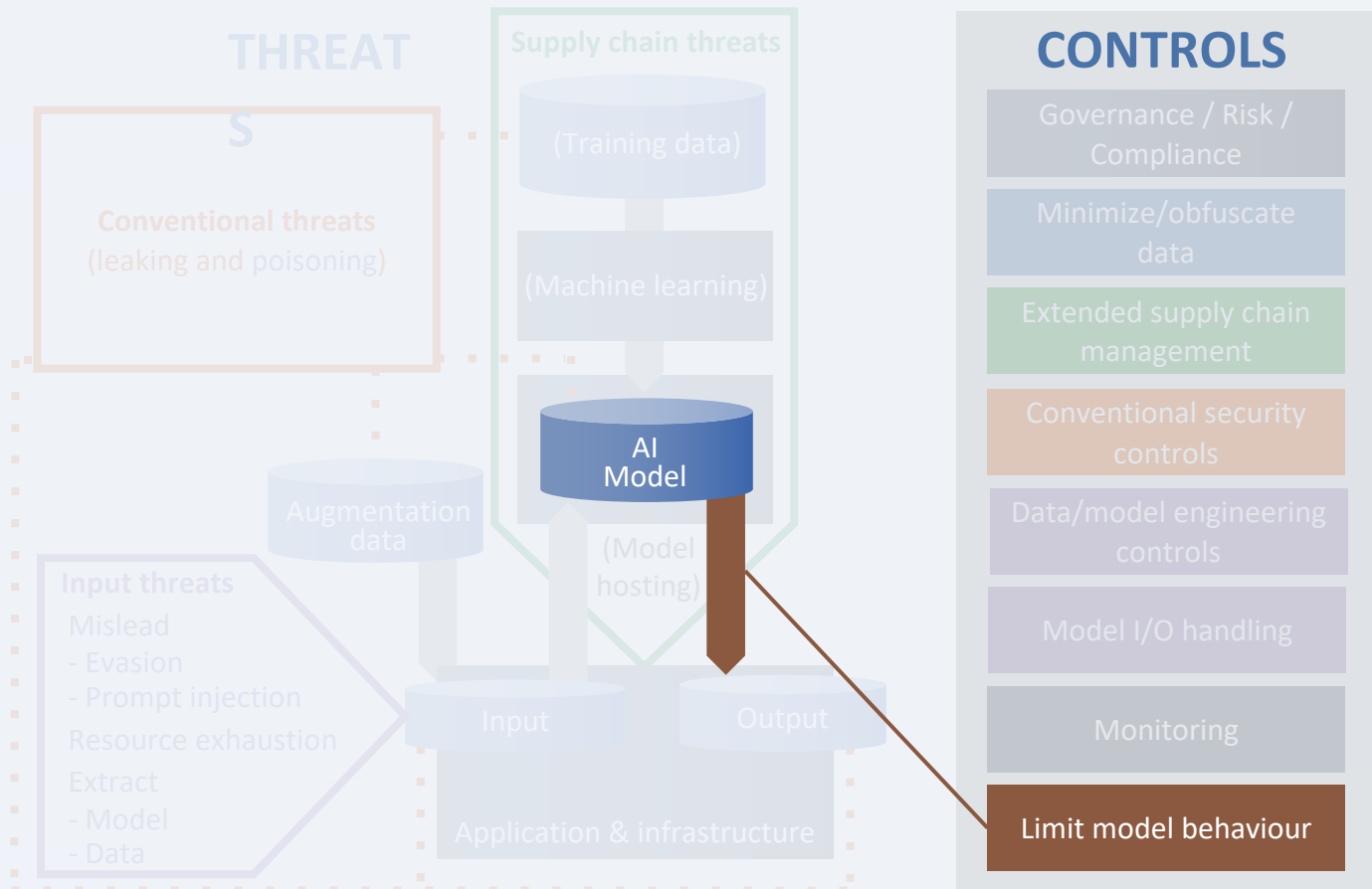


Controls - monitoring





Controls – limit behaviour





How to limit impact – blast radius control

Minimize/obfuscate data

Minimize data in rest and in transit

Obfuscate data where possible

Retain data as **brief** as possible

Minimize **technical details** (e.g. output but also: publications)

Ensemble models, federative learning

Limit model behaviour

Arrange human **oversight**

Detection - automated oversight

Assign minimal **privileges** to model actions (e.g. Agentic AI)

Be **transparent** / explain

Arrange proper correctness **testing**, continuous validation

Limiting unwanted bias testing





From controls to steps

CONTROLS

Governance / Risk /
Compliance

Minimize/obfuscate
data

Extended supply chain
management

Conventional security
controls

Data/model engineering
controls

Model I/O handling

Monitoring

Limit model behaviour

STEPS

1. **Organize the control of AI**
2. **Teach** AI engineers obfuscation strategies and minimization importance
3. Extend **supply management** to data, models & cloud AI
4. Add AI Assets and risks to ISMS **repository**
5. **Teach** AI engineers devsecops
6. **Teach** AI engineers AI security controls for model engineering and runtime + inform other teams on this
7. **Extend monitoring to observe AI-attack behaviour**
8. Implement model guardrails, oversight, least privilege



From steps to organizing

owaspai.org/go/organize/

CONTROLS

Governance / Risk / Compliance

Minimize/obfuscate data

Extended supply chain management

Conventional security controls

Data/model engineering controls

Model I/O handling

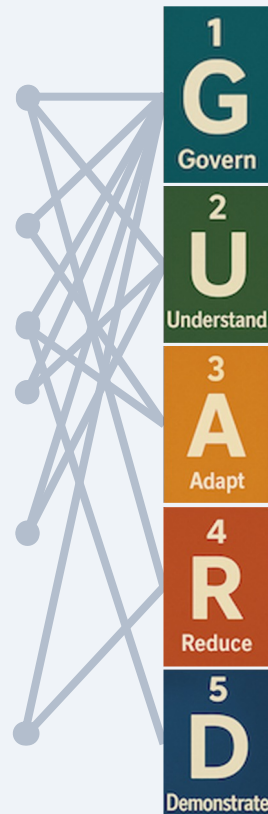
Monitoring

Limit model behaviour

STEPS

1. **Organize the control of AI**
2. **Teach** AI engineers obfuscation strategies and minimization importance
3. **Extend supply management** to data, models & cloud AI
4. **Add AI Assets and risks to ISMS repository**
5. **Teach** AI engineers devsecops
6. **Teach** AI engineers AI security controls for model engineering and runtime + inform other teams on this
7. **Extend monitoring to observe AI-attack behaviour**
8. **Implement model guardrails, oversight, least privilege**

ORGANIZE





// AI SECURITY CONTROLS

Controls grouped by goals



MANAGE

Governance / Risk /
Compliance

Extended supply chain
management

Conventional security
controls



RESILIENT MODELS

Data/model engineering
controls



WATCH

Model I/O handling

Monitoring



LIMIT

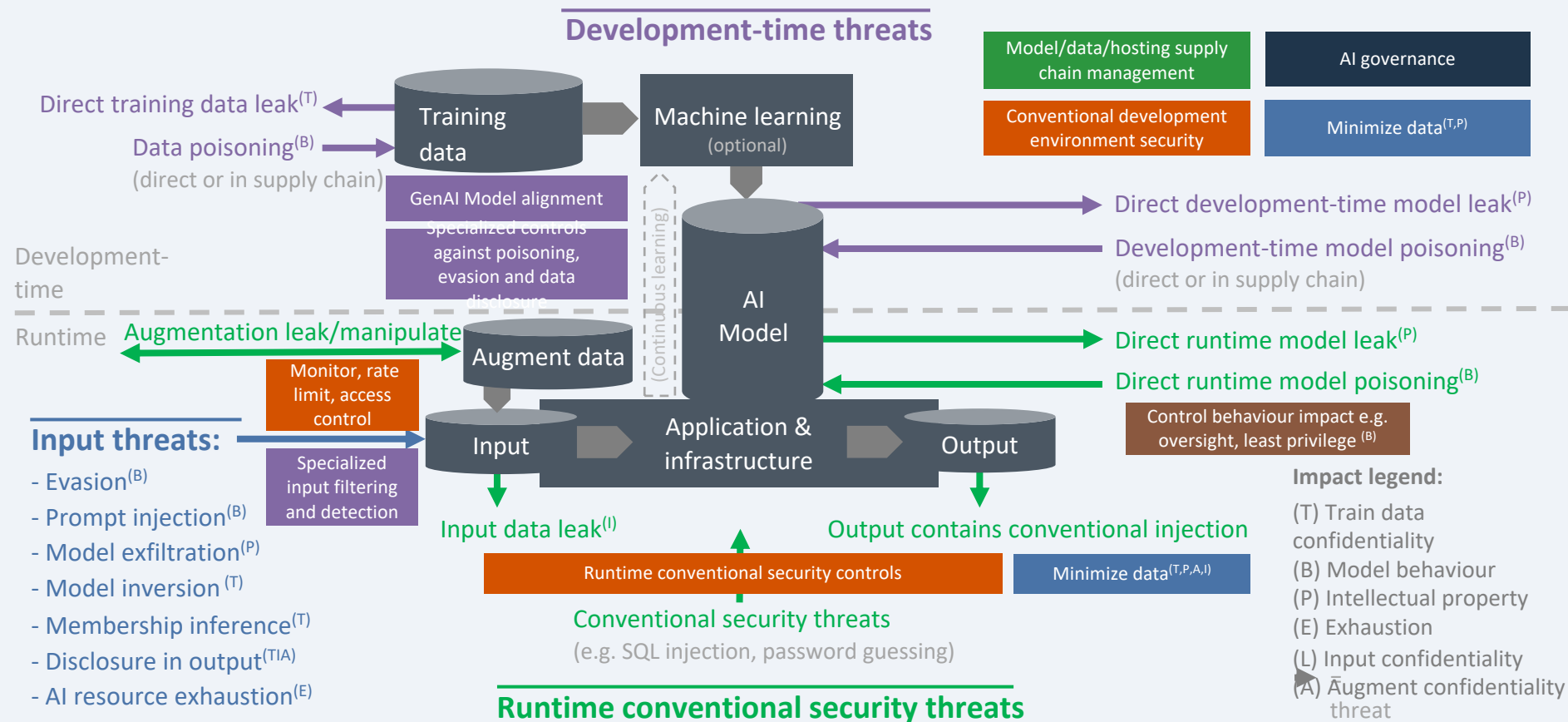
Minimize/obfuscate
data

Limit model behaviour



AI Exchange core threat model and control categories

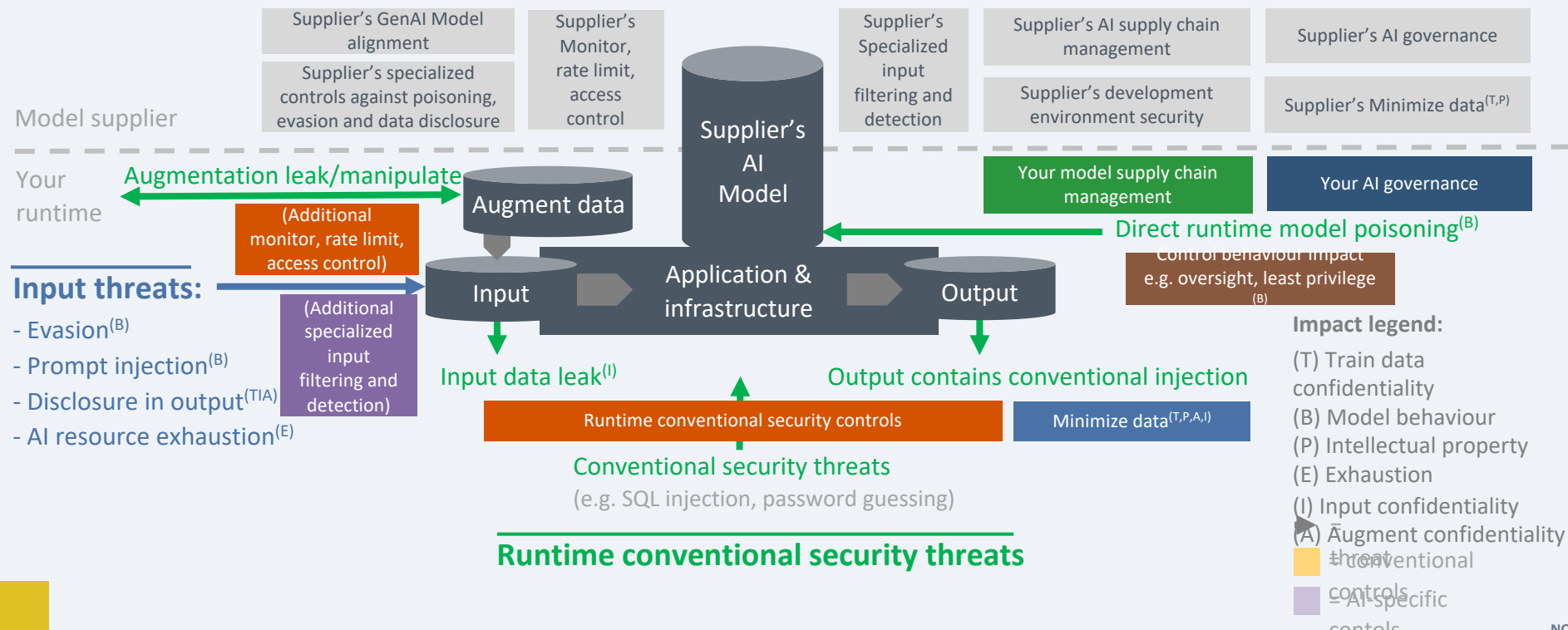
owaspai.org/go/controloverview/





AI Exchange core threat model – ready-made models, self-hosted

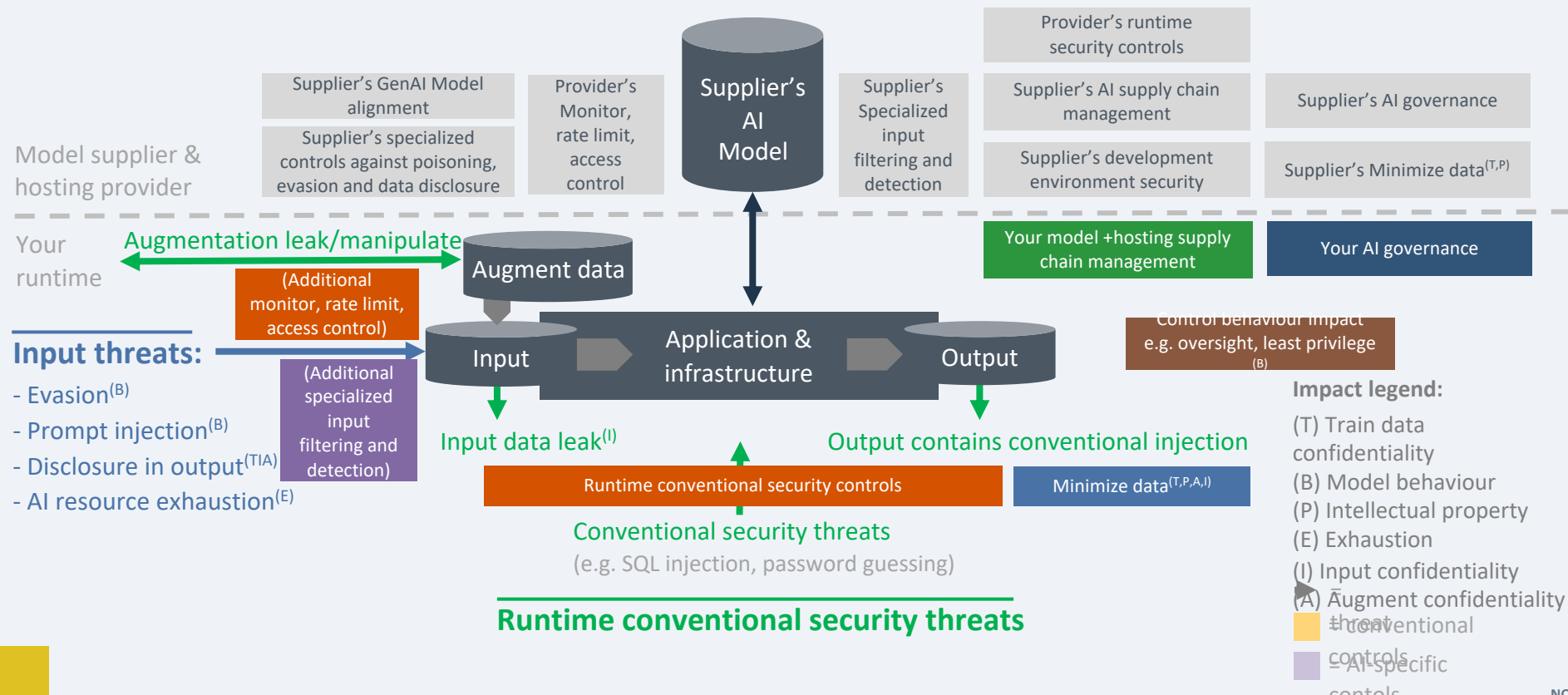
owaspai.org/go/readymademodel/





AI Exchange core threat model – hosted ready-made models

owaspai.org/go/readymademodel/





// AI SECURITY CONTROLS

The seven layers of Agentic prompt injection protection

owaspai.org/go/promptinjectionsevenlayers/

