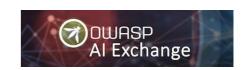
The periodic table of AI security



Found at: https://owaspai.org/goto/periodictable/

The table below, created by the OWASP AI Exchange, shows the various threats to AI and the controls you can use against them – all organized by asset, impact and attack surface, with deeplinks to comprehensive coverage at the <u>AI Exchange website</u> with further references to related standards.

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	Input validation, Input segregation Limit unwanted behavior, Monitor, rate limit, model access control plus:
		Evasion (e.g. adversarial examples)	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	Model poisoning development time	Limit unwanted behavior, Development environment security, data segregation, federated learning, supply chain management plus:
		Data poisoning of train/finetune data	model ensemble Limit unwanted behavior, Development environment security, data segregation, federated learning, supply chain management plus:
			model ensemble plus:
			More training data, data quality control, train data distortion, poison robust model
	Development - Supply chain	Model/data poisoning in supply chain	Limit unwanted behavior, Supplier: Development environment security, data segregation, federated learning
			Producer: supply chain management plus:
			model ensemble
Training data Confidentiality	Runtime - Model use	Data disclosure in model output	Sensitive data limitation (data minimize, short retain, obfuscate training data) plus:
			Monitor, rate limit, model access control plus:
		Model inversion / Membership inference	<u>Sensitive model output</u> <u>Sensitive data limitation</u> (data minimize, short retain, obfuscate training data) plus:
			Monitor, rate limit, model access control plus:
	Development - Engineering environment	Training data leaks	Obscure confidence, Small model Sensitive data limitation (data minimize, short retain, obfuscate training data)
			Development environment security, data
Model confidentiality	Runtime - Model use Runtime - Break into deployed model	Model theft through use (input-output harvesting)	<u>segregation</u> , <u>federated learning</u> <u>Monitor</u> , <u>rate limit</u> , <u>model access control</u>
		Direct model theft runtime	Runtime model confidentiality, Model obfuscation
	Development - Engineering environment	Model theft development- time	Development environment security, data segregation, federated learning
Model behaviour Availability	Model use	Denial of model service (model resource depletion)	Monitor, rate limit, model access control plus: Dos input validation, limit resources
Model input data Confidentially	Runtime - All IT	Model input leak	Model input confidentiality
Any asset, CIA	Runtime - All IT	Model output contains injection	Encode model output
Any asset, CIA	Runtime - All IT	Conventional runtime security attack on conventional asset	Conventional runtime security controls
Any asset, CIA	Runtime - All IT	Conventional attack on conventional supply chain	Conventional supply chain management controls