owaspai.org

OWASP AI EXCHANGE

AI security threats and controls navigator

LEGEND:

- Standard information security control (with attention points)
- Runtime AI engineer control
- Development-time AI engineer control
- Other control
- Threat (clickable)

Impact on Confidentiality, Integrity or Availability

1. General controls against all threats:

Governance:
- AI PROGRAM
- SEC PROGRAM
- SEC DEV PROGRAM
- DEV PROGRAM
- CHECK COMPLIANCE
- SEC EDUCATE

Sensitive data limitation:
- DATA MINIMIZE
- ALLOWED DATA
- SHORT RETAIN
- OBFUSCATE TRAININGDATA
- DISCRETE

Limit effect of unwanted behavior:
- OVERSIGHT
- LEAST MODEL PRIVILEGE
- MODEL ALIGNMENT
- AI TRANSPARENCY
- CONTINUOUS VALIDATION
- EXPLAINABILITY
- UNWANTED BIAS TESTING

2. Controls against Input threats:

Always against input threats:
- MONITOR USE
- RATE LIMIT
- MODEL ACCESS CONTROL
- ANOMALOUS INPUT HANDLING
- UNWANTED INPUT SERIES HDL.
- OBSCURE CONFIDENCE

Integrity of model behaviour

2.1 Against evasion:
- See Always
- EVASION INPUT HANDLING
- EVASION ROBUST MODEL
- TRAIN ADVERSARIAL
- INPUT DISTORTION
- ADVERSARIAL ROBUST DISTILLATION

2.2.1 Against direct prompt injection:
- PROMPT INJECTION IO HANDLING

2.2.2 Against indirect prompt injection:
- PROMPT INJECTION IO HANDLING
- INPUT SEGREGATION

Confidentiality of data

2.2 Against disclosure through use:

Against disclosure in output :
- See always
- SENSITIVE OUTPUT HANDLING

Against model inversion and membership inference:
- See always
- SMALL MODEL

Confidentiality of intellectual property

2.3 Against model exfiltration:
- See always
- MODEL WATERMARKING

Availability of model

2.4 Against AI resource exhaustion:
- See always
- DOS INPUT VALIDATION
- LIMIT RESOURCES

3. Controls against development-time threats:

Always against dev-time threats:
- DEV SECURITY
- SEGREGATE DATA
- CONF COMPUTE
- FEDERATIVE LEARNING
- SUPPLY CHAIN MANAGE

Integrity of model behaviour

3.1 Against broad model poisoning:
- See Always
- MODEL ENSEMBLE

Against data poisoning:
- See always
- MORE TRAIN DATA
- DATA QUALITY CONTROL
- TRAIN DATA DISTORTION
- POISON ROBUST MODEL

Against dev-time model poisoning:
- See always

Against supply-chain poisoning:
- See always

Confidentiality of data / ip

3.2 Against data leak development-time:

Against Train/test data leak:
- See Always

Against dev-time model leak:
- See Always

Against source code/config leak:
- See Always

4. Runtime conventional security controls:

All CIA risks

4.1 Against generic security threats:
- Technical appsec controls
- Operational security

Integrity of model behaviour

4.2 Against runtime model poisoning:
- RUNTIMEMODELINTEGRITY
- RUNTIMEMODELIOINTEGRITY

Confidentiality of intellectual property

4.3 Against runtime model leak:
- RUNTIMEMODELCONFIDENTIALITY
- MODELOBFUSCATION

CIA risks through injection

4.4 Against output contains conventional injection:
- ENCODEMODELOUTPUT

Confidentiality of input

4.5 Against input data leak:
- MODEL INPUT CONFIDENTIALITY

Confidentiality of augmentation data

4.6 Against augmentation data leak:
- AUGMENTATION DATA CONFIDENTIALITY

Integrity of augmentation data

4.6 Against augmentation data manipulation:
- AUGMENTATION DATA INTEGRITY

Threat model based on Software Improvement Group AI framework