Overview of AI security threats and controls at OWASP AI Exchange - owaspai.org

LEGEND:

- Standard information security control (with attention points)
- Runtime Datascience control
- Development-time Datascience control

1. General controls:

Governance:

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

Sensitive data limitation:

- DATAMINIMIZE
- ALLOWEDDATA
- SHORTRETAIN
- OBFUSCATETRAININGDATA
- DISCRETE

Limit effect of unwanted behavior:

- OVERSIGHT
- MINMODELPRIVILEGE
- AITRANSPARENCY
- CONTINUOUSVALIDATION
- EXPLAINABILITY
- UNWANTEDBIASTESTING

2. Controls against threats through runtime use:

Always:

- MONITORUSE
- RATELIMIT
- MODELACCESSCONTROL

Integrity of model behaviour

2.1 Against evasion:

- See Always
- DETECTODDINPUT
- DETECTADVERSARIALINPUT
- EVASIONROBUSTMODEL
- TRAINADVERSARIAL
- INPUTDISTORTION
- ADVERSARIALROBUSTDISTILLATION

Confidentiality of data

2.2 Against data disclosure by use:

Against data disclosure by model:

- See always
- FILTERSENSITIVETRAINDATA
- FILTERSENSITIVEMODELOUTPUT

Against model inversion and membership inference:

- See always
- OBSCURECONFIDENCE
- SMALLMODEL
- ADDTRAINNOISE

Confidentiality of intellectual property

2.3 Against model theft by use:

- See always

Availability of model

2.4 Against failure by use:

- See always
- DOSINPUTVALIDATION
- LIMITRESOURCES

3. Controls against development-time threats:

Always:

- DEVDATAPROTECT
- DEVSECURITY
- SEGREGATEDATA
- CONFCOMPUTE
- FEDERATIVELEARNING
- SUPPLYCHAINMANAGE

Integrity of model behaviour

3.1 Against broad model poisoning:

- See Always
- MODELENSEMBLE

Against data poisoning:

- See always
- MORETRAINDATA
- DATAQUALITYCONTROL
- TRAINDATADISTORTION
- POISONROBUSTMODEL

Against dev-time model poisoning:

- See always

Against transfer learning attacks:

- 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 application security controls:

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 intellectual property

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 data

4.7 Against leaking input data:

- MODELINPUTCONFIDENTIALITY

Threat model based on Software Improvement Group AI framework