AI security threats and controls navigator from the OWASP AI Exchange at owaspai.org

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

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

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