ISO/IEC 27090:2026-10 (E)

Cybersecurity - Artificial intelligence (AI) - Addressing security threats and compromises to AI systems

Contents

Page

Foreword v

Introduction vi

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Abbreviated terms 5

5 Application of information security 5
5.1 General 5
5.2 AI governance programme 6
5.3 Zero trust principles 7
5.4 Applying software engineering best practices to AI 8
5.5 Supply chain management 8
5.5.1 General 8
5.5.2 AI-specific supply chain threats 9
5.5.3 Complex AI supply chains 9
5.5.4 Specific attack vectors 9
5.5.5 Mitigation strategies 9
5.6 Threat modelling 10
5.7 Red teaming 10

6 Types of potential compromise to AI systems 10
6.1 General 10
6.2 Data poisoning attack 13
6.2.1 Description 13
6.2.2 Failures resulting from data poisoning attacks 14
6.2.3 Detection of data poisoning attacks 14
6.2.4 Mitigation of data poisoning attacks 15
6.3 Evasion attack 17
6.3.1 Description 17
6.3.2 Failures resulting from evasion attacks 18
6.3.3 Detection of evasion attacks 18
6.3.4 Mitigation of evasion attacks 18
6.4 Membership inference 20
6.4.1 Description 20
6.4.2 Failures resulting from membership inference 20
6.4.3 Detection of membership inference 21
6.4.4 Mitigation of membership inference 21
6.5 Model exfiltration 21
6.5.1 Description 21
6.5.2 Failures resulting from model exfiltration 22
6.5.3 Detection of model exfiltration 22
6.5.4 Mitigation of model exfiltration 23
6.6 Model inversion 23
6.6.1 Description 23
6.6.2 Failures resulting from model inversion 23

6.6.3 Detection of model inversion 24
6.6.4 Mitigation of model inversion 24
6.7 Direct model poisoning 24
6.7.1 Description 24
6.7.2 Failures resulting from direct model poisoning 25
6.7.3 Detection of direct model poisoning 25
6.7.4 Mitigation of direct model poisoning 25
6.8 Sensitive model output 25
6.8.1 Description 25

6.8.2 Failures resulting from sensitive model output 25
6.8.3 Detection of sensitive model output 25
6.8.4 Mitigations of sensitive model output 25
6.9 Prompt injection 26
6.9.1 Description 26
6.9.2 Failures resulting from prompt injection 26
6.9.3 Detection of prompt injection 26
6.9.4 Mitigation of prompt injection 27
6.10 Output contains injection attacks 27
6.10.1 Description 27
6.10.2 Failures resulting from output contains injection attacks 27
6.10.3 Detection of output contains injection attacks 27
6.10.4 Mitigations of output contains injection attacks 27
6.11 Considerations for AI systems 28
6.11.1 General 28
6.11.2 Reduce impact by data limitation 28
6.11.3 Governing and controlling the effect of unwanted model behaviour during the operation stage 29
6.11.4 AI system and data provenance 32

7 Mitigations and their interactions with threats and other applied mitigations 33
7.1 General 33
7.2 Conflicting interactions of mitigations 33
7.3 Continuity of mitigations across the AI life cycle 33
7.4 Model and mitigation deterioration over time 33
7.5 Logging and monitoring 34
7.6 Development environment protection 34
7.7 Malicious input detection and filtering 35
7.7.1 Detecting out-of-distribution inputs 35
7.7.2 Detecting adversarial inputs 35
7.8 Throttling model use 36
7.9 Optimising the size of the model 36

Annex A (informative) Mapping attacks to AI system life cycle and to assets 37

Annex B (informative) AI-specific versions of conventional attacks 44

Annex C (informative) Federated machine learning 49

Bibliography 51