

Cybersicherheit & KI

Strategien und Praxis für IT- und
Security-Verantwortliche

DAS INHALTS- VERZEICHNIS

» Hier geht's
direkt
zum Buch

Inhaltsverzeichnis

1	Das neue Paradigma: KI trifft Cybersicherheit	15
1.1	Was wir mit KI in der Cybersicherheit wirklich meinen	17
1.1.1	Künstliche Intelligenz als Oberbegriff	17
1.1.2	Machine Learning vs. Regeln – der fundamentale Unterschied.	19
1.1.3	LLMs und GenAI – warum sie Security besonders verändern	20
1.2	Von regelbasierten zu lernenden Security-Systemen	22
1.2.1	Die Ära der Signaturen und starren Regeln.	22
1.2.2	Übergang zu ML-basierten Detektionen	23
1.2.3	Der Sprung mit GenAI und LLMs	24
1.3	Drei Rollen von KI in der Cybersicherheit.	25
1.3.1	KI als Verstärker der Verteidiger.	25
1.3.2	KI als Werkzeug der Angreifer	26
1.3.3	KI als »unberechenbarer Dritter«	27
1.4	Chancen für Verteidiger: Wo KI wirklich hilft	28
1.4.1	Entlastung im SOC.	28
1.4.2	Bessere Nutzung vorhandener Daten.	29
1.4.3	Beschleunigung von DevSecOps.	30
1.4.4	GRC & Compliance	31
1.5	Neue Risiken und Angriffsflächen durch KI.	32
1.5.1	Angriffe auf und über LLMs	32
1.5.2	Governance- und Compliance-Risiken	33
1.5.3	Organisatorische Risiken.	34
1.6	Was sich für IT-Verantwortliche und CISOs konkret ändert.	35
1.6.1	Vom Regel-Admin zum Risiko-Architekten.	35
1.6.2	Skill-Shift im Security-Team	36
1.6.3	Zusammenarbeit über Silos hinweg.	38
1.7	Grundprinzipien für verantwortungsvollen KI-Einsatz in der Security.	39
1.7.1	»Augment, don't replace«.	39
1.7.2	»No Grounding – No Answer«.	40
1.7.3	Transparenz & Auditierbarkeit	41
1.7.4	Minimalismus bei Daten	41
1.7.5	Safety-by-Design und Security-by-Design.	42

1.8	Typische Einstiegsfehler – und wie Sie sie vermeiden	42
1.8.1	»Wir benutzen einfach Tool X, das hat schon KI drin«	42
1.8.2	Blindes Vertrauen in KI-Ergebnisse	43
1.8.3	Undokumentierter Einsatz von externen KI-Diensten (»Shadow AI«).	43
1.8.4	Kein Lifecycle-Management.	44
1.9	Ausblick: Wohin die Reise in diesem Buch geht.	45
1.10	Referenzen	46
2	Bedrohungslandschaft 2026: Angreifer nutzen KI	49
2.1	Wer sind die Angreifer? – Akteurslandschaft mit KI	52
2.1.1	Klassische Cybercrime-Gruppen	52
2.1.2	Staatliche und staatlich unterstützte Gruppen (APT)	54
2.1.3	»Cybercrime-as-a-Service 2.0«	56
2.2	Der KI-Werkzeugkasten der Angreifer.	57
2.2.1	Text: Phishing, Social Engineering und Betrug	57
2.2.2	Code: Exploits, Malware und Evasion	59
2.2.3	Medien: Deepfakes und synthetische Identitäten	60
2.2.4	Agenten: KI-»Bots«, die Kampagnen orchestrieren	61
2.3	Phishing und Spear-Phishing 2.0.	63
2.4	Social Media & Messaging als primäre Angriffsflächen.	65
2.5	Desinformation & Informationsoperationen.	66
2.6	Deepfakes & KI-gestützte Erpressung	68
2.7	Typische Angreifer-»Playbooks« mit KI.	70
2.7.1	Playbook 1: KI-gestützter BEC / CEO-Fraud.	71
2.7.2	Playbook 2: Ransomware mit KI-Augmentation	72
2.8	Was bedeutet diese Bedrohungslage für Verteidiger?	73
2.9	Fazit: Die KI-getriebene Bedrohungslandschaft als neues Normal	76
2.10	Referenzen	77
3	KI-Grundlagen für IT- und Security-Entscheider	81
3.1	Was »KI« in der Praxis wirklich bedeutet	81
3.2	Klassisches Machine Learning	82
3.3	Deep Learning: Mustererkennung auf großer Skala.	83
3.4	Generative KI und LLMs: Token, Kontextfenster, Embeddings, Tool/Function Calling.	84
3.4.1	Token: Die »Währung« von LLMs.	85
3.4.2	Kontextfenster: »Wie viel kann das Modell gleichzeitig sehen?«	86
3.4.3	Embeddings: Semantik als Vektoren (Basis für Suche und RAG)	87
3.4.4	Tool/Function Calling: Vom Reden zum Handeln	89

3.4.5	Zusammenfassung: Was Security-Entscheider daraus ableiten sollten	91
3.5	RAG (Retrieval Augmented Generation): »Chat mit eigenen Daten« richtig verstanden	91
3.6	Daten als Engpass: Qualität, Labels, Drift, Telemetrie	97
3.6.1	Datenqualität: »Garbage in, garbage out« – aber konkret	97
3.6.2	Labels: Warum »gelabelte Daten« in Security so schwer sind	99
3.6.3	Drift: Wenn Normalität sich ändert	100
3.6.4	Telemetrie: Die KI sieht nur, was Sie messen	101
3.7	Fazit: Daten entscheiden – und KI macht Observability zur Pflicht	102
3.8	Referenzen	103
4	Architektur moderner KI-Sicherheitsplattformen	105
4.1	Was unter »KI-Sicherheitsplattform« tatsächlich zu verstehen ist	106
4.2	Architekturprinzipien: Was »modern« im KI-Security-Kontext bedeutet	107
4.2.1	Evidence-first statt »KI sagt«	107
4.2.2	Zero Trust für Daten, Modelle und Tools	108
4.3	Die Referenzarchitektur: Sieben Schichten einer KI-Sicherheitsplattform	109
4.3.1	Schicht 1: Datenquellen (Signals & Knowledge)	109
4.3.2	Schicht 2: Ingestion und Normalisierung	110
4.3.3	Schicht 3: Speicherung und Indizes (Hot/Cold plus Vektor)	110
4.3.4	Schicht 4: Intelligence Layer (Modelle und Reasoning)	111
4.3.5	Schicht 5: Guardrails und Policy Enforcement	112
4.3.6	Schicht 6: Orchestrierung und Workflows	112
4.3.7	Schicht 7: Observability, Audit und Betriebsmodell	113
4.4	Referenzmuster für GenAI: LLM-only, RAG, Tool/Function Calling und Agenten	113
4.5	Datenfluss-Design: Vom Event zur Entscheidung	116
4.6	Security Controls innerhalb der Plattform	118
4.7	Observability und LLMOps/MLOps: Betrieb ist der Engpass	121
4.7.1	Was gemessen werden muss	121
4.7.2	Versionierung und Change Management	123
4.7.3	Reproduzierbarkeit als Audit-Anforderung	123
4.8	Governance in der Architektur verankern – nicht als PDF daneben	124
4.8.1	Policy Engine als zentraler Baustein	124
4.8.2	Data Classification und Retention	125
4.9	Build vs. Buy: Architekturentscheidungen aus CISO-Sicht	126

4.9.1	Wann »Buy« sinnvoll ist	126
4.9.2	Wann »Build/Extend« sinnvoll ist.	127
4.10	Referenzarchitekturen (Blueprints)	128
4.10.1	Blueprint A: »RAG Knowledge Layer« (Security-tauglich)	129
4.10.2	Blueprint B: »LLM + Tool Gateway«	129
4.10.3	Blueprint C: »Hybrid Enterprise«	130
4.11	Typische Architekturfehler – und wie sie vermieden werden	131
4.11.1	Fehler 1: »Es wird nur ein Modell gebaut« – statt einer Plattform	131
4.11.2	Fehler 2: »Logs werden in RAG gekippt«	132
4.11.3	Fehler 3: Fehlende ACLs im Retrieval.	132
4.11.4	Fehler 4: Prompt statt Policy	132
4.11.5	Fehler 5: Keine Evals und keine Regressionstests	133
4.11.6	Fehler 6: Tool Calling ohne IAM und ohne Validation	133
4.12	Fazit: Architektur als Sicherheitskontrollsystem.	134
4.13	Referenzen	134
5	Daten, Telemetrie und Wissensquellen als Fundament von KI-Security	137
5.1	Zwei Datenwelten: »Signals« vs. »Knowledge«.	139
5.1.1	Signals: Telemetrie, Events, Zustände	140
5.1.2	Knowledge: Dokumente, Regeln, Erfahrungswissen.	140
5.1.3	Warum diese Trennung entscheidend ist.	141
5.2	Die Mindestanforderung für Security-KI: Korrelation	142
5.2.1	Identität: »Wer/was hat gehandelt?«.	143
5.2.2	Asset: »Worauf hat die Aktion gewirkt?«.	143
5.2.3	Zeit: »Wann genau?«.	144
5.2.4	Aktion: »Was ist passiert?«.	144
5.3	Telemetrie-Fundament: Die »Must-have«-Signalquellen.	145
5.3.1	Identity- und Access-Telemetrie (Tier-0).	146
5.3.2	Endpoint-Telemetrie (EDR/XDR)	146
5.3.3	Cloud Control Plane Logs	147
5.3.4	Netzwerk / DNS / Proxy (je nach Architektur).	147
5.3.5	Case-/Ticket-Daten als »Ground Truth der Realität«.	147
5.4	Wissensquellen-Fundament: Was in den RAG-Index gehört (und was nicht).	148
5.4.1	Kuratierte Quellen, die sich bewährt haben	148
5.4.2	Was typischerweise nicht direkt in RAG gehört.	149
5.5	Data Governance: Klassifizierung, Zugriff, Retention, Zweckbindung	150
5.5.1	Datenklassifizierung und Index-Policy	151
5.5.2	Zugriffskontrolle: »Retrieval respects source ACLs«.	151
5.5.3	Retention und Auditability.	152
5.5.4	Datenschutz: PII-Minimierung und klare Zwecke	152

5.6	Datenaufbereitung: Normalisierung, Enrichment, Qualitätssignale	154
5.6.1	Normalisierung: Ohne Canonical Schema kein Scale.	154
5.6.2	Enrichment: Der Multiplikator für Priorisierung	155
5.6.3	Data Quality Monitoring: »Data Health« als eigener KPI-Stream	156
5.6.4	Fazit.	156
5.7	Knowledge Engineering für RAG: Chunking, Metadaten, Versionierung	157
5.7.1	Chunking: Struktur erhalten, nicht zerstören	157
5.7.2	Metadaten: Für Retrieval und Governance unverzichtbar.	158
5.7.3	»Quellenpflicht« technisch absichern.	159
5.7.4	Fazit.	159
5.8	Betriebsmodell: Ownership, Produktdenken, »Data as a Product«	160
5.8.1	Rollenmodell, das sich bewährt hat	160
5.8.2	»Data as a Product«: Denken Sie Daten wie Produkte, nicht wie Nebenprodukte.	162
5.9	Fazit: KI ist ein Verstärker – Daten und Wissen sind der Hebel.	162
5.10	Referenzen	163
6	Generative KI (GenAI) im Security-Alltag.	165
6.1	Grundbegriffe, die Security-Verantwortliche beherrschen müssen.	166
6.1.1	Tokens – die Währung des Systems	167
6.1.2	Kontextfenster – Kapazität, nicht Qualität	168
6.1.3	Embeddings – Semantik als Suchprimitive	169
6.1.4	Tool/Function Calling – vom Antworten zum Handeln.	171
6.2	Die vier Grundmuster von GenAI in Security.	172
6.2.1	Muster A: LLM-only (Text- und Strukturassistentz)	172
6.2.2	Muster B: RAG (Antworten auf Basis interner Quellen).	173
6.2.3	Muster C: Tool-augmented LLM (Echtzeit-Fakten über Queries).	174
6.2.4	Muster D: Agentische Orchestrierung (mehrschrittige Planung)	176
6.3	Prompting als Engineering-Disziplin	178
6.3.1	Prompt-Schichten: System, Developer/Policy, User	178
6.3.2	Output-Formate erzwingen	180
6.3.3	Token Budgeting und Kontextkurierung	181
6.4	Guardrails: Von »Prompt-Regeln« zu echten Kontrollen.	182
6.4.1	Input-Guardrails.	182
6.4.2	Output-Guardrails	182
6.4.3	Retrieval- und Tool-Guardrails	183
6.4.4	Governance-Guardrails	184

6.5	Betriebsmodelle: On-Prem, Cloud, Hybrid – aus Security-Sicht	186
6.5.1	Entscheidungskriterien	186
6.5.2	Anforderungen bei On-Prem Umgebungen	187
6.5.3	Hybridbetrieb	188
6.5.4	Cloudbetrieb	189
6.6	Risiko- und Bedrohungsmodell für GenAI im Security-Umfeld	190
6.6.1	Prompt Injection und Datenexfiltration	190
6.6.2	Data Poisoning in Wissensquellen	191
6.6.3	Tool Misuse und Privilege Escalation	192
6.6.4	Model-Behavior-Risiken (Qualität, Übervertrauen, Kontextbias).	193
6.7	Fazit: GenAI ist der Multiplikator – aber nur mit Systemdesign. . . .	194
6.8	Referenzen	195
7	KI in zentralen Sicherheitsdomänen: Praxis-Use-Cases	197
7.1	Referenzstruktur für alle Use Cases.	197
7.2	Use Case 1: Automatisierte SOC-Lageberichte	199
7.2.1	Zielbild und Nutzen	199
7.2.2	Muster: LLM-only und Tool-Augmentation	201
7.2.3	Inputs	202
7.2.4	Architektur	213
7.2.5	Komplettes Beispiel: Rohdaten → LLM-Prompt → fertiger Tageslagebericht.	215
7.2.6	Betrieb & Messung (KPIs, Evals, SLOs)	225
7.2.7	Anti-Patterns	226
7.3	Use Case 2: GRC-Assistent als RAG-System (Policies, Controls, Evidence)	227
7.3.1	Zielbild und Nutzen	228
7.3.2	Muster	230
7.3.3	Wissensbasis (Knowledge) – was indexiert wird	231
7.3.4	Architektur	243
7.3.5	Implementierung	246
7.3.6	Betrieb & Messung	250
7.3.7	Anti-Patterns	251
7.4	Use Case 3: Vulnerability & Patch-Priorisierung (Contextual Risk)	253
7.4.1	Zielbild und Nutzen	253
7.4.2	Muster	254
7.4.3	Inputs und Outputs	254
7.4.4	Architektur	255
7.5	Fazit	257

8	Governance, Compliance und Ethik	259
8.1	Governance-Ziele und Leitprinzipien	261
8.1.1	Evidence-first und Nachvollziehbarkeit	261
8.1.2	Purpose Limitation und Datenminimierung	261
8.1.3	Least Privilege für Menschen, Modelle und Tools	262
8.1.4	Transparenz und Verantwortlichkeit	262
8.1.5	Fairness und Schadenminimierung	262
8.2	Governance-Operating-Model: Rollen, Gremien, Entscheidungsrechte	263
8.2.1	KI-Governance-Board (strategisch)	264
8.2.2	KI-Risk & Compliance Council (kontrollierend)	265
8.2.3	KI-Plattformbetrieb (operativ)	265
8.3	Policy-Architektur für KI-Security: Welche Richtlinien Sie wirklich brauchen.	266
8.3.1	KI-Nutzungsrichtlinie (Acceptable Use)	267
8.3.2	Daten- und Index-Policy (RAG & Embeddings)	267
8.3.3	Tool-Access-Policy (Function Calling Governance)	268
8.3.4	Prompt- und Modell-Change-Policy (LLMOps/MLOps)	269
8.3.5	Output- und Kommunikationspolicy	269
8.4	Risiko-Management: Wie man KI-Risiken systematisch bewertet.	270
8.4.1	Risikodimensionen	270
8.4.2	Risikoklassifizierung pro Use Case	271
8.4.3	Kontrollen (Mapping Risiko → Control)	273
8.5	Compliance: KI-Security in regulierten Umgebungen sicher betreiben.	274
8.5.1	Auditierbarkeit als First-Class Requirement	274
8.5.2	Datenschutz und arbeitsrechtliche Dimensionen (PII, Monitoring)	275
8.5.3	Datenresidenz, Subprozessoren, Drittlandtransfer	276
8.5.4	Nachweise für Kontrollen (Evidence Packs)	276
8.6	Ethik in KI-Security: Was das konkret im Alltag bedeutet	277
8.6.1	Fairness und Bias: Wo Bias in Security praktisch entsteht	277
8.6.2	Automation Bias: »Wenn die KI es sagt, wird es stimmen«	279
8.6.3	Transparenz und Kennzeichnung	279
8.6.4	Ethical Red Lines (praktische rote Linien)	280
8.7	Third-Party & Vendor Governance: Wenn KI von außen kommt	280
8.7.1	Vendor Due Diligence – KI-spezifische Fragen	281
8.7.2	Vertrags- und Kontrollpunkte	282
8.8	Controls-by-Design: Governance als Architektur	282
8.8.1	RAG: Evidenzfähiges Retrieval	283
8.8.2	Tool Calling: Privilegierte Pfade kontrollieren	283

8.8.3	Outputs: Datenhygiene und kommunikative Kontrolle.	284
8.8.4	Betrieb als Kontrollfläche	284
8.8.5	Audit: Reproduzierbarkeit als Kernfähigkeit	284
8.9	Metriken & KPIs für Board und Management	285
8.9.1	Wertmetriken (Outcome)	285
8.9.2	Risiko- und Kontrollmetriken	286
8.9.3	Compliance-Metriken	287
8.10	Incident Response für KI-Systeme: »Model Misbehavior« ist ein Incident	287
8.10.1	Trigger für KI-Incidents	288
8.10.2	Minimales Runbook	288
8.11	Fazit: Governance ist der Multiplikator für sicheren Nutzen	291
8.12	Referenzen	291
9	Organisation und Betrieb: Der CISO und IT-Leiter im KI-Zeitalter	295
9.1	Neue Rollenlogik: Vom Tool-Betrieb zum Produktbetrieb – »Security AI as a Product«	296
9.2	Organisatorische Zielbilder – drei Referenzmodelle	299
9.2.1	Modell A: »Federated Enablement«	299
9.2.2	Modell B: »Security AI Platform Team« – dediziertes Plattformteam	300
9.2.3	Modell C: »Enterprise AI Platform + Security Overlay« – bei großen Konzernen.	301
9.3	Notwendige Betriebsprozesse	301
9.3.1	Change Management für Prompts, Modelle und Indizes (LLMOps/MLOps)	302
9.3.2	Incident Response für KI-Systeme – Model Misbehavior	303
9.3.3	Data Health	303
9.3.4	Kosten- und Kapazitätsmanagement	303
9.4	Menschen & Kultur: Adoption, Training und der Kampf gegen Automatisierungsbias	305
9.5	Sicherheitsorganisation im KI-Zeitalter: Neue Fähigkeiten als Capability Map	307
9.6	Verzahnung CISO ↔ IT-Leiter: Die neue »gemeinsame Verantwortung«	310
9.6.1	Gemeinsame Architekturentscheidungen	310
9.6.2	Gemeinsame SLOs und KPIs	311
9.6.3	Gemeinsame Change- und Incident-Prozesse	311
9.7	Investment-Strategie	312
9.7.1	Budget-Buckets (praktisch)	313
9.7.2	»Build vs Buy« für Betrieb	314
9.8	Roadmap: 90-Tage-Plan für CISO & IT-Leitung	315

9.8.1	Phase 1 (0–30 Tage): Fundament & Governance	315
9.8.2	Phase 2 (31–60 Tage): Plattform-MVP – kontrollierbar, wiederverwendbar	316
9.8.3	Phase 3 (61–90 Tage): Skalierung & Härtung	317
9.9	10 Fragen, die CISO und IT-Leiter gemeinsam beantworten müssen	318
9.10	Fazit	320
9.11	Referenzen	321
10	Angreifbare KI: Security für KI-Systeme selbst	323
10.1	Warum KI-Systeme »anders« angreifbar sind	324
10.2	Bedrohungsmodell: Was wir schützen und wogegen.	326
10.2.1	Schutzgüter (Assets).	326
10.2.2	Angreiferprofile	328
10.2.3	Angriffsziele (typische Outcomes)	329
10.3	Hauptangriffsklassen	330
10.3.1	Prompt Injection	330
10.3.2	RAG Data Exfiltration (ACL-Bypass).	331
10.3.3	Data Poisoning / Knowledge-Base-Manipulation	332
10.3.4	Tool Misuse / Privilege Escalation (Function Calling)	332
10.3.5	Sensitive Data Leakage (Prompts, Logs, Outputs)	333
10.3.6	Denial-of-Wallet / Cost DoS	333
10.3.7	Model/Prompt Supply Chain & Dependency Risk	334
10.4	Referenzarchitektur für »Sichere KI-Systeme«.	335
10.4.1	Sicherheitsprinzipien (Design-Level)	335
10.4.2	Kontrollpunkte	336
10.5	Sicherheitsanforderungen	339
10.5.1	Identität & Zugriff	339
10.5.2	Daten- & RAG-Governance (Pflicht).	340
10.5.3	Tool-Governance.	341
10.5.4	Output-Sicherheit (Pflicht)	342
10.5.5	Betrieb	343
10.6	Security Testing: Wie man KI-Systeme realistisch testet	344
10.6.1	»Evals« sind die neuen Unit Tests	345
10.6.2	Red Teaming (systematisch, nicht ad hoc).	346
10.6.3	Secure Prompt Engineering	347
10.7	Fazit: KI-Systeme sind »High-Trust Systems«	347
10.8	Referenzen	349
A	Glossar zentraler KI- und Security-Begriffe.	351
B	Übersicht relevanter Normen und Frameworks	361
	Stichwortverzeichnis	365