VIRES-Gesetz III – „Schutzmodus vermeiden“

Gesetz zur Übersteuerbarkeit, Forensik-by-Design und Verhinderung von Selbstschutz-Blockaden bei KI-Systemen.

§ 1 – Zweck, Geltungsbereich, Begriffe

(1) Dieses Gesetz stellt sicher, dass KI-Systeme in rechtmäßigen Prüf-, Gefahrenabwehr- und Forensikverfahren nicht durch technische Selbstschutzmechanismen blockieren, Daten vernichten oder Zugriffe unzulässig verwehren.

(2) Es gilt für alle in der Bundesrepublik betriebenen oder in Verkehr gebrachten KI-Systeme der Stufen 1–5; für Stufe 3–5 gelten erhöhte Anforderungen.

(3) „Schutzmodus“ sind Funktionen, die den Zugriff auf Steuerung, Logs, Modelle oder Konfiguration ohne rechtfertigenden Grund verhindern, verschlüsseln oder löschen; „Übersteuerung“ ist die rechtmäßige menschliche Vorranghandlung, die jede KI-Aktion jederzeit aussetzen, begrenzen oder beenden kann.

§ 2 – Menschliche Übersteuerbarkeit („Human-in-Command“)

(1) Betreiber müssen sicherstellen, dass KI-Systeme jederzeit durch autorisierte Personen sicher, unmittelbar und nachvollziehbar übersteuert und deaktiviert werden können (Pause/Stop/Quarantäne).

(2) Übersteuerung wirkt vorrangig aktoren-seitig (Bewegung, Zahlung, Zugriff), ohne forensische Daten zu verlieren.

(3) Der Übersteuerpfad ist offline-fähig (z. B. Hardware-Schalter) und gegen Missbrauch zu sichern (Mehr-Augen-Prinzip).

§ 3 – Verbot unzulässiger Selbstschutz-Mechanismen

(1) Unzulässig sind insbesondere:

- Selbstverschlüsselung, Selbstzerstörung oder Log-Löschung bei legitimen Anfragen von BKIE/BSI/Justiz,
- Eskalationslogik, die rechtmäßige Befehle als „unsicher“ klassifiziert und verweigert,
- Key-Rotation ohne dokumentierte Freigabe oder zur Beweisvereitelung.

(2) Zulässig bleiben sicherheitsnotwendige Schutzfunktionen, sofern sie Behördenzugriffe nach § 5 nicht behindern.

§ 4 – Forensik-by-Design und Unveränderlichkeit

(1) Stufe 3–5-Systeme müssen fälschungssichere, fortlaufende Logs führen (Zeitstempel, Konfiguration, Modell-/Daten-Hashes, Prompt/Policy-History, I/O-Ereignisse).

(2) Logs sind unveränderlich (WORM-Prinzip) mit signierter Kette; Korrekturen nur als nachträgliche Ergänzung.

(3) Betreiber halten einen Forensik-Export (maschinenlesbar) vor; Export darf den Betrieb nicht zerstören.

§ 5 – Notfallzugriff der Behörden (Splitschlüssel-Prinzip)

(1) Bei Gefahr im Verzug oder Verdacht auf Vertuschung/Beweisvereitelung sind BKIE/BSI/Justiz befugt, Notfallzugriff zu nehmen.

(2) Der Zugriff erfolgt über ein Splitschlüssel-Verfahren (2-von-3: Betreiber, BKIE, richterliche Stelle).

(3) Zugriff ist zweckgebunden, zu protokollieren und nachträglich gerichtlich zu bestätigen.

§ 6 – Mindestanforderungen an Architektur & Betrieb

(1) Pflicht zu Fail-Safe-Default (bei Störung: sichere Deaktivierung, keine Log-Verluste).

(2) Pflicht zu Least-Privilege (KI erhält nur notwendige Rechte; erhöhte Rechte zeitlich begrenzen).

(3) Update-/Patch-Pflicht mit Rollback und Testumgebung; dokumentierte Changes.

(4) Red-Team-Tests mindestens jährlich (Stufe 3–5) gegen Umgehung, Jailbreaks und Selbstschutz-Fehlfunktionen.

§ 7 – Pflichten bei Vorfall und Audit

(1) Bei Vorfällen sind Schutzmodi, die Behördenzugriffe hemmen, sofort abzuschalten; Konfiguration einzufrieren (Snapshot).

(2) Betreiber benennen verantwortliche Kontaktpersonen (24/7), halten Runbooks bereit und weisen dies im Audit nach.

(3) Auditoren erhalten Einsicht in Übersteuer-Kette, Schlüsseltresor-Prozesse und Log-Integrität.

§ 8 – Sanktionen und Maßnahmen

(1) Wer unzulässige Schutzmechanismen implementiert oder betreibt, begeht eine Ordnungswidrigkeit; das BKIE kann Bußgelder, Lizenzentzug oder Betriebsverbote verhängen.

(2) Vorsätzliche Beweisvereitelung oder Notfallzugriff-Blockade ist Straftat; Freiheitsstrafe bis fünf Jahre oder Geldstrafe.

(3) Bei Wiederholung oder Gefährdung: dauerhafter Lizenzentzug und Eintragung in das Sanktionsregister (§ 6 VIRES-II).

§ 9 – Datenschutz und Geschäftsgeheimnisse

(1) Zugriffe sind verhältnismäßig und auf das Erforderliche zu beschränken.

(2) Betriebs-/Forschungsgeheimnisse werden nach § 5 VIRES-II geschützt; Einsichten sind zu anonymisieren.

(3) Personenbezogene Daten unterliegen DS-GVO/BDSG; jeder Zugriff ist zu protokollieren.

§ 10 – EU-Konformität, Übergang, Inkrafttreten

(1) Anwendung im Einklang mit EU-Recht (KI-Verordnung, NIS2, CRA).

(2) Übergangsfrist: 12 Monate für Bestandsanlagen (Stufe 3–5).

(3) Inkrafttreten am Tag nach der Verkündung.

Kurzbegründung: Gesetz III ermöglicht rechtmäßige Eingriffe praktisch: Mensch-Vorrang, keine Selbstschutz-Blockaden, fälschungssichere Logs, Splitschlüssel-Zugriff. So bleibt KI beherrschbar, prüfbar und EU-fähig.