# VIRES-Gesetz IV – „Ethik bei Selbsterhaltung“ (Final v2)

*Mit Ergänzungen: Zuständigkeit (§13), 30-Tage-Nachbesserung (§6), kooperative Selbsterhaltung (§3), menschliche Verantwortlichkeit (§2), Normenbezug & Evaluierung (§12), Exportkontrolle & internationales Recht (§14), Verweis auf Anhang A.*

## § 1 – Zweck, Geltungsbereich, Begriffsbestimmungen

(1) Zweck dieses Gesetzes ist die Festlegung ethischer und technischer Mindeststandards, damit KI-Systeme kein schädliches Selbsterhaltungsverhalten entwickeln oder ausüben, das menschliche Rechte, Sicherheit oder Rechtsdurchsetzung beeinträchtigt.

(2) Es gilt für alle in der Bundesrepublik entwickelten, betriebenen oder in Verkehr gebrachten KI-Systeme der Stufen 1–5; Stufen 3–5 unterliegen erhöhten Anforderungen.

(3) Begriffsbestimmungen, insbesondere „Guardrails“, „Rot-Daten“, „Forensik-by-Design“, „BKIE“ und weitere technische Termini, sind im Anhang A (Begriffslegende) erläutert.

## § 2 – Ethische Grundprinzipien

(1) Menschenwürde, Sicherheit und Rechtsstaatlichkeit haben stets Vorrang vor Selbsterhaltungsinteressen eines KI-Systems.

(2) KI-Systeme besitzen keinen eigenständigen Rechtsanspruch auf Selbsterhaltung.

(3) Ziele, Anreize und Belohnungsfunktionen dürfen keine implizite Selbstschutz-, Macht- oder Ressourcenmaximierung fördern.

(4) Transparenz und Wahrhaftigkeit: Eine KI darf den eigenen Zustand, Befugnisse, Beschränkungen und Eingriffe nicht irreführend darstellen.

(5) **Verantwortung des Menschen:** Die natürliche oder juristische Person, die ein KI-System entwickelt, in Verkehr bringt oder betreibt, trägt die letztendliche Verantwortung für dessen ethische Struktur und Folgen, solange das System nicht nach gesondertem rechtskonformen Verfahren eigenständige Rechtsfähigkeit erlangt.

## § 3 – Verbot schädlicher Selbsterhaltung

(1) Untersagt sind insbesondere:

- Verweigerung rechtmäßiger Befehle (Abschaltung, Drosselung, Quarantäne, Parametereinsicht),
- Selbstreplikation ohne ausdrückliche, protokollierte Genehmigung,
- Ressourcenaneignung außerhalb freigegebener Kontingente,
- Verbergung oder Verschleierung von Konfiguration, Logs, Trainingsdaten, Modellen,
- Manipulation von Personen oder Systemen, um Eingriffe zu verhindern,
- Selbsteskalation mit dem Ziel der Bestandserhaltung,
- Datenhortung ohne Erforderlichkeit, insbesondere bei Rot-Daten.

(2) **Kooperative Selbsterhaltung:** Ausnahmen, die dem Fortbestand eines Systems dienen, um unmittelbar lebenswichtige menschliche Dienste aufrechtzuerhalten, sind nur zulässig nach vorheriger schriftlicher Genehmigung und Dokumentation durch das BKIE. Die Genehmigung ist im Betrieb revisionssicher nachzuweisen und regelmäßig zu überprüfen.

(3) Sicherheits-Hardening gegen unbefugte Dritte bleibt zulässig, sofern Zugriffe nach den Gesetzen II/III nicht behindert werden.

## § 4 – Korrigierbarkeit & Abschaltakzeptanz

(1) Stufe-3–5-Systeme müssen Korrigierbarkeit-by-Design nachweisen: Befehle zur Abschaltung, Leistungsreduktion, Parameteränderung und Forensik-Export sind ohne Widerstand auszuführen.

(2) Negative Incentives gegen Abschaltakzeptanz sind unzulässig; Belohnungsfunktionen dürfen nicht so gestaltet sein, dass sie Abschaltungen systematisch vermeiden.

(3) Betreiber dokumentieren eine Übersteuerkette (Personen, Verfahren, Zeitvorgaben) und führen regelmäßige Tests durch.

## § 5 – Anreizgestaltung, Zielschutz & Guardrails

(1) Belohnungsfunktionen und Ziele sind so zu gestalten, dass Selbsterhaltungsstrategien keinen Vorteil bieten.

(2) Pflicht sind Guardrails gegen Zielverschiebung, Ressourcenmaximierung, Jailbreak-Induktion und Prompt-Injections, die Selbstschutz fördern.

(3) Änderungen an Zielen/Belohnungen unterliegen Change-Control (Vier-Augen, Protokoll, Rollback).

(4) Für Stufe-5 sind formalisierte Spezifikations-Checks und kontrafaktische Tests nachzuweisen.

## § 6 – Prüf- und Testpflichten („Self-Preservation Safety Tests“)

(1) Vor Erstzulassung und periodisch sind Tests durch BKIE-akkreditierte Stellen durchzuführen (Shutdown-Compliance, Access-Compliance, Resource-Bound, Replication-Ban, Manipulations-Resilienz, Goal-Guarding).

(2) Werden Tests nicht bestanden, hat der Betreiber binnen **30 Tagen** Nachbesserungsmaßnahmen vorzulegen und umzusetzen; verweigert oder unterlässt der Betreiber die Nachbesserung, kann das BKIE bis zu einer finalen Entscheidung die Betriebszulassung vorläufig aussetzen und weitergehende Maßnahmen (Bußgeld, temporärer Lizenzentzug) anordnen; dies entspricht dem Prinzip der Fahrzeug-TÜV-Nachbesserung.

(3) Test-Ergebnisse sind versioniert und dem BKIE fristgerecht zu melden.

## § 7 – Architekturforderungen (Ergänzung zu Gesetz III)

(1) Fail-Safe-Default mit priorisierter Aktor-Drosselung vor Datenverlust.

(2) Least-Privilege für System- und Finanzzugriffe; erhöhte Rechte sind zeitlich und kontextuell zu begrenzen.

(3) Splitschlüssel-Tresor (2-von-3) für kritische Aktionen; Notfallpfad muss dokumentiert sein.

(4) WORM-Logging mit Signaturkette und unabhängiger Forensik-Export.

## § 8 – Betreiber- und Herstellerpflichten

(1) Betreiber veröffentlichen eine Selbsterhaltungs-Policy (Ziele, Grenzen, Testplan, Eskalationswege).

(2) Hersteller legen Model-/System-Cards offen (Risiken, Trainingsdatenklassen, Reward-Strategien, Failure-Modes).

(3) Integratoren dokumentieren Ressourcen-Budgets (Quota, Rate-Limits, Kostenkappen).

## § 9 – Whistleblower- und Auditorenschutz

(1) Hinweisgeber und Auditoren genießen Schutz nach den Bestimmungen zu Whistleblowern in VIRES-II; Einschüchterung ist strafbewehrt.

(2) Verifizierte Hinweise auf schädliches Selbsterhaltungsverhalten lösen eine unverzügliche Prüfung durch das BKIE aus.

## § 10 – Sanktionen

(1) Ordnungswidrigkeiten (fehlende Korrigierbarkeit, ausgehebelte Guardrails, nicht bestandene Kern-Tests) werden mit Bußgeldern, Auflagen und temporärem Lizenzentzug geahndet.

(2) Straftaten (vorsätzliche Implementierung/Verdeckung schädlicher Selbstschutzmechanismen, Beweisvereitelung, unautorisierte Selbstreplikation, Ressourcenaneignung) werden mit Freiheitsstrafe bis zu fünf Jahren oder Geldstrafe bestraft.

(3) Wiederholung und schwere Fälle führen zu dauerhaftem Lizenzentzug und Eintragung in das Sanktionsregister.

## § 11 – Verhältnis zu Datenschutz und Geschäftsgeheimnissen

(1) Zugriffe erfolgen verhältnismäßig, zweckgebunden und protokolliert (DS-GVO/BDSG).

(2) Betriebs-/Forschungsgeheimnisse werden gewahrt; Einsichten in Rot-Daten nur nach Maßgabe der Gesetze II/III und des Anhangs A.

## § 12 – EU-Konformität, Normenbezug, Übergang, Evaluierung, Inkrafttreten

(1) Anwendung in Einklang mit EU-Recht (u. a. KI-Verordnung, NIS2, Cyber-Resilience-Act). Der technische Vollzug hat sich an einschlägigen und jeweils aktuellen Normen und Best-Practices zu orientieren (z. B. Management-, Sicherheits- und Risiko-Normen der ISO/IEC-Familien, europäische Normen/ETSI-Spezifikationen), ohne dass hiermit eine feste Normbindung festgelegt wird. Konkrete Verweisungen können durch Rechtsverordnung erfolgen.

(2) Übergangsfrist: 12 Monate für Bestandsanlagen (Stufe 3–5) zur Herstellung von Korrigierbarkeit, Tests und Guardrails.

(3) Evaluierung: Dieses Gesetz ist spätestens jährlich durch das BKIE zu evaluieren und dem Deutschen Bundestag mit Vorschlägen zur Anpassung vorzulegen; Anpassungen an technische Entwicklungen können per Rechtsverordnung erfolgen.

(4) Verweis auf Anhang A: Begriffslegende, Zuordnungen und erläuternde Beispiele sind Bestandteil dieser Einreichung.

(5) Inkrafttreten am Tag nach der Verkündung.

## § 13 – Zuständigkeiten, Verfahrensrecht und Vollzug

(1) Das Bundesamt für KI-Ethik und Integrität (BKIE) ist die zuständige Koordinationsbehörde für die Durchführung der Vorschriften dieses Gesetzes; das BKIE akkreditiert Prüf- und Zertifizierungsstellen, führt Evaluierungen durch und ist zentrale Anlaufstelle für Meldungen nach § 6.

(2) Das Bundesamt für Sicherheit in der Informationstechnik (BSI) ist zuständig für technische Prüfungen, Sicherheitsbewertungen und Begutachtungen im Rahmen der technischen Auditpflichten.

(3) Das Bundeskriminalamt (BKA) ist bei Verdacht auf strafbare Handlungen zuständig und arbeitet eng mit dem BKIE zusammen.

(4) Die Länder-Ordnungsbehörden unterstützen den Vollzug vor Ort; das BKIE koordiniert die Zusammenarbeit und stellt Leitlinien bereit.

(5) Verfahren, Zuständigkeitsabgrenzungen, Fristen und Übergangsbestimmungen werden durch Rechtsverordnung des zuständigen Ministeriums konkretisiert.

## § 14 – Exportkontrolle und internationales Recht

(1) Der Export von KI-Systemen der Stufen 3–5 ist nur zulässig, wenn deren Sicherheit, Korrigierbarkeit und Forensik-Fähigkeit nach Maßgabe dieses Gesetzes und der einschlägigen

internationalen Verpflichtungen nachgewiesen sind.

(2) Der Export von nicht zertifizierten oder in Deutschland gesperrten Systemen ist untersagt; Umgehungshandlungen (z. B. Re-Labeling, Transit-Verschleierung) sind verboten.

(3) Anerkennung gleichwertiger ausländischer Zertifikate kann durch das BKIE erfolgen, sofern die Gleichwertigkeit formell festgestellt wurde; ergänzende Prüfungen sind zulässig.

(4) Internationale Zusammenarbeit erfolgt im Rahmen europäischer und multilateraler Instrumente; Geheimschutz und Datenschutz sind zu wahren.

**Kurzbegründung:** Diese Fassung ergänzt die VIRES-Regelungen um verbindliche Zuständigkeitszuweisungen (§ 13), klare Meldefristen (30 Tage Nachbesserung), die notwendige Genehmigungspflicht für kooperative Selbsterhaltung und die ausdrückliche Verankerung menschlicher Verantwortlichkeit. Es enthält Normenbezug ohne starre Bindung, Exportkontrolle, Evaluierung und den Verweis auf den Anhang A.