

Richtlinie zur Einhaltung des Urheberrechts für Allgemeine KI-Modelle

Wirksamkeitsdatum: 00.00.0000

1. Begriffsbestimmungen

In dieser Richtlinie zur Einhaltung des Urheberrechts haben die nachstehenden Begriffe die folgende Bedeutung:

„**Wir**“ oder „**Unser**“ bezeichnet [UNTERNEHMENSNAME].

„**Allgemeines KI-Modell**“ oder „**GPAI-Modell**“ hat die gleiche Bedeutung wie in Verordnung (EU) 2024/1689, Artikel 3(63) (KI-Verordnung) definiert.

„**Rechtmäßig zugängliche Inhalte**“ bezeichnet Inhalte, die rechtmäßig zugänglich sind, ohne technische Schutzmaßnahmen zu umgehen.

„**Rechtevorbehalt**“ bezeichnet den Ausdruck von Rechteinhabern, der Text-/Daten-Mining oder die Wiederverwendung mittels maschinenlesbarer Verfahren untersagt, z. B. (aber nicht ausschließlich) robots.txt Dateien.

„**Liste wiederholter Rechtsverletzer**“ bezeichnet eine Liste von Websites, die von EU-/EWR-Behörden oder Gerichten als wiederholt an kommerziellen Urheberrechtsverletzungen beteiligt identifiziert wurden.

2. Zweck

Diese Urheberrecht Richtlinie legt die Governance, Verfahren und technischen Kontrollen fest, die erforderlich sind, um sicherzustellen, dass die **Allgemeinen KI-Modelle (GPAI)** von [UNTERNEHMENSNAME] in voller Übereinstimmung mit dem Urheberrecht der Europäischen Union und der Mitgliedstaaten entwickelt, trainiert, bereitgestellt und gewartet werden, insbesondere (aber nicht ausschließlich) gemäß Richtlinie 2001/29/EG, Richtlinie (EU) 2019/790 und Richtlinie 2004/48/EG sowie nach Artikel 53 der Verordnung (EU) 2024/1689.

Dieses Werk wurde unter der Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication in die Public Domain freigegeben.

3. Geltungsbereich

Diese Richtlinie gilt für alle unsere GPAI-Modelle, die auf dem Binnenmarkt der EU platziert oder verfügbar gemacht werden. Sie gilt ferner für alle unsere Datensätze, Werkzeuge und Verfahren, die für das Training, Fine-Tuning, die Evaluierung oder Inferenz der Modelle verwendet werden, und ist für alle unsere Mitarbeiter, Auftragnehmer und Drittpartner verbindlich, die an der Entwicklung oder dem Betrieb von GPAI beteiligt sind.

4. Nutzung ausschließlich rechtmäßig zugänglicher Inhalte

Unsere Webcrawler und/oder andere Technologien greifen nur auf Inhalte zu, reproduzieren oder extrahieren ausschließlich Inhalte, die rechtmäßig zugänglich sind, und umgehen keine technischen Schutzmaßnahmen, wie sie von Rechtsinhabern durch Zugangskontroll- oder Schutzmechanismen (z. B. Verschlüsselung, Scrambling oder andere Transformationen des Werkes oder Kopierschutzmechanismen) implementiert werden, die dem in Artikel 6(3) der Richtlinie 2001/29/EG definierten Schutzzweck dienen.

[UNTERNEHMENSNAME] wird vom Crawlen, Zugreifen und Indexieren zu Trainingszwecken seiner GPAI-Modelle alle Websites ausschließen, die von EU- oder EWR-Gerichten oder Behörden als wiederholte gewerbliche Urheberrechtsverletzer anerkannt sind.

5. Einhaltung von Rechtevorbehalten

Unsere Webcrawler und/oder anderen Technologien respektieren Robot-Exclusion-Protokolle (robots.txt), wie sie in der Internet Engineering Task Force (IETF) Request for Comments Nr. 9309 und ähnlichen Standardprotokollen festgelegt sind, die Rechtevorbehalte im Sinne von Artikel 4(3) der Richtlinie (EU) 2019/790 ausdrücken. Wir werden außerdem andere maschinenlesbare, dem Stand der Technik entsprechende Protokolle für Rechtevorbehalte beobachten und übernehmen, sobald sie aus Standardisierungsbemühungen hervorgehen, und öffentlich Informationen über die eingesetzten Webcrawler und/oder Technologien, deren robots.txt-Verhalten und sonstige Mechanismen zur Identifizierung und Einhaltung von Rechtevorbehalten offenlegen.

6. Unberührtheit von Rechtevorbehalten

Nichts in dieser Richtlinie zur Einhaltung des Urheberrechts darf so ausgelegt werden, dass es die Rechte der Rechtsinhaber einschränkt oder beeinträchtigt, die Nutzung ihrer Werke oder anderer geschützter

Dieses Werk wurde unter der Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication in die Public Domain freigegeben.

Inhalte ausdrücklich vorzubehalten, sei es durch maschinenlesbare Maßnahmen für Online-Inhalte oder durch andere rechtmäßige Mechanismen. Dies gilt auch für Inhalte, die von Dritten gesammelt oder gecrawlt und möglicherweise von uns genutzt werden.

7. Abmilderung rechtsverletzender Ausgaben

Die Dokumentation, welche unseren Open-Source-Modellen beigefügt ist, weist die Nutzer ausdrücklich darauf hin, dass urheberrechtsverletzende Nutzungen untersagt sind.

8. Ansprechpartner & Beschwerdeeinreichung

Die benannte Kontaktperson für Urheberrechtsinhaber, welche uns wegen potenzieller Verletzungen erreichen möchten, ist:

HERR / FRAU XYZ

[UNTERNEHMENSNAME]

ADRESSE

E-MAIL-ADRESSE

STADT / LAND

Rechtsinhaber und ihre bevollmächtigten Vertreter, einschließlich Verwertungsgesellschaften, können ausreichend präzise und hinreichend begründete Beschwerden über unsere angebliche Nichteinhaltung über das vorgesehene Kontaktformular einreichen, das auch den Upload weiterer Dokumente ermöglicht. Wir werden eingehende Beschwerden sorgfältig, unparteiisch und innerhalb einer angemessenen Frist bearbeiten. Beschwerden, die offensichtlich unbegründet sind oder bereits identisch beantwortet wurden, werden ignoriert.

9. Haftungsausschluss zur Rechtskonformität

Wir erkennen an, dass diese Richtlinie die Einhaltung des EU-Urheberrechts oder der nationalen Urheberrechtsgesetze unterstützt, diese jedoch nicht ersetzt. Die tatsächliche Rechtskonformität bleibt unsere Verantwortung.

Dieses Werk wurde unter der Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication in die Public Domain freigegeben.

10. Verhältnismäßigkeit und Anwendbarkeit

Unsere Maßnahmen sind dem Umfang und der Art unserer Geschäftstätigkeit angemessen. Durch die Umsetzung dieser Richtlinie wollen wir uns an den freiwilligen, aber rechtlich bedeutsamen Erwartungen orientieren, die im EU-Verhaltenskodex für GPAI-Modelle dargelegt sind.

ANNEX (SAMPLE)

Crawler & Rights Reservation Detection

A. Crawler Configuration Requirements

1. Crawler Identification

- All crawlers must:
 - Use a **unique User-Agent string** identifying [Company Name] and a contact URL/email.
 - Publish crawler technical specifications and robots.txt compliance behavior at a public endpoint (e.g., [https://\[company\]/crawler-info](https://[company]/crawler-info)).

Example (HTTP request header):

User-Agent: CompanyCrawler/1.2 (+<https://example.com/crawler-info>; contact@example.com)

2. Crawl Scope Enforcement

- Implement a **domain allowlist** and **infringer blocklist** at the crawler configuration level.
- Blocklist is synchronized daily from:
 - EUIPO or national IP authority “persistent infringer” lists.
 - Any internal takedown or complaint resolutions that require source blocking.

Example configuration:

crawler:

```
obey_robots_txt: true
allow_domains:
- example.edu
- example.gov
- example.org
block_domains:
- infringer-site1.com
- infringer-site2.net
max_depth: 3
rate_limit: 1 req/sec
```

3. Access Control Compliance

- Crawler must refuse access to:
 - Pages requiring authentication, CAPTCHAs, or tokens.
 - URLs returning HTTP 401, 402, or 403.
- Crawler may not employ headless browser automation to bypass restrictions without explicit license.

B. Rights-Reservation Detection Workflow

1. Primary Signal: robots.txt Parsing

- Detect User-agent: * rules with Disallow: / or specific path rules.
- Parse X-Robots-Tag HTTP header for nodm, noindex, or equivalent tags.

Sample Python snippet:

```
import requests
from urllib.parse import urljoin

def check_rights_reservation(domain):
    robots_url = urljoin(domain, '/robots.txt')
    resp = requests.get(robots_url, timeout=5)
    if resp.status_code == 200 and ('noai' in resp.text.lower() or 'disallow' in resp.text.lower()):
        return True
    return False
```

2. Secondary Signal: Embedded Metadata

- Check page HTML and media metadata for:
 - IPTC “AI Data Mining” prohibition flags.
 - C2PA manifest claims with “No AI Training” directives.
 - Dublin Core rights and license fields.

Example JSON from C2PA manifest:

```
{
  "assertions": [
    {
      "label": "c2pa.datamining",
      "value": "prohibited"
    }
  ]
}
```

3. Storage & Logging

- Log all rights-reservation detections with:
 - Timestamp
 - URL
 - Detected signal type (robots.txt, HTTP header, metadata)
 - Crawling decision (allow/deny)
- Retain logs for **5 years** in an immutable audit store.

C. API Endpoints for Rights-Reservation Verification

1. Check Single URL

GET /api/v1/rights-reservation?url={url}

Response:

```
{
  "url": "https://example.com/article",
  "rights_reservation": true,
  "signals": ["robots.txt:nodm", "X-Robots-Tag:noindex"]
}
```

2. Bulk Check

POST /api/v1/rights-reservation/bulk

Body:

```
{
  "urls": [
    "https://site1.com/page",
    "https://site2.com/page"
  ]
}
```

D. Audit & Review Procedures

- **Daily:** Automated scan of updated rights-reservation protocols from standardization bodies.
- **Quarterly:** Manual audit of crawler logs to confirm:
 - No overridden blocklist entries.
 - No training ingestion from rights-reserved sources.

Annually: Third-party technical audit with reproducible test URLs.