Copyright Compliance Policy for General-Purpose AI Models

Effective Date: 00.00.0000

1. Definitions

In this Copyright Compliance Policy, the below terms have the following meanings:

" **Our**" or "**We**" means COMPANY NAME.

"G**eneral Purpose AI**" or "**GPAI**" model has the same meaning as defined in Regulation (EU) 2024/1689, Article 3(63) (AI Act).

"**Lawfully Accessible Content**" means content that can be accessed legally without circumventing technological protection measures.

" **Rights Reservation**" means an expression by rightholders prohibiting text/data mining or reuse by machine-readable means, such as but not limited to robots.txt files.

" **Persistent Infringer List**" means a list of sites identified by EU/EEA authorities or courts as repeatedly engaging in commercial-scale infringement.

2. Purpose

This policy establishes the governance, procedures, and technical controls necessary to ensure that [COMPANY NAME]'s General-Purpose AI (GPAI) models are developed, trained, deployed, and maintained in full alignment with European Union and EU Memberstate copyright laws, in particular but not limited to Directive 2001/29/EC, Directive (EU) 2019/790 and Directive 2004/48/EC, as per Regulation (EU) 2024/1689, Article 53.

3. Scope

This policy applies to all our GPAI models placed on or made available in the EU single market. It further applies to all our datasets, tools, and processes used for model training, fine-tuning, evaluation, or inference and is binding for all our personnel, contractors, and third-party partners engaged in GPAI development or operation.

4. Use of Lawfully Accessible Content Only

Our web-crawlers and / or other technologies will only access, reproduce or extract content that is lawfully accessible and not circumvent technological protection measures, such as those controlled by the rightholders through the application of an access control or protection process like encryption, scrambling or other transformation of the work or other subject matter or other copy control mechanism, which achieves the protection objective, as defined in EU Directive 2001/29, Article 6(3).

[COMPANY NAME] will exclude from crawling, accessing and indexing for its GPAI model training purposes any websites that are recognized by EU or EEA courts or public authorities as Persistent Infringers of copyright on a commercial scale.

5. Rights Reservations Compliance

Our web-crawlers and /or other technologies will respect robot exclusion protocols (robots.txt) as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309 and similar standards-protocols expressing rights reservations per Directive (EU) 2019/790, Article 4(3). We will further monitor and adopt other machine readable, state-of-the-art protocols for rights reservation as they emerge from standardization efforts and publicly disclose information about the web-crawlers and /or other technologies used, the robots.txt behaviors, and any other relevant mechanisms to identify and comply with rights reservations.

6. Rights Reservation Unaffected

Nothing in this Copyright Compliance Policy shall be construed as limiting or prejudicing the rights of rightholders to expressly reserve the use of their works or other protected subject matter by any means, including, without limitation, through machine-readable measures for online content or by any other lawful mechanism.This shall include content scraped or crawled by third parties potentially used by us.

7. Mitigation of Infringing Outputs

Documentation accompanying our open source models alerts users that copyright infringing uses are prohibited.

8. Point of Contact & Complaint Submission

The designated point of contact for copyright holders to reach our organization about potential infringements is:

MR / MRS XYZ
COMPANY NAME
ADDRESS
EMAIL ADDRESS
CITY / COUNTRY

Rightsholders and their authorised representatives, including collective management organisations, may submit sufficiently precise and adequately substantiated complaints concerning our alleged non-compliance by using the designated contact form, which allows for the upload of further documents. We will act upon complaints received in a diligent, non-arbitrary manner and within a reasonable time. A complaint that is manifestly unfounded or has already been responded to an identical complaint will be ignored.

9. Legal Compliance Disclaimer

We acknowledge that this policy supports, but does not replace, compliance with EU copyright law or national copyright legislation. Actual legal compliance remains our responsibility.

10. Proportionality and Applicability

Our measures are proportionate to the size and nature of our operations. By implementing this policy, we aim to align with the voluntary yet legally significant expectations laid out in the EU Code of Practice for GPAI models.

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).

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.