Wat kun je doen om je werk zoveel mogelijk te beschermen tegen ongewenst gebruik door AI?

Als maker van een werk heb je daarop meestal het auteursrecht. Dit betekent dat jouw toestemming nodig is om je werk te gebruiken. Of dit ook geldt voor de training van AI-systemen, is op dit moment geen uitgemaakte zaak.

De Auteurswet regelt in ieder geval dat je als maker mag 'opt-outen' voor AI's met commerciële doeleinden. Je geeft hiermee aan dat je werk niet mag worden gebruikt voor het trainen van AI. De wet geeft geen toelichting over wat de opt-out inhoudt en hoe dit eruit moet zien, alleen dat deze machinaal leesbaar moet zijn. Hoe een opt-out precies moet worden gedaan, is op dit moment onduidelijk. Er is (nog) geen standaard voor.

Met onderstaande stappen voldoe je zo veel mogelijk aan de opt-out zoals in de wetgeving is genoemd. Mocht er in de toekomst meer over het auteursrecht en de manier van opt-outen bekend worden, dan laten we dat hier weten.

Blokkeer crawlers met Robots.txt

Robots.txt is een tekstbestand dat in de backend-code van een website kan worden geplaatst. Hiermee wordt crawlers verteld wat ze wel en niet kunnen scannen.

Je kunt robots.txt gebruiken om het *crawlen* te beperken of volledig te blokkeren. Je kiest er zelf voor welke bots je uitsluit. Google maakt ook gebruik van *crawling*. Om te voorkomen dat je onvindbaar bent voor zoekmachines zoals Google, is dus het belangrijk goed op te letten welke bots je kiest.

Robots.txt is afkomstig uit het Robot Exclusion Protocol, een richtlijn om websites af te schermen tegen bots. Het protocol is slechts een richtlijn en dus niet dwingend. Dit heeft tot gevolg dat een bot de richtlijn kan naleven, maar ook kan zijn geprogrammeerd om dat niet te doen. Kwaadwillende bots zullen het robots.txt-bestand negeren. Maar door het robots.txt-bestand toe te voegen, maak je je voorbehoud expliciet en machinaal leesbaar en zet je een goede stap.

<u>Hoe gebruik je robots.txt?</u>

Als je een websitebouwer hebt, vraag dan of deze een robots.txt-bestand toevoegt. Beheer je je eigen website, voeg dit bestand dan toe aan de *root* van je website. Zie hiervoor het stappenplan in Bijlage I. Grote websitebouwers hebben handleidingen beschikbaar gesteld over hoe je de robots.txt installeert.[1]

De voorbeelden in Bijlage II geven je inzicht in welke bots grote websites zoal uitsluiten.

Voor de meer gevorderde websitebouwers kan er ook worden gekeken naar de initiatieven van W3C (TDM-protocol)[2] en het integreren van NoAI-tags in HTML.[3]

Meld je af voor AI trainingsdatasets

---

[1] **Wordpress**: https://kinsta.com/blog/wordpress-robots-txt/
Joomla:https://docs.joomla.org/Robots.txt_file
Wix: https://support.wix.com/nl/article/het-robotstxt-bestand-van-je-website-bewerken
Webflow: https://university.webflow.com/lesson/disable-search-engine-indexing#generating-a-robots-txt-file
[2] https://www.w3.org/2022/tdmrep/
[3] https://help.raptive.com/hc/en-us/articles/13764527993755-NoAI-Meta-Tag-FAQs

Op de website Have I Been Trained (www.haveibeentrained.com) kun je niet alleen controleren of jouw afbeelding is gebruikt voor het trainen van datasets, maar deze hiervoor ook afmelden. Met deze opt-out wordt je werk niet meegenomen in de volgende trainingsrondes.

Have I Been Trained doorzoekt de LAION datasets. Dit zijn de datasets achter grote AI systemen als Stable Diffusion en Imagen. Ook is dit een open software database waarvan veel AI-ontwikkelaars gebruik maken.

Have I Been Trained neemt nieuwe datasets op zodra deze worden vrijgegeven en werkt zo veel mogelijk samen met andere organisaties die afbeeldingslinks verzamelen. Zij hoopt op termijn als eenmalige opt-out tool te kunnen dienen.

Ook DALL-E 3 (https://openai.com/dall-e-3) kan worden gevraagd afbeeldingen uit de trainingsgegevens te verwijderen. DALL-E 3 is een AI systeem van Open AI, het AI-bedrijf dat ook zit achter Chat GPT. Helaas is dit (nog) een vrij lastig en arbeidsintensief opt-out proces.

Hou er rekening mee dat de opt-out slechts geldt voor deze datasets waar je deze inzet. Andere AI-systemen, zoals Midjourney, werken met een eigen dataset. Bovendien is het onzeker of eenmaal geleerde informatie echt uit de dataset kan worden verwijderd. Wanneer een werk eenmaal in een dataset zit, is het moeilijk om deze weer te ‘ontleren’.

Manipuleer je afbeelding

Onderzoekers van de Universiteit van Chicago ontwikkelden een applicatie genaamd Glaze. Deze verandert de afbeelding zodanig dat AI-systemen er geen bruikbare informatie uit kunnen halen. Voor bezoekers blijft de afbeelding bijna hetzelfde maar voor AI is de afbeelding onbruikbaar. Glaze is te downloaden via de volgende link: https://glaze.cs.uchicago.edu/downloads.html.

Een vergelijkbare tool, Mist, is te downloaden op: https://mist-project.github.io/index_en.html.

Deze tools zijn ook in te zetten bij het plaatsen van afbeeldingen op sociale media.

Plaats een AI-clausule op je website

Het plaatsen van een clausule op je website, waarin je aangeeft dat je je auteursrecht voorbehoudt en dat de gegevens niet door AI-systemen mogen worden gebruikt, is een manier om je met behulp van een opt-out op je auteursrecht te beroepen. Je plaatst hiervoor op een duidelijk zichtbare plek een voorbehoud voor de bezoekers daarvan.

Hoewel het de vraag is of een bot dit kan lezen, geeft het wel aan dat jij als rechthebbende niet wil dat jouw werk voor AI wordt gebruikt. Ook kan het van belang zijn voor je rechtspositie als in de toekomst meer duidelijk wordt over het gebruik van auteursrechtelijk beschermd werk voor het trainen van AI.

Een voorbeeld: “© Copyright reserved. No automated text and data mining is permitted on this website.”

BIJLAGE I – STAPPENPLAN

Voor een uitgebreide uitleg zie: https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt

Stap 1:

Controleer of je een robots.txt-bestand hebt door je domeinnaam in te voeren gevolgd door /robots.txt. Bijvoorbeeld: https://kunstenbond.nl/robots.txt

Stap 2:

Als je nog geen robots.txt-bestand hebt op je website en de websitebouwer biedt dit niet aan, kun je zelf een robots.txt-bestand creëren. Open hiervoor Notepad (Windows) of TextEdit (Mac).

Stap 3:

In je robots.txt-bestand kun je specifieke bots uitsluiten. Om de bots van OpenAI uit te sluiten, kun je de volgende tekst gebruiken:

```
User-agent: GPTBot
Disallow: /

User-agent: CCBot
Disallow: /
```

Sla het bestand op als robots.txt, in kleine letters. Het robots.txt-bestand moet worden opgeslagen met een UTF-8 codering.

Stap 4:

Upload het bestand naar de *root* van je website. Voor het uploaden van het robots.txt-bestand is geen standaardadvies, dit is namelijk sterk afhankelijk van de architectuur van de website. Neem contact op met de hosting company van je website of zoek op het internet hoe je dit moet doen.

LET OP: Elke dag verschijnen er nieuwe bots en sommige AI's geven niet vrij welke bots ze gebruiken. Daarom is het essentieel om ook de andere stappen te volgen.

BIJLAGE II - VOORBEELDEN

Ter inspiratie zijn hieronder voorbeelden verzameld van een aantal grote websites die een robots.txt-bestand hebben toegevoegd aan hun websites. Deze sites kun je ook gebruiken om up-to-date te blijven met je robots.txt-bestand.

Amazon, https://www.amazon.com/robots.txt:

```
User-agent: EtaoSpider
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: CCBot
Disallow: /
```

New York Times, https://www.nytimes.com/robots.txt:

```
User-agent: CCBot
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: ia_archiver
Disallow: /

User-Agent: omgili
Disallow: /

User-Agent: omgilibot
Disallow: /
```

The Guardian, https://www.theguardian.com/robots.txt:

```
User-agent: NewsNow
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: CCBot
Disallow: /
```

Shutterstock, https://www.shutterstock.com/robots.txt:

```
# Disallow User-Agents
User-agent: CCBot
Disallow: /

# Disallow public gptbot
User-agent: GPTBot
Disallow: /
```

Tumblr, https://www.tumblr.com/robots.txt:

```
# OpenAI's crawler
User-agent: GPTBot
Disallow: /

# Common Crawl's cralwer
User-agent: CCBot
Disallow: /

# SentiBot's cralwer
User-agent: sentibot
Disallow: /
```

Vimeo, https://vimeo.com/robots.txt:

```
User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: CCBot
Disallow: /
```

Pixabay, https://pixabay.com/robots.txt:

```
User-agent: 008
User-agent: MJ12bot
User-agent: sitebot
User-agent: dotbot
User-agent: AhrefsBot
User-agent: Ocelli
User-agent: sistrix
User-agent: ShopWiki
User-agent: WBSearchBot
User-agent: Riddlerbot
User-agent: linguatools
User-agent: www.integromedb.org/Crawler
User-agent: CCBot
Disallow: /
```

NRC, https://www.nrc.nl/robots.txt

```
User-agent: *
Disallow: /api/wordproof/
Disallow: /data/
Disallow: /de/data/s3/
Disallow: /login/
Disallow: /nieuwsbrieven/preview/
Disallow: /paywall-api/
Disallow: /search/

User-agent: CCBot
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: Google-Extended
Disallow: /
```

Volkskrant, https://www.volkskrant.nl/robots.txt

```
# Alle auteurs-, naburige en databankrechten die op de inhoud
en opmaak van de DPG Media websites
# en DPG Media apps rusten, worden door DPG Media BV
uitdrukkelijk voorbehouden. De inhoud van de
# DPG Media websites en apps is uitsluitend voor persoonlijk,
niet-commercieel gebruik en het is
# niet toegestaan om gegevens van de website of uit de apps
door middel van screen scraping
# (of een andere geautomatiseerde werkwijze) te vergaren.
# Zie ook de Gebruikersvoorwaarden van DPG Media B.V. op
www.dpgmedia.nl/gebruiksvoorwaarden

# All copyrights, neighbouring rights and database rights in
the content and layout of the
# DPG Media websites and DPG Media apps are explicitly reserved
by DPG Media BV. The content of the DPG Media
# websites and DPG Media apps is for personal, non-commercial
use only and it is not allowed to
# collect data from the website or from the apps by means of
screen scraping (or any other
# automated method).
# See also the terms of use of DPG Media B.V. at
www.dpgmedia.nl/gebruiksvoorwaarden

User-agent: Twitterbot
Allow: /

User-agent: *
Allow: /
Disallow: /*?otag*
Disallow: /*?abo_type*
Disallow: /*?URL_referrer*
```

```
Disallow: /auth/
Disallow: /temptation/
Disallow: /*utm_campaign=shared_earned*
Disallow: /*redirectUri=*
Disallow: /zoeken?query=*
Disallow: /search?query=*

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: CCBot
Disallow: /

User-agent: anthropic-ai
Disallow: /
```

De Groene Amsterdammer, https://www.groene.nl/robots.txt

```
User-agent: *
Disallow: /*?*q=*
Disallow: /campagnes/
Disallow: /abonneren/
Disallow: /agenda/182
Disallow: /agenda/183
Disallow: /agenda/184

User-agent: CCBot
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: anthropic-ai
Disallow: /
```

Vogue, https://www.vogue.com/robots.txt

```
User-agent: CCBot
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: LinkCheck by Siteimprove.com
Disallow: /
```

Depositphotos, https://depositphotos.com/robots.txt:

```
User-agent: Charlotte
Disallow: /
User-agent: Speedy
Disallow: /
User-agent: 008
Disallow: /
User-agent: sitebot
Disallow: /
User-agent: MJ12bot
Disallow: /
User-agent: sistrix
Disallow: /
User-agent: ShopWiki
Disallow: /
User-agent: WBSearchBot
Disallow: /
User-agent: Riddlerbot
Disallow: /
User-agent: linguatools
Disallow: /
User-agent: www.integromedb.org/Crawler
Disallow: /
User-agent: CCBot
Disallow: /
User-agent: SemrushBot
Disallow: /
User-agent: SemrushBot-SA
Disallow: /
User-agent: SemrushBot-BA
Disallow: /
User-agent: SemrushBot-SI
Disallow: /
User-agent: SemrushBot-SWA
Disallow: /
User-agent: SemrushBot-CT
Disallow: /
User-agent: SemrushBot-BM
Disallow: /
```