

Nightshade: KI-Modelle vergiften, eigene Werke schützen

[Originalartikel](#)

[Backup](#)

```
<html> <figure class=„aufmacherbild“><img
src=„https://heise.cloudimg.io/width/700/q75.png-lossy-75.webp-lossy-75.foil1/_www-heise-de_/imgs/
18/4/5/2/9/6/6/0/shutterstock_1146470141-6c0fdbcb0f0f9ce9.jpeg“
srcset=„https://heise.cloudimg.io/width/700/q75.png-lossy-75.webp-lossy-75.foil1/_www-heise-de_/im
gs/18/4/5/2/9/6/6/0/shutterstock_1146470141-6c0fdbcb0f0f9ce9.jpeg 700w,
https://heise.cloudimg.io/width/1050/q75.png-lossy-75.webp-lossy-75.foil1/_www-heise-de_/imgs/18/4/
5/2/9/6/6/0/shutterstock_1146470141-6c0fdbcb0f0f9ce9.jpeg 1050w,
https://heise.cloudimg.io/width/1500/q75.png-lossy-75.webp-lossy-75.foil1/_www-heise-de_/imgs/18/4/
5/2/9/6/6/0/shutterstock_1146470141-6c0fdbcb0f0f9ce9.jpeg 1500w,
https://heise.cloudimg.io/width/2300/q75.png-lossy-75.webp-lossy-75.foil1/_www-heise-de_/imgs/18/4/
5/2/9/6/6/0/shutterstock_1146470141-6c0fdbcb0f0f9ce9.jpeg 2300w“ width=„5155“ height=„2897“
sizes=„(min-width: 80em) 43.75em, (min-width: 64em) 66.66vw, 100vw“ alt=„Gift in einer Flasche
mit Zettel, auf dem ein Totenkopf gemalt ist“ class=„img-responsive“ referrerpolicy=„no-referrer“
/><figcaption class=„akwa-
caption“>(Bild:&#160;busliq/Shutterstock.com)</figcaption></figure><p><strong>Mit Nightshade
lassen sich die Trainingsdaten gro&#223;er KI-Modelle vergiften. Sie machen beispielsweise aus einer
Katze einen Hund.</strong></p><p>KI-Systeme wie Bildgeneratoren sind mit frei
zug&#228;nglichem Material aus dem Internet trainiert worden. Viele K&#252;nstler sehen dadurch
ihr Urheberrecht verletzt. Nun gibt es ein Tool, mit dem sie sich nicht nur abschotten, sondern sogar
den KI-Modellen schaden k&#246;nnen. Nightshade ver&#228;ndert Werke so, dass es f&#252;r das
menschliche Auge nicht zu erkennen ist. In den Trainingsdaten wird dadurch aber aus einer Katze
potenziell ein Hund oder aus einer Kuh ein Portemonnaie.</p><p><a
href=„https://www.heise.de/news/Gift-fuer-Trainingsdaten-Neues-Tool-soll-Bilder-vor-KI-Bildgenerator
en-schuetzen-9343354.html“><strong>Vorgestellt wurde Nightshade bereits im vergangenen Herbst
[1]</strong></a>, nun ist das Werkzeug auch tats&#228;chlich verf&#252;gbar. Entwickelt wurde
es an der Universit&#228;t von Chicago. Die Wissenschaftler haben bereits ein &#228;hnliches Tool
namens Glaze im Angebot. Wobei der Vorg&#228;nger darauf abzielt, dass K&#252;nstler ihre Werke
sch&#252;tzen k&#246;nnen. Konkret, in dem sie ihre Bilder so verf&#228;lschen, dass KI-Modelle
nicht in der Lage sind, ihren Stil zu lernen oder gar wiederzugeben. Daf&#252;r werden Pixel
ver&#228;ndert &#8211; f&#252;r das menschliche Auge unsichtbar, f&#252;r die Maschine
entsteht jedoch etwas Anderes. Die Entwickler sehen es als gegeben, dass sich KI-Anbieter teilweise
dar&#252;ber hinwegsetzen, wenn es um <a
href=„https://www.heise.de/news/Crawler-fuer-Training-von-GPT-ausschliessen-OpenAI-stellt-Code-fue
r-robots-txt-vor-9237417.html“><strong>Opt-Out-Listen und Hinweise geht [2]</strong></a>, dass
Inhalte nicht f&#252;r das Training genutzt werden d&#252;rfen.</p><h3 class=„subheading“
id=„nav_mit_nightshade0“>Mit Nightshade wird die Kuh zum Portemonnaie</h3><p><a
href=„https://nightshade.cs.uchicago.edu/whatis.html“ rel=„external noopener“
target=„_blank“><strong>Nightshade soll das ganze etwas weitertreiben [3]</strong></a>. Nutzt
man dieses Tool, und kommen die Crawler trotz Verbot an den eigenen Werken vorbei, werden die
Trainingsdaten vergiftet. Zeigt das Bild n&#228;mlich ein Portemonnaie statt einer Kuh und ist
dennoch als Kuh ausgezeichnet, lernt das Modell, Portemonnaies zu zeigen, wenn jemand per Prompt
eigentlich eine Kuh sehen will. Das gesamte KI-Modell kann dadurch Schaden nehmen &#8211;
f&#252;r alle Nutzer. In Tests mit einer eigenen Version vom Bildgenerator Stable Difussion reichten
50 vergiftete Bilder aus, dass Hunde begannen komisch auszusehen, 300 manipulierte Bilder
```

später waren aus Hunden Katzen geworden. Um ein großes Modell zu schaffen, bedarf es laut der Entwickler deutlich mehr vergifteter Materialien. Damit ist Nightshade also eine Erweiterung von Glaze. Künstler können freilich selbst entscheiden, ob sie beide Tools einsetzen wollen oder nur eines der beiden. Wobei der gemeinsame Einsatz eine faire Chance für Anbieter von KI-Modellen darstellt, die sich zunächst an das Verbot halten können und die Bilder nicht in das Training fließen lassen können. Erst nach dem Bruch würde dann die Vergiftung starten. Ob beide Tools zusammen funktionieren, wird allerdings erst noch getestet. Nightshade soll dann eine Erweiterung von Webglaze werden, also der Webversion von Glaze. Nightshade soll auch als Open-Source zur Verfügung gestellt werden. Dann können freilich auch andere Menschen in anderen Settings versuchen, damit den KI-Anbietern zu schaden. Um die Software nutzen zu können, muss man sie zunächst installieren. Dafür, <https://nightshade.cs.uchicago.edu/userguide.html> so heißt es in einer Erklärung [4], müssen einige Machine-Learning-Bibliotheken und vortrainierte Modelle heruntergeladen werden. Grundsätzlich gibt es einige Bedingungen an die Hardware. Aktuell scheint es zu sehr langen Ladezeiten zu kommen. ()

URL dieses Artikels:

`https://www.heise.de/-9604398`

Links in diesem Artikel:

[1] `https://www.heise.de/news/Gift-fuer-Trainingsdaten-Neues-Tool-soll-Bilder-vor-KI-Bildgeneratoren-schuetzen-9343354.html`

[2] `https://www.heise.de/news/Crawler-fuer-Training-von-GPT-ausschliessen-OpenAI-stellt-Code-fuer-robots-txt-vor-9237417.html`

[3] `https://nightshade.cs.uchicago.edu/whatis.html`

[4] `https://nightshade.cs.uchicago.edu/userguide.html`

[5] `mailto:emw@heise.de`

Copyright © 2024 Heise Medien

From: <https://schnipsl.qgelm.de/> - Qgelm

Permanent link: https://schnipsl.qgelm.de/doku.php?id=wallabag:wb2nightshade_-ki-modelle-vergiften_-eigene-werke-schutzen

Last update: 2025/06/27 11:17

