European Writers' Council

Doing the Rights thing since 1977

LIBRARY OF CONGRESS

U.S. Copyright Office

101 Independence Ave. S.E.,
Washington, D.C. 20559-6000
[Docket No. 2023–6; COLC–2023–0006]
Artificial Intelligence and Copyright: Call for Comments
Submitted via: https://www.regulations.gov/commenton/COLC-2023-0006-0001

With reference to https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf as well as the extension of the deadline https://www.govinfo.gov/content/pkg/FR-2023-09-21/pdf/2023-20480.pdf

Brussels, 30th of October 2023

The European Writers' Council (EWC) is the world's only and largest representation of writers in the book sector and of all genres (fiction, non-fiction, academic, children's books, poetry, etc.). With 49 organisations and professional guilds from 31 countries of the EU, the EEA and of non-EU areas, the EWC represents 220.000 writers and translators.
These individuals write and publish in 34 languages, including in the original or as a translation in the US territories and corresponding legal frameworks. For them, as well as for the book writers affected worldwide and in Europe, who are directly and immediately hit by the consequences of so-called "AI" in the book sector and in particular by the production and use of generative informatics, we have a corresponding duty to respond on their behalf to this important initiative of the U.S. Copyright Office.

We thank you for the opportunity to comment and respond as follows according to the questionnaire, with a focus on synthetic, algorithm based automated text (re)generators ("GAI") in the trade book sector and its effects on book writers.

General Questions

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

(1a) For the book sector: Generative, analytical and assistive informatics, sub-areas of so-called artificial "intelligence", threatens numerous jobs and fields of application in the book sector. Synthetic text and image generators, synthetic robot voice cloning and algorithmic informatics to analyse text and data, will replace some professions by machines in the medium term; be it in the areas of text, editing, proofreading, production, cover design, illustration, translation, selection and editing of original and translated manuscripts, audio book production or in the promotion and distribution of books.
Already, numerous criminal and damaging "AI business models" have threatened the book sector - with fake authors, fake books and also fake readers. It has been demonstrated that the foundations for major language models such as GPT, Meta, StableLM, BERT have been generated from book works whose sources are shadow libraries such as Library Genesis (LibGen), Z-Library (Bok), Sci-Hub and Bibliotik – bit torrent piracy sites.

Without legal regulation, generative, algorithmic, synthetic, and reproducing technologies accelerate and enable the expansion of exploitation, legitimisation of copyright infringement, climate harm, discrimination, information and communication distortion, identity theft, reputational damage, blacklisting, royalty fraud and collective licensing remuneration fraud.

At the same time, a close look and assessment is needed to categorise and regulate the individual aspects of advanced informatics; because not all smart software is "AI", not every application is equally risky. Assisting or analysing software of advanced informatics are already widely used with the book sector – for accounting, logistics, meta data management, citation glossaries, anti-plagiarism control or editing workflow.

(1b) For the public or the users of applications: generative text informatics is a high-risk communicator and unreliable source of information. "Hallucinating" is the vocabulary used to describe generative text systems[1] that **completely invent or incorrectly plug together data, events[2], court decisions[3] or biographies**, contradict themselves when asked questions, or need to be constantly corrected by users with *reinforcement learning from human feedback* (RHLF)[4]. In the process, users teach the system what its developers did not. At the same time, generative text software makes it easier for actors such as propaganda farms to rapidly and cheaply spread disinformation and hate speech; and creates fake authors who flood social networking platforms or market players such as Amazon[5] with GPT output and artificial communication[6]. The lack of or inadequate security checks to save costs[7] and the lack of test and correction series prior to publication mean that **generative text informatics must be assessed as fundamentally untruthful.** At the same time, however, the "faith" and lack of sensitivity towards artificial content of many of the over 100 million users are so high that they do not recognise these hallucinations - or do not even suspect that the output is false or fictitious. Basically, GAI needs original, "fresh", human texts in order not to go crazy, as a study from Stanford University found out: If synthetic content (AI output) is used as training[8], the system collapses.

(1c) "GAI" (re)produces bias and reinforces intersectional discrimination [9, 10].
Stable Diffusion, an image-generating ("text to image") computer science, knows no black members of parliament, no female doctors, and poses as cleaners basically Asian women. Text (re-)generators reproduce sexist and gender-stereotypes - as they draw on material that comes from a particular more Western, male, white-oriented canon[11] or "learned" misogyny from the comment sections of the internet. A bias (false prejudice) can refer not only to gender or skin color; but to places, ages, social classes, professions, medical conditions, cultures or the classification of facts, of concepts such as "success" or "happiness" – and political opinion.
Effect: Users of a synthetic, algorithmic, (re)generating text software adopt the bias[12] and reinforce it. As a result, people are pigeonholed even more quickly and, above all,

[1] https://www.beamex.com/resources/for-a-safer-and-less-uncertain-world/generative-ai/
[2] https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html
[3] https://www.morningbrew.com/daily/stories/2023/05/29/chatgpt-not-lawyer?mbcid=31642653.1628960&mblid=407edcf12ec0&mid=964088404848b7c2f4a8ea179e251bd1&utm_campaign=mb&utm_medium=newsletter&utm_source=morning_brew
[4] https://www.telusinternational.com/insights/ai-data/article/rlhf-advancing-large-language-models
[5] https://www.vice.com/en/article/v7b774/ai-generated-books-of-nonsense-are-all-over-amazons-bestseller-lists
[6] https://www.independent.co.uk/tech/ai-author-books-amazon-chatgpt-b2287111.html
[7] https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html
[8] https://futurism.com/ai-trained-ai-generated-data
[9] https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf
[10] https://www.bloomberg.com/graphics/2023-generative-ai-bias/
[11] https://crfm.stanford.edu/2023/06/15/eu-ai-act.html
[12] https://www.nyu.edu/about/news-publications/news/2022/july/gender-bias-in-search-algorithms-has-effect-on-users--new-study-.html

unquestioningly, which can have an impact on social and professional access, education, housing, health care and creditworthiness.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

Yes. GAI harms human authors, their income and reputation through fake authors, fake books, fake readers - and identity theft:

(2a) Uncontrolled AI output is being pushed into the bestseller lists with click farms: For months, the global self-publishing service provider Amazon has been flooded with bogus books by fake authors whose text and visual content have been cobbled together by generative speech and image output software. AI bots from click farms "read" these nonsense works and push them into the bestseller lists[13]. This leads to a rapid decline in revenue for human writers fed by shared-revenue models, such as Kindle KDP and its distribution on a shared-revenue basis (A pot of revenue divided by pages read and distribution beneficiaries, similar to Spotify). At peak times, 80 out of 100 Kindle KDP bestsellers are AI editions.

(2b) Identity theft and name deception: The world's most important review platform, Goodreads, like Amazon, is flooded with GAI "books" published under the illegitimately used names of real human authors or slightly altered spellings of real known names. These "books" are listed as new releases in the authors' profiles and entice readers to buy them. However, the income from these fakes flows to unknown sources. Human authors who are cheated out of their earnings must in turn spend money to defend themselves with lawyers - and thus pay twice. So far, neither Goodreads nor Amazon have stopped this identity theft, which damages the reputation of human authors when a (low-quality) GAI book is associated with their name.

(2c) Unauthorised AI translations open up foreign-language markets and channel sales to unknown foreign users: We have cases of works being illegally transferred from, for example, the original English language into other languages by means of robot translation without a license and published under a different name, usually in Amazon Self-publishing and often still equipped with an AI-made cover. The author names, in turn, deliberately resemble well-known names. The revenues flow to unknown sources.

(2d) Publishing **services only against payment by the author:** Publishers are increasingly also producing AI-generated covers. **We have cases where authors requested human designers and were then asked to pay**. This practice is considered indecent. However, the authors, as weaker contractual partners, hardly have the courage to refuse this, out of well-founded fear of being considered "difficult" or of being rejected by publishers for future cooperation. They are pressured into accepting a technology that harms their own profession at its core.

(2e) Illegal remuneration claims to collective management organisations and media customers like press or photo stock: In any case, it cannot be ruled out that both automatically generated and machine translated (MT) press articles and automatically translated books, or even automatically produced images, already "enjoy" private copying remuneration from collective managements organisations (CMOs), as there is no legal labelling obligation yet; or automated texts, MT translations and AI-generated images flow into the media on a royalty basis.

[13] https://www.vice.com/en/article/v7b774/ai-generated-books-of-nonsense-are-all-over-amazons-bestseller-lists

(2f) Machine voices replace humans - and lead to the loss of license fees: DeepZen has been working on clone voices since 2013 and offers its repertoire to publishers to save on fees for human narrators; numerous publishers, including renowned ones, have already resorted to this. The dislocation continues in the question of revenue distribution: **if** there is no audiobook narrator where does his calculated share go? The job situation for professional narrators is sinking rapidly[14]. To professionally produce a voice clone (of human people) costs less than 2.000 Dollar in a professional studio. It is even cheaper with programmes like Murf, Lobo, Respeacher, Voice.Ai or Overdub. After a few seconds of recording, a voice clone is generated with which you can make "Anyone" say "Anything", no matter how immoral or fraudulent[15],[16].

In 2022, Google introduced its services for publishers in six countries, in early January 2023[17] **Apple** introduced a series of AI voices named such as Madison and Jackson. Authors and publishers who sell their books through Apple Books are supposed to make use of these (and sign a confidentiality clause to this effect).

The areas of application of clone voices or synthetic "voices" range from dubbing to audio books to trick calls for fraudsters or for deep-fake interviews etc. **Actors and audio book narrators are increasingly confronted with having to agree to voice cloning in contracts for work if they want to continue to be employed. This leads to the gradual elimination of narrators.** In addition, there are isolated cases in which voice clones were created without the consent of the human speakers. Or to be replaced by purely synthetic voices of advanced devices (example: "Tonie Box", where synthetic robot voices read automatically generated texts to children for goodnight[18]**). AI dubbing also becomes relevant when e-books are read aloud by devices and voice clones, but the author has neither granted a license nor receives remuneration.**

All in all, all these new "AI business models" lead to the following paradox: those who made the existence of generative programmes possible in the first place are not remunerated. But those who use the software profit monetarily. This transfer of value as a form of exploitation cannot be intended in a democracy with social and just values.

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

All relevant papers are mentioned within the footnotes, if one of the answers relate to them (studies, surveys, facts and findings).

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? [40] How important a factor is international consistency in this area across borders?

Yes, there is, in Europe, a legislative act that should be avoided in the United States. Within the Directive 2019/790 (EU) on Copyright and Relates Rights in the Digital Single Market

[14] https://www.voanews.com/a/7092661.html
[15] https://www.podcast.de/episode/609495902/deepfake-bei-anruf-klon
[16] https://www.deutschlandfunkkultur.de/audio-deepfakes-was-wenn-wir-unseren-ohren-nicht-mehr-100.html
[17] https://www.theguardian.com/technology/2023/jan/04/apple-artificial-intelligence-ai-audiobooks
[18] https://rp-online.de/nrw/staedte/duesseldorf/duesseldorf-tonies-testet-geschichten-mit-kuenstlicher-intelligenz_aid-90005417

("CDSM Directive"), adopted in spring 2019 and since 7th of June 2021 in force, Articles 3 and 4, mentions within Art. 4 a non-remunerated exception for TDM (Text and Data Mining) for general purposes.

The preliminary breaking points of this:

- **Unclear legal situation:** It is at least uncertain whether legal permissions for TDM (based in national legislations of EU Member States on Art 4, 2019/790 EU) allow the use of copyrighted works as "training data" for scraping, copying, storage and unsupervised machine learning. **Even more, it is considered in the European as well as International area of IP experts, that machine learning for generative informatics like synthetic and algorithm-based texts, images or voice (re-)generators, is a total new form of usage, needs the authorisation by the originator, and has in any case to be handled within a volunteer and remunerated licensing system and does not fall under no circumstances under the fair use doctrine, as well as not under the Art 4 exception of the CDSM Directive.**
- **Even more, it is seriously considered that Art 4 is not in line with the Paragraph 9.1 of the Berne Convention. Art. 4 of the CDSM Directive is also not and do not cover any TDM for machine training done before July 2021.**
- **Non-remuneration:** output of generative text bot systems and synthetic text generators compete with a market and harms the interest of authors and other rightsholders. The European Commission fatally has not run a 3-step-test incl. the risk assessments, and the damage expected to a market, and to moral, personal, and economic rights. The non-remuneration is a transfer of value towards tech, and for the illegitimate profit of users. A violation of the Berne Convention can be assuemd,because it includes forbidding any exception without compensation whenever there is harm caused to the legitimate interests of rightsholders.

- In any case, however, the opt-out provided for TDM (Text and Data Mining) in the CDSM Directive is in no way practicable. And this is not only due to the lack of contractual routines, in which authors could declare the opt-out when transferring rights of use – as there is no customary practice to declare if writers or translators agree to TDM or not. **None of the contracts** concluded until 2021, when the Directive came into force, included declarations on TDM; only now, in the year 2023, some publishing houses and writers start to include TDM opt-out declarations in their bilateral agreements and contracts.
- **No sector standard for meta data, ONIX or other machine readable opt-outs or rights reservation protocols:** There is no standardisation to make an **opt-out** machine-readable within works that are "available online"; also, according to contracts none of the AI development companies have asked so far, to be quite sure. It is also unclear what "available online" means and where to draw the line; it is highly likely that scrapers also obtain book works behind paywalls of online book retailers.
- And although several other, machine readable opt-outs are available they have demonstrably been actively ignored by AI data scrapers. The opt-out declarations so far used are: Terms Of Use, on-page copyright information within the imprint, CMI metadata embedded in image files, rights and attribution information in filenames and image descriptions, rights information and signature (watermark) visible in the image itself, copyright notices embedded in source texts and code.
  But these opt-outs are ignored.

- **No technical international standard for an declaration for opt-out in sight**:

Even though the W3C group is working on developing solutions to flag an opt-out in e-books (see July 2023 report[19]), currently only for URL and metadata of EPub3, authors remain unprotected until an indefinite time.

- **Some good news: the ISO standard ISCC (International Standard Content Code)** is being tested now for approval (previous standards in the book sector are ISBN, ISSN, ISNI, ISTC, DOI); opt-out declarations and a "fingerprint" of the file with this new identifier could be machine-read by special software - if AI developers were interested in rights clearance ... but in any case, the ISCC identifier can help solving the problem of making data sets transparent and also document training data.
- But: It remains completely unclear how an opt-out can be explained for **analogue works.**
- **It is also an open question whether an opt-out also applies to works that have already been used for TDM in the past.**
- Equally open is how to deal with out-of-print works and out of commerce works, when they are digitised by libraries or archives: who is implementing the opt-out? What is with scanned print books, for instance illegally distributed by the Internet Archive, and misused for machine learning?

International Factor: Certainly, the most detrimental legislation in the world for any author is the TDM exception in Japan, which does not even allow an opt-out and makes piracy possible without sanctions. Therefore, this should be avoided completely in the United States from where all the currently relevant AI tech companies are based, and stealing globally works, books, data, images and private information.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

In general, the full respect of international IP law, esp. Art 9.1 of the Berne Convention, is mandatory. No further limitations or exceptions shall come into force, and no Fair Use doctrine.

GAI systems like synthetic text (re-)producing products, shall only be trained of cultural works following the ART principle: Authorisation – Remuneration – Transparency. Only under a voluntary opt-in-regime with proportional remunerated licenses, and a fully documentation of usage and purpose, shall the works of authors and data of citizens be considered to be used.

- On the input: In principle, we are of the opinion that scraping, reproduction, storage and use for machine learning as a basis for large language or image models are (new) types of use that are reserved to the author according to Art 9.1 of the Berne Convention. This leads to the effect of considering volunteer opt-in licensing regimes.

Transparency is key here: corpora-builders, TDM (Text & Data Mining) institutions, developers, distributors, and providers of foundation models should be obliged to provide details of ALL copyrighted materials used, for what purpose and from where and when on this was collected. Illegal access to protected works and data shall be sanctioned.

- On the output: not fully or partly (re-)generated text, image, or audio product by a synthetic so-called generative AI system, shall be granted copyright or any related sui generis right.
- Labeling: Individuals and entities should be always and with no exception informed when and to what extent works are AI generated, to promote trust in human made labour, to avoid disinformation, and to ensure legal claims for remuneration.

[19] https://www.w3.org/community/tdmrep/ and https://www.w3.org/2022/tdmrep/

This includes to develop and deploy reliable content authentication and provenance mechanisms such as meta data, ISCC identifiers, watermarking, visible labels, and sector specific standard techniques to enable users to identify AI-generated content, and to manage at the same time necessary rights reservation protocols.

- Also, a system of liability is needed, that takes effect in cases of disinformation, violations of personal rights, and distortion of competition.

Training

If your comment applies only to a specific subset of AI technologies, please make that clear.

We comment on text generators (book sector, partly press and websites).

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

Between 2018 and mid-2021 seven large language models were released to public, including the predecessors of today's mostly known generative (Natural Language Processors) NLP specific model, ChatGPT. Its basic models (today GPT-4, previously corresponding to 1-3) have been developed since 2015, in parallel to Google's BERT. The concept of deep machine learning, which is based on large corpora of books, and takes place without human supervision or preparation or curation, was developed in 2017.
The corresponding "collections" are called BookCorpus, Books3[20], C4[21], or The Pile.
This means, that the unlawful scraping and (mis)using is taking place at least since a decade.
Plus: data sets are collected, stored, and used since the 1960ties; with two relevant boosts in 1999 and 2008, as the timeline[22] of the developments of data sets for machine learning shows.

Relevant Sources:

- All 12 large Foundation Model Providers refuse to disclose copyrighted data (Standford[23]).
- "Data archaeology analysis"[24] revealed insights into Open AI's GPT-3.5 used in-copyright published book works before 2021 from all over the world and in at least three dozen languages.
- BookCorpus, one of BERT's (Google's LLM) training sources in 2015, contains in-copyright materials by published authors; but also, by unpublished ones via the website Smashwords.
- The sources are assumed to be bit torrent piracy sites[25]. Shadow library websites like Library Genesis (LibGen), Z-Library (Bok), Sci-Hub, and Bibliotik make large collections of copyrighted books available for bulk download through torrent systems. Thus, chain of illegal acts could be traced, with AI companies being conscious of the at least unethical behaviour.
- The Researchers of AI Safety Camp decoded over 200.000 titles (full list), among them also books by European Authors or Publishing Houses, and used for training within Book3 or The Pile, before any TDM exception came into force (2023/07).

[20] https://github.com/psmedia/Books3Info , https://public.tableau.com/app/profile/peter.schoppert/viz/ISBNSfromtheBooks3database/ByImprintbubble?ref=psmedia.asia#1
[21] https://arxiv.org/pdf/2104.08758.pdf
[22] https://devopedia.org/text-corpus-for-nlp#:~:text=A%20plain%20text%20corpus%20is,it%20may%20produce%20better%20results.

[23] https://crfm.stanford.edu/2023/06/15/eu-ai-act.html
[24] https://arxiv.org/pdf/2305.00118.pdf
[25] https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?itid=lk_inline_manual_54

Conclusion: AI companies have been pulling copyrighted book works from bit torrent piracy sites since 2013[26],[27]. The corpus Book3 and The Pile was proven to contain 194,000 identified titles; under investigation by volunteer research teams of e.g., the Dutch based AI Safety Camp, are 1.2 million more copyrighted titles. At the end of September 2023, this led to a lawsuit by 17 US authors[28] such as George R. R. Martin and Jodi Picoult, among others.

Unlawful scraping: In addition, there is ample evidence that machine-readable robots.txt opt-out statements on html websites and URL are simply ignored by scrapers or unsupervised machine learning crawlers. This includes press articles not behind a "pay wall", Fan Fiction Websites, but also Twitter posts, forums, or Facebook comments.

The works used for "input" are not curated; the unsupervised deep learning takes place on the full material. Only after a large language model (LLM) is build, additional tokens and labels are set by labelers, to develop an LLM for further specific purposes. ChatGPT is built from GPT LLM.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

We don't know, as AI developers do not make their sources fully transparent. There are voices who claim that the scanned Google's Books works are part of the corpora. As well and as mentioned above, sources were also piracy websites, and websites which contains protected text works. It cannot be precluded that companies also collect data from territories where there is no regulation or operate in territories which, like Japan, do not have any copyright or opt-out for this case.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

If "training material" means: for unsupervised neuronal deep machine learning as foundation for commercial large language models and generative (text) informatics: None were licensed (or remunerated).

6.3. To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

Until 2018, the Google translator's learning system was taught from sample texts such as the Bible, instruction manuals, Wikipedia, or texts from the UN or EU Commission on their websites.

Further sources of LLM: e.g., Wikipedia, websites of the European Commission, Project Gutenberg, Reddit, or Goodreads, press articles without a pay wall, Wattpad, any website with text by a blogger – although publicly available and not always in the public domain, copyrighted text works were also used without licensing, without consent, without information and, of course, without remuneration and consequently illegally.

In September 2023 certain AI developers started to hire writers and poets to (a) teach the synthetic regenerating text software how to write "better", and (b) to write texts for learning purposes. So, writers train generative systems with their very own replacement in sight.

[26] https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?itid=lk_inline_manual_54
[27] https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/
[28] https://apnews.com/article/openai-lawsuit-authors-grisham-george-rr-martin-37f9073ab67ab25b7e6b2975b2a63bfe

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

Due to the technical working of these models – from scraping, copying, storing, unsupervised self-learning, broken up into tokens –, neither personal information nor copyright protected works can be removed. This leaves authors and other rightsholder in a trap. We consider that this non-ability to forget is inherently non-compliant with the DMCA Safe Harbor provisions or any GDPR Right To Be Forgotten and Schrems-II.
AI tech companies claim that the inputs disappear once "learned" from the machines: even though the first copy of the work is substituted by another within the data sets and broken up into tokens, n-grams or other transformed pieces, the fact that the AI system can recompile and re-produce in full copyright protected works it from its learning is like having another copy of the protected work in storage and use.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in:
7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

As highlighted in the statement by the German "Initiative Urheberrecht": "Künstliche Intelligenz braucht Leitplanken" ("Artificial Intelligence Needs Guard Rails | Initiative Urheberrecht") of 28th April 2023[29], **there are three levels: INPUT – PROCESSING – OUTPUT.**
At the INPUT level, a distinction must be made between the selection and acquisition of data (SCRAPING) and TRAINING; this internal differentiation is new to the system of classification. In computer science, PROCESSING (computation) is consistently described as a black box; not even the operators of AI systems know exactly what happens during the learning process – and they do not control it. The products of generative AI are compiled at the OUTPUT level.

In the exchange between AI-researchers from various German universities and the on informatics specialised Fraunhofer Institute, as well as IT-lawyers from guilds, trade unions and collective management organisations, the Initiative Urheberrecht is working on an ongoing evaluation of the copyright status quo of current generative AI-Systems.
The findings about the training-process of generative text robotics[30]: INPUT consists of two steps: SCRAPING and TRAINING.
Initially, all kinds of data, including significant amounts of copyrighted works are collected and stored so that they can be used to train the AI system in the next step. This process is called SCRAPING, and it is undoubtedly a copyright-relevant process. Specifically, the works and performances collected are (a) copied and (b) stored in a database so that they can be made available for training.
According to leading computer scientists specialised on AI, however, the relevant databases, or rather their contents, are not usually deleted. There are several reasons for this, not least of which is the possibility of subsequent re-processing to determine comparability.
In terms of copyright, the processes at the INPUT level can be described as follows:
SCRAPING involves mass copying, a copyright-relevant process. All parties involved (authors, rights holders, users, AI platform operators etc.) agree that this is the case from a technical and

[29] https://urheber.info/diskurs/artificial-intelligence-needs-guard-rails
[30] https://urheber.info/media/pages/diskurs/current-iu-position-paper-on-generative-ai/1750ecf67e-1697140215/230919_iu_position_ai-act_september2023_en.pdf

copyright perspective. All works and data are copied. Scraped data of all kinds is stored in a database as a basis for TRAINING.

During TRAINING, the second step at the INPUT level, models are taught from the previously copied and stored works and content that predict probabilities (such as certain character, pixel, or word sequences). Current models are generally based on machine learning (including neural networks/deep learning), and are neither supervised by a human, nor curated or prepared.
From a computer science perspective, TRAINING itself does NOT result in a usable database "per se"; the computed model cannot and is not intended to function like a traditional database. The PARAMETRIZATION of the trained data results in a highly abstract REPRESENTATION or MANIFESTATION of the content within the model. From a copyright perspective, there is no database since the data is not "individually accessible by electronic or other means".

For more technical insights into the different methods of neuronal machine learning, like preprocessing, tokenization, stemming, de- or encoding, transforming etcetera, is to find in the comprehensive paper by the Initiative Urheberrecht[31] and further sources in the footnote[32].

7.2. How are inferences gained from the training process stored or represented within an AI model?
We quote again from the analysis paper[33] by the Initiative Urheberrecht: After training, the data in the model is not available as copies in the "traditional" sense. The GAI model no longer uses the database created previously for its output, but rather only uses the parameters selected from that database. Based on the current state of the art, however, it is not possible to provide an unambiguous description of how exactly the partly trillions of parameters[34] in the model are classified, even from a technical point of view. In the copyright debate, the question of whether reproductions in the sense of copyright law still exist after the training has been completed is the subject of some debate. However, there is much to suggest that even the trained GAI model (at the 2nd level) still contains reproductions in a copyright-relevant sense, since it is undoubtedly possible for systems like ChatGPT to reproduce poems "in the style of", or other copyrighted texts. Even if the reproduction of the respective text is based on the probability of stringing together the respective passages based on the respective user requests ("prompts"), the work is still part of the model in this way.

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material?
No, see response to Q 6.4. Unlearning or forgetting what has been put in the machine, is not possible according to the current state of technology and statements by leading AI scientists. There is therefore a risk of substantial claims for damages. If one of the pending lawsuits against generative AI providers in the U.S. is successful, their entire MODEL would have to be deleted and the training process would have to be restarted. Also, even though the first copy is "broken up", the machine is still able to "memorise verbatim" the protected work or parts of it; meaning that the "unlearning" process is non-existent.

If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?
A clean slate of all existing GAI applications, as they are based on theft.

[31] https://urheber.info/media/pages/diskurs/current-iu-position-paper-on-generative-ai/7a9938d31c-1697140215/230919_iu_position_ai-act_september2023_en.pdf
[32] https://medium.com/@memrhimanshu/generative-ai-for-text-classification-1ceee4a0da79
[33] https://urheber.info/media/pages/diskurs/current-iu-position-paper-on-generative-ai/7a9938d31c-1697140215/230919_iu_position_ai-act_september2023_en.pdf
[34] https://www.linkedin.com/pulse/large-language-models-power-billions-parameters-isabel-hong?utm_source=rss&utm_campaign=articles_sitemaps&utm_medium=google_news

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

Yes. Among others, the Stanford University proved the memorizing ability of complete sentences and chapters of copyright protected books and press articles, and also researchers from the Department of Computer Science of the **University of Copenhagen and the University of Electronic Science and Technology of China**[35] **proved verbatim reproduction**. Quote from the paper, released on 20th of October 2023: "*This work explores the issue of copyright violations and large language models through the lens of verbatim memorization, focusing on possible redistribution of copyrighted text. We present experiments with a range of language models over a collection of popular books and coding problems, providing a conservative characterization of the extent to which language models can redistribute these materials. Overall, this research highlights the need for further examination and the potential impact on future developments in natural language processing to ensure adherence to copyright regulations*."

On the other hand, there are corpora built since the 1990-ties, which are re-used for today's purpose. They include Google Books, Amazon reviews, IMDB summaries, copyrighted book works, press and magazine articles, etcetera and the whole content of the web, both protected works, but also data, personal and private information, etc. Most of them are highly likely not fully legally obtained, and without consent or knowledge by authors, citizens and rightsholders. An overview of offered corpora and data bases for NLP machine learning for large language models is to find in the footnote[36].

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

In no means, as the scraping, copying, storing, and re-using for machine learning are (new) types of usage which fall under Art 9.1. of the Berne Convention.

8.1. In light of the Supreme Court's recent decisions in Google v. Oracle America [41] and Andy Warhol Foundation v. Goldsmith,[42] how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning,[43] raise different considerations under the first fair use factor?

8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training?

8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes.[44] How should the fair use analysis apply if AI models, or datasets are later adapted for use of a commercial nature? [45] Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

8.1-8.4: **We refer to the position of the Authors' Guild of the United States.**

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? [46] Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

[35] https://arxiv.org/pdf/2310.13771.pdf
[36] https://devopedia.org/text-corpus-for-nlp#:~:text=A%20plain%20text%20corpus%20is,it%20may%20produce%20better%20results

The core goal of generative AI is to substitute existing works crafted by human authors. The purpose of a textgenerator model is to generate output that competes directly against the used works and all future works of human writers. The character of the use is producing a source for production of second-order derivatives of marketplace substitution grade - we are consequently dealing with a 100% demand for works which, unremunerated, are used to occupy the same market through synthetic output also at 100%.
Accordingly, the 3-step-test must include the risk assesment of competition and harming the legitimate interests of authors and rightholders.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

As the scraping, copying, storing and usage for machine learning fall under Art 9.1. of the Berne Convention, the voluntary opt-in-regime with remunerated, time limited licenses and under a full transparency regulation for purpose, outcome, and profit, we recommend to never take opt-out as a similar solution into consideration. The author has to be asked.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses? [47]
Definitely yes, this is also necessary under the Berne Convention.

9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses? [48]

First it must be considered that besides owning the copyright, every author has the moral right as well as the protection of the Berne Convention Art 9.1.
In any case, some feasible ways for functional opt-out systems, although we do not recommend them as a good solution:
(a) Contractual declaration routines for text, image, voice and translation to not transfer any right of TDM, or copying, storing and use for machine learning (b) meta data rights reservation protocol within the file, e.g., Onix, ISCC identifier, TDM and machine learning rights reservation protocol (c) opt-out "flag" on the website of the publisher and book retailer online, incl. opt-out for TDM, but also opt-out for scraping, copying, storing and machine learning (d) full transparency disclosure by developers, deployers, distributors of GAI, as well as from corpora developers and Text and Data Mining entities, to track and trace the copyrighted material.
Any opt-out regime shall NOT be free of remuneration or compensation.

The technical tools are in the making, please see response under question 4: Meta Data and Onix solutions developed by W3C, and the International Standard Content Code (ISCC) identifier.
In any case, opt-outs should NOT only be accepted when machine readable, but also declared and valid within imprints, general terms and conditions or similar declarations within the work itself as well as in digital means. They shall consider the new forms of usage, as TDM does not cover the process for machine learning.

9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?

Informed consent is a must. Practicality shall never been put over a fundamental right of consent. Even AI could be used to develop the smooth-running logistics.
What is needed is the respect of AI developers, providers, corpora builders, deployers and distributors – which is currently missing. Example from the visual artistry:

There is an opt-out registry covering one billion works already[37], provided by Spawning.ai. OpenAI ignored these declarations while building the synthetic diffusion image reproducer, DallE3. This is an insult to authors and rightholders, to ignore their specific rights reservation.

9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?
(a) Sanctions against the denying company (b) financial charge to be paid to the author and, where applicable, further rightsholders, for the damage done in the past.
As there are no existing remedies at all for the theft of works to develop commercial high profit systems which are competing in the same market they stole works from, only a clean slate and a shutdown of all existing GAI systems and business models are the correct legal and ethical way.

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

Sure, within a contractual routine. There shall be no fear or disadvantage for those who object. We please also ask to name authors, artists or originators as such; as there are not "creators' rights" or similar, and the wording creator has no legal framework.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained?
On a strictly volunteer basis, made possible for instance via collective management organisations, and with corresponding contractual routines within author and the first counterpart, as well as with the CMO.

10.1. Is direct voluntary licensing feasible in some or all creative sectors?
In all sectors where an author and artist as originator exists.

10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? [49] Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses?

Yes, voluntary collective licensing is a feasible approach, via CMOs, RROs and further similar entities, who are managing primary rights as well as secondary rights.
What they need are bilateral agreements to represent an international repertoire.
If any antitrust or competition laws object to the operating of CMOs or RROs, the adjustment of specific US laws is recommended.

10.3. Should Congress consider establishing a compulsory licensing regime? [50]
No. This is against the Berne Convention 9.1, and it is not covered by a 3-step-test as well as there are no urgent needs to bend authors' rights as well as copyright to an extent from which only a few techs monopolies benefit.
If so, what should such a regime look like?
What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported, and distributed?

10.4. Is an extended collective licensing scheme [51] a feasible or desirable approach?
Yes.
10.5. Should licensing regimes vary based on the type of work at issue?

[37] https://spawning.substack.com/p/our-seed-round-of-funding-to-build

Yes.
Draft steps to develop collective and volunteer licenses for the different types of use, e.g., but not limited to, Text and Data Mining, but especially for Scraping, Copying, Storage, pre-processing for Machine Learning, and, as in our case, usage in Large Language Models, and other purposes in the field of advanced generative informatics, need to be developed properly.
New forms of use arise, and it is always on the author to decide to transfer or not by virtue of his or her intellectual property rights, and types of usage.
This is where collective management comes in to set up and represent machine learning licenses - for specific, or general use.
In principle, the ART principle should be respected:
Authorisation
Remuneration
Transparency
This means that it is only up to the author as such to voluntarily consider opting in. He should not suffer any disadvantage if he refuses to agree to a use that in the long run will provide competing products that will partially or completely replace his labour.

Collective licenses could be developed along the following lines:
(1) Volunteer opt in for: TDM; Volunteer Opt-in for: Machine learning for specific or general purposes; Opt-in for the steps between scraping, copying, storing and the output of LLM.
(2) Differentiated licenses, e.g., limited in time, used only for certain fields, and subject to conditions such as transparency, tracking and documentation of works used; furthermore, they should ensure proportional and appropriate remuneration. A "blank" license should be avoided.
(3) Calculating an appropriate remuneration is based on numerous factors: what is the estimated loss of income through the use of generative text robotics, in particular through the damaging business models that have already emerged, such as fake books and illegal translations; what is the competitive disadvantage for the author when texts are produced automatically "in the style of" (him/herself); what is the role of other copying perceptions; to name but a few.
(4) To be defined: who are the debtors (corpus producers, AI developers, AI providers or end users?).

11. What legal, technical, or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model, and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

The complete AI value chain needs a routine of a clear, transparent, and functioning rights reservation protocol; this is like every chain with different stakeholders. The developer of foundation models and corpora builders, if not the same entity, have the duty to obtain lawfully accessed licenses and document every source, and make bilateral agreements with the follow-up entity. May it be with the ISCC fingerprint or within meta data.
Every entity in the value chain shall have the same mandatory duty to guarantee, as well to reach out for the guarantee, that material is legally licensed. Aspects like commercial or non-commercial do not play a role, as the chain of rights transfer remains the same.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

As every work makes the existence of any foundation model possible, it contributes 100 per cent to its function. It needs a distinction to the output, as this is dependent on the prompts.
In principle, you have to look on the input, which makes it possible to replace with its synthetic, re-generating mash up remix the achievement of human work.

But: Proof of whether certain specific works have been used for machine learning or for the creation of the foundation models cannot be provided at the INPUT level alone, but must often be provided also at the OUTPUT level, in addition to the transparency which shall be required about the type and amount of training data used: For example, if a prompt asks for the style of a particular writer or poet, and the output is very close to that style ("proximity"), then it can be concluded that the works of that poet were used for training.
If the work is no longer present as a copy in the model, the corpus of training results, but is represented abstractly, then this may constitute a NEW TYPE OF USE. If it can be proven at all that the specific work exists and can be found, the question of whether it manifests itself in abstract vectors or in bits and bytes is irrelevant. We are dealing with technology that allows REPRODUCTION.
In many cases, no pixel in the OUTPUT product, such as images, is identical to the original, which raises the question as to whether proximity should be determined technically or based on perception.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

A healthy and ethical and legally correct one. The goal must be to remunerate every participating part, from originators and authors to labelers, fair and appropriate. If a company has no capacity to finance this, it should look for opportunities to calculate better, and to not take the short cut to theft.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.
Unlicensed copying and other copyright infringements occur constantly at the INPUT level.
As noted above, a significant amount of the scraping that has occurred to date took place without authorisation, remuneration or transparency – and over two decades. At least – as data sets for machine learning have been developed since the 1960s but had a boost within the digital evolution and material available on the world wide web.
These are blatant copyright violations that the U.S. shall not accept, if only for economic reasons; after all, the content digested has been used to train systems that are preparing to replace the commercial production of new works and labour by human authors.
It is essential to find solutions for past illegal use of this works and data that are acceptable to the authors and right holders.

Transparency & Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?
Yes, both entities.

15.1. What level of specificity should be required?, 15.2. To whom should disclosures be made? 15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?
Introduce a comprehensive transparency obligation that, in addition to its direct copyright relevance, also allows for market monitoring and clear impact assessment. Authors, performing artists and rights holders must be able to find out whether and to what extent their works and

performances are being used for training at the INPUT level and the extent to which they are being used as a basis at the OUTPUT level.
Introduction of a fundamental duty to label products originating from generative AI, facilitating the unambiguous, comprehensible identification of machine-generated content. However, it is possible that a total and comprehensive duty to label may not apply in certain rare cases due to constitutional requirements.
We again recommend the W3C approach of meta data to be inserted into data set documentation, as well as the ISCC identifier, as a sort of "fingerprint" of each work, to be tracked and traced easily and containing all relevant information.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

The impact would be to finally comply with standards of rights, dignity, and ethics, and to no longer benefit from theft.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?
In principle, authors and rightsholders shall be informed before, and not after.
But if after, especially on the usage from 2008 on, there needs to be an official international and national trusted entity and registry, with the option to have the work deleted from foundation models. If this is not manageable: the system in full has to be deleted and taken off the market.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?
We refer to the position of the Authors' Guild of the United States.

Generative AI Outputs
If your comment applies only to a particular subset of generative AI technologies, please make that clear.
We comment on text generators (book sector, partly press and website).

Copyrightability
18. Under copyright law, are there circumstances when a human using a generative AI system should be considered the "author" of material produced by the system?

No. Copyright protection requires "individual intellectual creation" associated with a natural person. As this is not the case with autonomously generated AI products, these types of products cannot be given the status of a work and therefore cannot be given copyright protection. Nor can the person formulating the prompts claim any rights with respect to the result on the basis of the prompts alone, because the mere formulation of the task and the choice between several results proposed by the AI system is not a creative or protectable act.

If so, what factors are relevant to that determination? For example, is selecting what material an AI model is trained on and/or providing an iterative series of text commands or prompts sufficient to claim authorship of the resulting output?
No, it is not. As ideas are not protected, neither prompt nor text commands are sufficient.

19. Are any revisions to the Copyright Act necessary to clarify the human authorship requirement or to provide additional standards to determine when content including AI-generated material is subject to copyright protection?
We refer to the position of the Authors' Guild of the United States.

20. Is legal protection for AI-generated material desirable as a policy matter? Is legal protection for AI-generated material necessary to encourage development of generative AI technologies and systems? Does existing copyright protection for computer code that operates a generative AI system provide sufficient incentives?

No, a legal protection for machine generated output is not desired, as it is not man-made. There shall be no exception.

We additionally refer here to the Universal Declaration of Human Rights[38] and Article 27(2): " Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."

If this fundamental right, this achievement of the Enlightenment, and the manifestation of the protection of the human individual and its very own form of expression, were to be counteracted by a similar or equivalent "machine right", this would be tantamount to denying human creativity and will.

20.1. If you believe protection is desirable, should it be a form of copyright or a separate sui generis right? If the latter, in what respects should protection for AI-generated material differ from copyright?

It shall have no own right at all. Copyrights and authors' rights are attached to human labour, individuality, responsibility, originality, will, and freedom of speech as well as having access to make art under the Universal Declaration of Human Rights. To bend these human rights for a machine is an approach which will lead to the demolition of art and culture.

21. Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material? Would such protection "promote the progress of science and useful arts"? [52] If so, how?

We object against the thesis that any generative AI is a friend of the arts. No, any uplifting of machine produced output, with no self-control, no liability, no intrinsic motivation, no origin, no labour, is a blind enthusiasm, which oversees the legal, the economic, the ethical harm. If machines do human mental work, in the long run the next generations will lose the techniques of writing, composing, researching, communicating and making art.

The IP system is in its heart also a safety net for investment; especially of the private risk-investment and non-paid work of authors and artists, which depend, that EVERY USE OF THEIR WORK is remunerated.

It should be borne in mind that intellectual property rights for further rightsholders, also such as those of the sound carrier or film producer or even publishing houses, are equally based on the idea of investment protection. Providers of AI services use the AI infrastructures of companies such as Microsoft, for example, meaning that the investments that may be worthy of protection are not made by the individual providers.

Infringement

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

See Section 7 and relevant responses to 7.1-7.4.

Based on the referred answer in section 7, exclusive rights of preexisting copyrighted works are implicated both at the input level and at the output level.

At the input level, exclusive rights are implicated because all the inputs nourishing the AI systems are copies of (legally or even illegally obtained) protected works. This copy is a "use"

[38] https://www.un.org/en/about-us/universal-declaration-of-human-rights

of the protected work to the benefit of the AI system. Without the copy of the protected work, the AI system is empty. This use, in legal terms, is called "a reproduction" of the protected work that requires the authorisation of the author or its grantee/licensee under the Berne Convention, which clearly and unequivoqually states in its article 9:
"Authors of literary and artistic works protected by this Convention shall have the exclusive right of authorizing the reproduction of these works, in any manner or form".

At the output level, exclusive rights are implicated because the outputs are generated by "putting together" pieces of preexisting copyrighted works. Without "reusing" the preexisting protected work or pieces of the preexisting protected work, no generated product from AI would be possible:

- Whenever these pieces of protected works are sufficiently enough in volume to able the public to recognise a protected work, there is no discussion as to whether the exclusive rights are implicated: they are implicated as the protected work is used to compile the generated AI product and the public can designate the origin of the generated product, the originated author or the originated protected work. Counterfeiting acts are analyzed by courts in Europe by comparing similarities and proximity (and not by making differences stand out): the more similarities there are, the more the counterfeiting acts are acknowledged by courts.
- Whenever these pieces of protected works are not sufficiently enough in volume to be recognised a protected work, exclusive rights are still implicated since the AI system is rebuilding a work / product, from the use of millions of pieces of the protected work, which is supposed to be "one and only" and should not be "divided" or "scattered in million pieces" without the authorisation of its author. Consequently, this is when transparency is unavoidable as it is necessary to know which pieces from which author were used in order to generate the AI product. In the end, there might not be any counterfeiting act, but other exclusive rights are implicated in countries following authors' rights legislations and integrating moral rights: imitating the style of an author could be seen as the use of trends in the art of the author, modifying the protected work / distorting the protected work in order to avoid obvious counterfeiting but still suggesting the pre-existing works to content the public.

As explained by AI tech companies, the work of the AI system is to predict how to construct a work from the neuronal network. Without the pieces of the protected works, the AI system cannot build its prediction.

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?
See responses to questions 7.4 and 12.
Yes, the substantial similarity test is adequate when the pre-existing protected work is recognisable, meaning that the public can relate the generated product to a pre-existing protected work (even though the generated product by the AI system is not to considered as a protected work under copyright legislations). The similarity test allows a comparison that will able the author from the pre-existing protected work to argue that, even though the generated AI product is not a complete copy of the pre-existing work, the exclusive rights are still implicated, and the author's authorisation was required.
When the protected work is not recognisable in the generated output product, the similarity test could be completed by other standard tests and, in countries acknowledging moral rights, by scrutinizing whether the protected pre-existing work has been either decompiled or recompiled differently that would harm the authors' integrity right.

Plus, every time a prompt is made to order a piece "in the style of," and the outcome is within a clear proximity, the rights of the author are damaged.

In addition, in principle, every work used for training must be identified via an international standard, like the ISCC identifier or ONIX data; the output regulation needs to be addressed with a strict labelling regime of fully and partly generated texts incl. translations, and with a labelling and recording by the provider of prompts based on order "write in the style of...", to guarantee that the original human writer or his legal successors are remunerated each time a text is generated on the basis of their work and compete in the same market.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?
See responses to 7.4 and 12.
Unfortunately, authors only have the way to claim and go to court, which puts the burden again on the victims of this mass theft.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

Everyone in the AI value chain, starting from scrapers and data basis developers, to developers, deployers and providers, up to everyone who uses this, incl. the end-user. If the end-consumer does not want to be held liable for infringement, he should not use it – and providers of any GAI model shall warn their end-consumer of the risk.

25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs? [53]
26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?
27. Please describe any other issues that you believe policymakers should consider with respect to potential copyright liability based on AI-generated output.
We refer to the position of the Authors' Guild of the United States.

Labeling or Identification

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

Yes. Individuals and entities should be always and with no exception informed when and to what extent works are AI generated, to promote trust in human made labour, to avoid disinformation, and to ensure legal claims for remuneration.
Also, labelling is the basis for a need and to be legally implemented system of liability that takes effect in cases of disinformation, violations of personal rights, and distortion of competition. Proof and liability must be clarified.
Lawmakers and sectors together have the duty to develop and deploy reliable content authentication and provenance mechanisms such as meta data, ISCC identifiers, watermarking, visible labels, and sector specific standard techniques to enable users to identify AI-generated content, and to manage at the same time necessary rights reservation protocols.

Plus: if AI products and outputs are not labelled as such, and re-used via scraping and training, the risk of the total AI system collapse is high: the LLM is poisoned by itself.

28.1. Who should be responsible for identifying a work as AI-generated?

Developers, providers, and end-users (individual as well as using or distributing entity, e.g., press, publishers, online portals, and platforms, etc.), distributors, deployers, incl. the sectors itself when GAI is used, e.g., for AI manipulated movie effects, machine translations and even summarizer of text works. The granulation can be considered.

28.2. Are there technical or practical barriers to labeling or identification requirements?

No.

28.3. If a notification or labeling requirement is adopted, what should be the consequences of the failure to label a particular work or the removal of a label?

For the sake of differentiation, and in view of the danger of manipulation and misinformation, GAI products should be labelled clearly and comprehensively – automatically from the moment of their creation, if necessary. For example, the ISCC standard[39] developed with EU funding could be helpful, especially since it is decentralized and non-proprietary. The deletion of such a label and the separation of any metadata records potentially associated with the file or its contents from the file or its contents should be prohibited – similarly to the prohibition of circumvention of copy protection measures. To not apply or to even remove a label shall lead to financial sanctions, including taking the product (software and output) off the market.

29. What tools exist or are in development to identify AI-generated material, including by standard-setting bodies? How accurate are these tools? What are their limitations?

Additional Questions About Issues Related to Copyright
30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

Personal rights (face, voice, appearance), performing rights (voice), competition and label rights (name), data protection rights, neighboring rights, digital rights.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI-generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

We refer to the position of the Authors' Guild of the United States.

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works "in the style of" a specific artist)? Who should be eligible for such protection? What form should it take?

Imitation and plagiarism are sanctionable. Making something "in the style of" is based on the fact that these works are already included in the basic model and is in itself already a multiple violation of personal rights, copyrights, moral rights, and fields of competition. Anyone who

39
Information about the ISCC can be found at: https://iscc.codes /// ISCC currently has the status of "Draft International Standard" ISO/DIS 24138; ISO project page: https://www.iso.org/standard/77899.html

orders something "in the style of" already runs high risks of plagiarism and copyright infringement. Producing imitations of existing copy-protected works must be prohibited. For further input see responses in section 7.1-7.4. on “proximity.”

33. With respect to sound recordings, how does section 114(b) of the Copyright Act relate to state law, such as state right of publicity laws? [54] Does this issue require legislative attention in the context of generative AI?

--

34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

When voices are separated from people in voice cloning, when actors and actresses are replaced by their own clones, when prompts as “write in the style of...” competes with the works and reputation of the real human author, when inputting a visual artist's name into a text-to-image model returns countless products that give the impression that they were created by that very artist, but also when the protagonists of journalistic and documentary media have statements put into their mouths that they never made and would never make, this is a profound encroachment on the personal rights of those affected.

The past actors' and screenwriters' strike in the U.S. is proof of the urgency of this aspect. There is a need for clear and enforceable rules, including a right of prohibition, to protect the personality, to which – especially in copyright law – the livelihood of most stakeholders is closely tied.

Any attempt to use contractual agreements to allow the unrestricted use of input for the production and operation of systems that compete directly with the authors of the training content should be prohibited entirely.

Politicians must consider that the value creation of the entire national creative industries takes place and is accounted for locally, while the profits generated by AI providers – together with the cultural heritage, world knowledge, innovative power, and identity-forming personal intellectual creations of all European knowledge workers in their entirety – are only for the benefit of a few. Which leads to the fact, that the local and national strength of the sectors are declining, which will lead in the long run to unemployment, raise of retirement poorness, and to a burden of social costs in counties and municipalities.

As the backbone, blood, and brains of the entire producing book sector worldwide, and its further exploiters including and especially AI developers and companies who are ignoring rights, risks, and responsibilities, we ask you to please incorporate our comments in your further reflections and actions. There will not be another opportunity to stand up for writers and therefore human cultural value any time soon, and in this way to sustainably secure the future of the book sector.

We remain available for any requests or concerns and look forward to a fruitful exchange.
With our kindest regards:

Nicole Pfister Fetz
Secretary General

Nina George
Commissioner

Maïa Bensimon
Vice President