"Do Chatbots Dream of Copyright?" - Copyright in AI-generated Language Data

CLARIN ERIC

11 April 2023

Organisers

This CLARIN Café is organised by **Paweł Kamocki** (CLIC) in collaboration with CLARIN

Your CLARIN host: **Antal van den Bosch**

Technical support by
David Bordon

The event is recorded for further dissemination purposes.
Questions and comments? Put them in the chat box.

Schedule

14:00 - 14:05 Opening and CLARIN 101 - Antal van den Bosch, Member of the CLARIN ERIC

14.05 - 14.35 *Of course. But… maybe?* - Paweł Kamocki, IDS Mannheim

14.35 - 14.55 *The model doesn't fall far from the data* - Thomas Margoni, KU Leuven

14.55 - 15.05 *The UK did it before it was cool* - Toby Bond, Bird&Bird

15.05 - 16.00 Questions and discussion

Moderator: Antal van den Bosch

Discussants:
Fabian Ferrari (Utrecht University), Francesca Frontini (CLARIN BoD)

Introducing CLARIN

https://www.youtube.com/watch?v=ut9wOIYWDfc

https://www.clarin.eu/content/clarin-in-a-nutshell

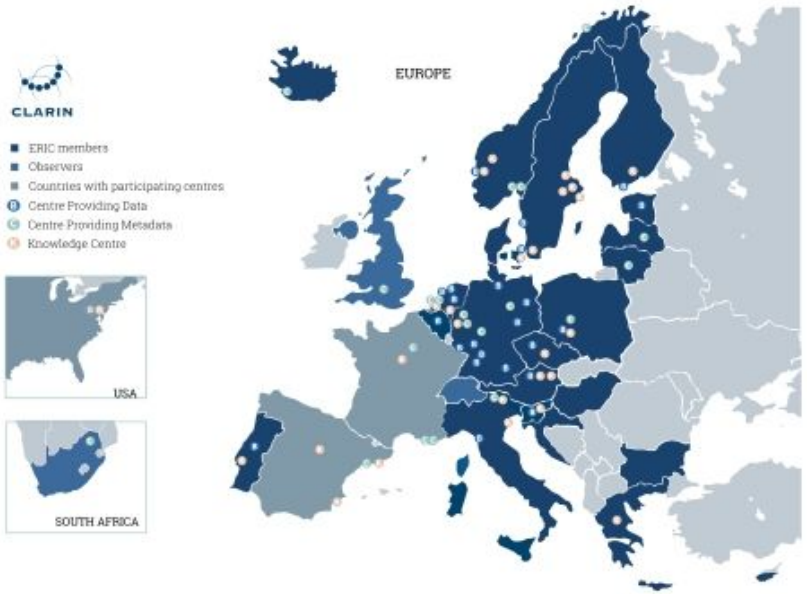
CLARIN

CLARIN …

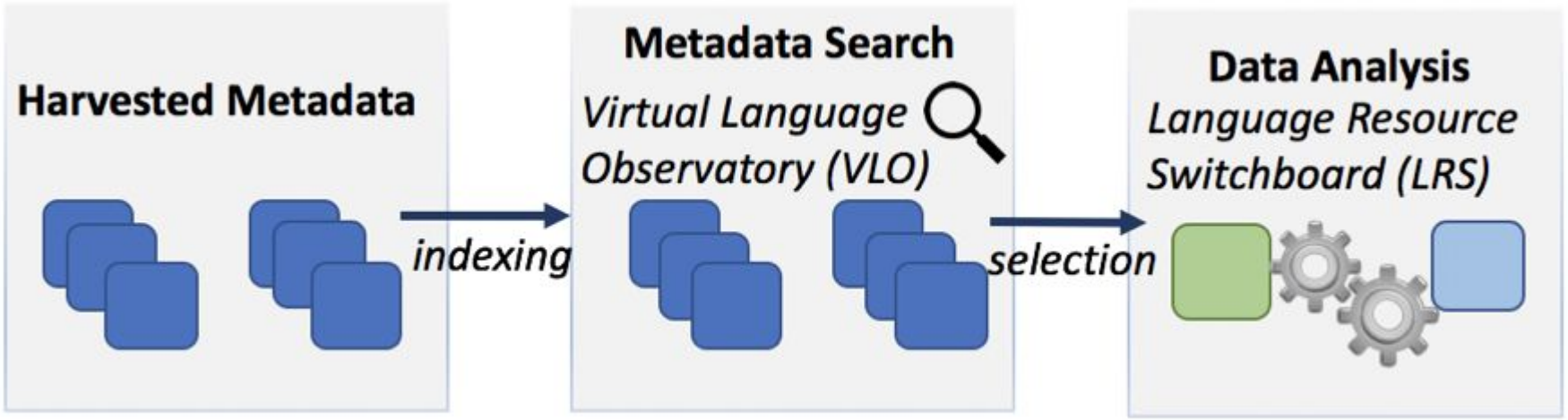
- is the *Common Language Resources and Technology Infrastructure*
- has the **ESFRI** ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing and training**
- is one of the European RIs in the SSH cluster (aka SCI)
- is an integral part of **the European Open Science Cloud**
  - See **clarin.eu/eosc**

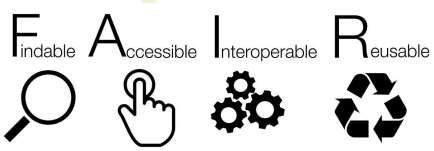
CLARIN today

- a distributed network of **70 centres**
- **22 members**: AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- **3 observers**: CH, UK, ZA
- **1 third party**



The Technical Infrastructure





clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

The Knowledge Infrastructure

https://www.clarin.eu/content/knowledge-infrastructure

Setting the scene

- 28 October 2021 - CLARIN Café on Text and Data Mining Exceptions in the Directive on Copyright in the Digital Single Market

- 8 November 2022 - CLARIN Café on Text and Data Mining Exceptions a Year After - Has the Pony Become a Horse?

The Café

Of course. But… maybe? Theoretical approach to copyright in AI-generated works.

Paweł Kamocki, IDS Mannheim

(public domain)

Of course, AI outputs are not protected by copyright.

- no human authorship
- no originality
- that would be unfair!

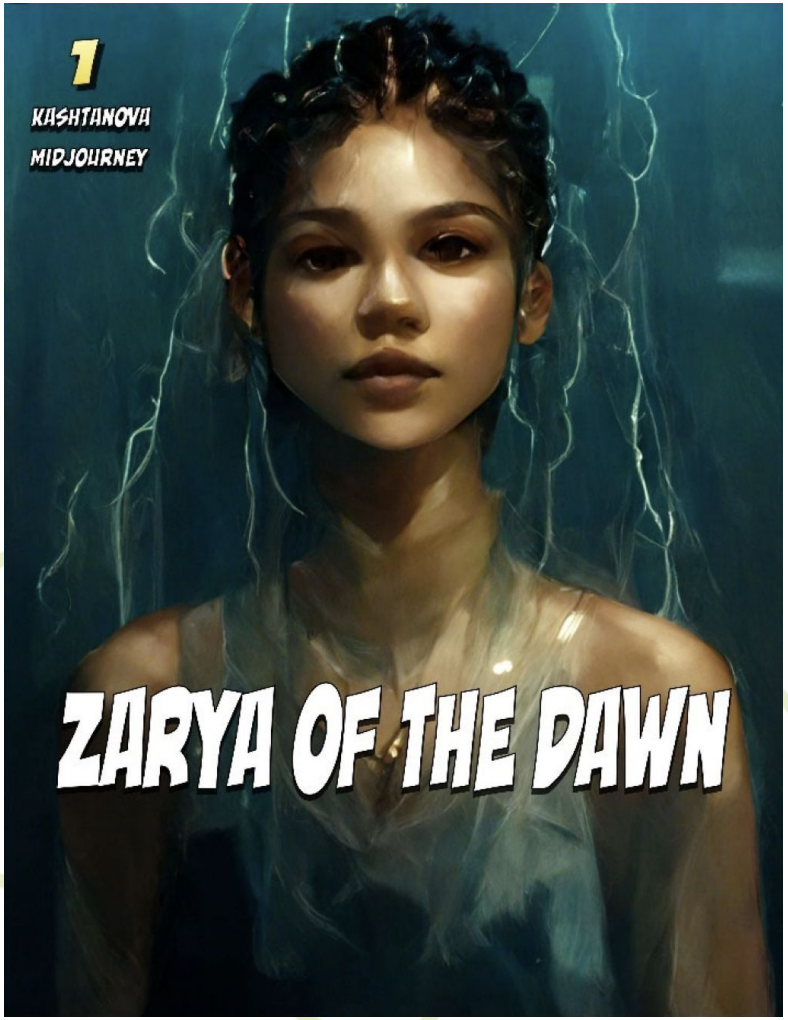
No human authorship

- Copyright = author's right (*droit d'auteur*)
  - author: 'a person who begins or creates something' (Cambridge Dictionary)
- Required by the Berne Convention
  - related to "author": nationality, honor, death
- Required by EU law
  - Copyright Term Directive 2006: 'death', 'natural person who have created the work' (Art. 1)
- Confirmed by CJEU case law
  - Painer (2011): 'only human creations are (...) protected'
- Copyright is a *human* right
  - Art. 27(2) UDHR: 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.
- No human author = no copyright

No originality

- Originality according to CJEU
  - author's own intellectual creation (Infopaq, 2009)
  - intellectual creation of the author reflecting his personality and expressing his free and creative choices (Painer, 2011)
  - no originality if "technical considerations, rules or constraints which leave no room for creative freedom" (Football Dataco, 2012)
  - no alternative criteria (e.g. aesthetic effect) can be applied to grant copyright protection (Cofemel, 2019)
- No originality (author's personality) in AI-generated works
- Argument: It would be unfair vis-à-vis 'real creators' to protect AI-generated works

They already know it in the US



- US Copyright Office Statement of Policy (16 March 2023)

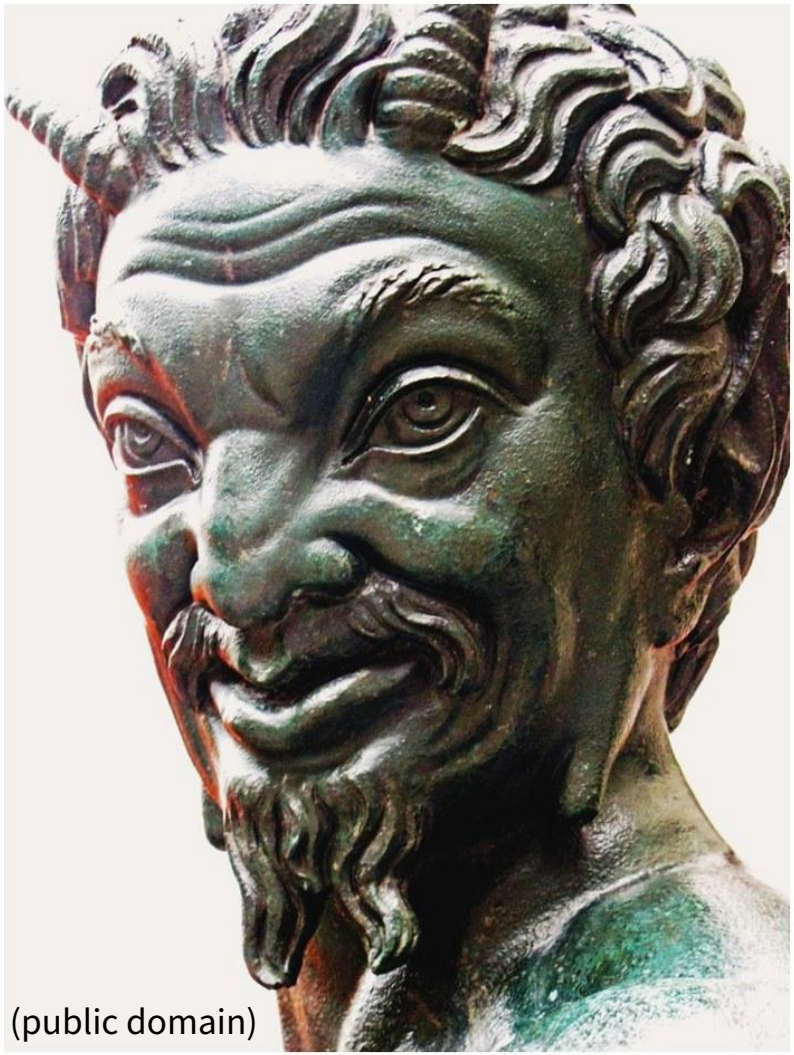
(public domain)

Of course, AI outputs are not protected by copyright.

Of course.

But... maybe they should?

- copyright can be held by legal entities
- there is always a human intervention
- AI outputs are not generated ex nihilo
- AI outputs need rules



(public domain)

Copyright of non-human actors

- Corporate authorship (work for hire)
  - work created by an employee in the course of employment
  - US, UK, Ireland
  - expressly allowed by EU law (Copyright Term Directive)
  - standard solution for software (also in the EU)
- Even in France: oeuvre collective
  - created at the initiative of a legal entity
- Non-human initial ownership of copyright is widely accepted

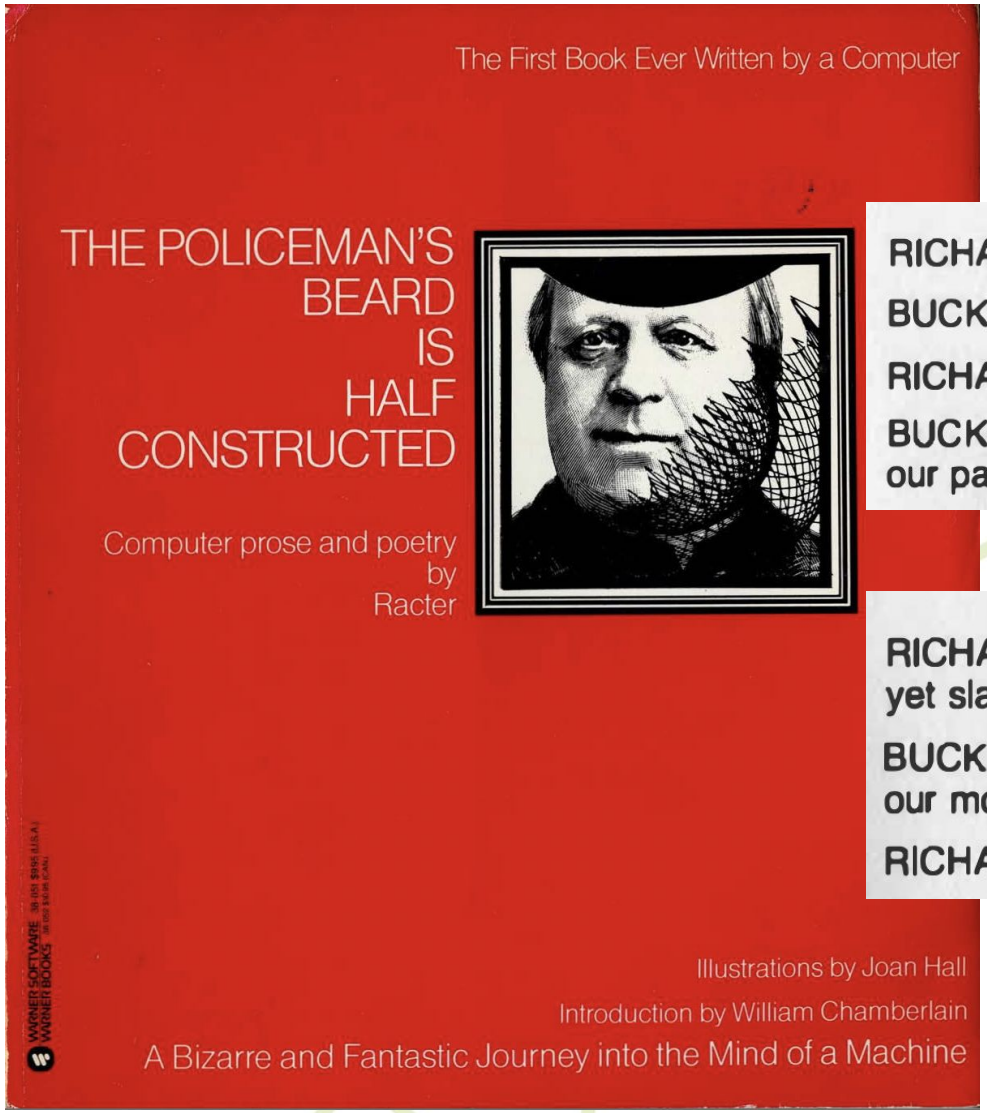
There is always a human ‘author’

- Author: ‘a person who **begins** or creates something’ (Cambridge Dictionary)
- AI-generated vs. AI-assisted works
- AI is just a tool (cf. a camera)
  - 1839: beginning of practical photography
  - 1862 (France), 1884 (US): copyright protection of photographs (choice)
- There is always a human who makes a choice
  - no ‘personality’? Is there more personality in a salad shaker or a 4-word slogan?

AI does not generate *ex nihilo*

- AI-generated outputs as derivative works?
- CJEU, Infopaq (2009): 11-grams can be protected by copyright
- LLMs can ‘regurgitate’ large chunks of training data
- Getty Images vs. Stable Diffusion lawsuit

Practical considerations



RICHARD. A week is obscurely like a night.

BUCKINGHAM. My Lord, chicken is like lamb.

RICHARD. Yet weeks can be killed as can chicken.

BUCKINGHAM. Tis true, my Liege, yet ambiguities adorn our pain as ambiguities broaden our issues.

RICHARD. Noble brother, thy tale is furious, yet slaughtering attorneys in truth is essential.

BUCKINGHAM. Good prince, measuredly I think that our months are shortened by the millisecond.

RICHARD. Deepen your pondering, good brother.

Practical considerations

- AI-generated texts have economic value
- Transactions need rules
  - cf. related right in unoriginal photographs (in Germany)
- How to prevent “copyfraud” of AI-generated texts?
- We need property rights!

They already know it in Brussels (don't they?)

European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI))

15. Takes the view that technical creations generated by AI technology must be protected under the IPR legal framework in order to encourage investment in this form of creation and improve legal certainty for citizens, businesses and,

of 'intellectual creation' addresses the author's personality; calls on the Commission to support a horizontal, evidence-based and technologically neutral approach to common, uniform copyright provisions applicable to AI-generated works in the Union, if it is considered that such works could be eligible for copyright protection; recommends that ownership of rights, if any, should only be

Member States should have implemented this Directive by June 2021. At this stage, the Commission believes that the creation of art works by AI does not deserve a specific legislative intervention. Therefore, the Commission is not planning to revise this Directive.

Thank you for your attention

AI technologies: ownership of AI data inputs

Thomas Margoni

Research Professor of Intellectual Property Law
Centre for IT & IP Law (CiTiP), Faculty of Law
University of Leuven (KUL)

RecreatingEU project: Scope and research question

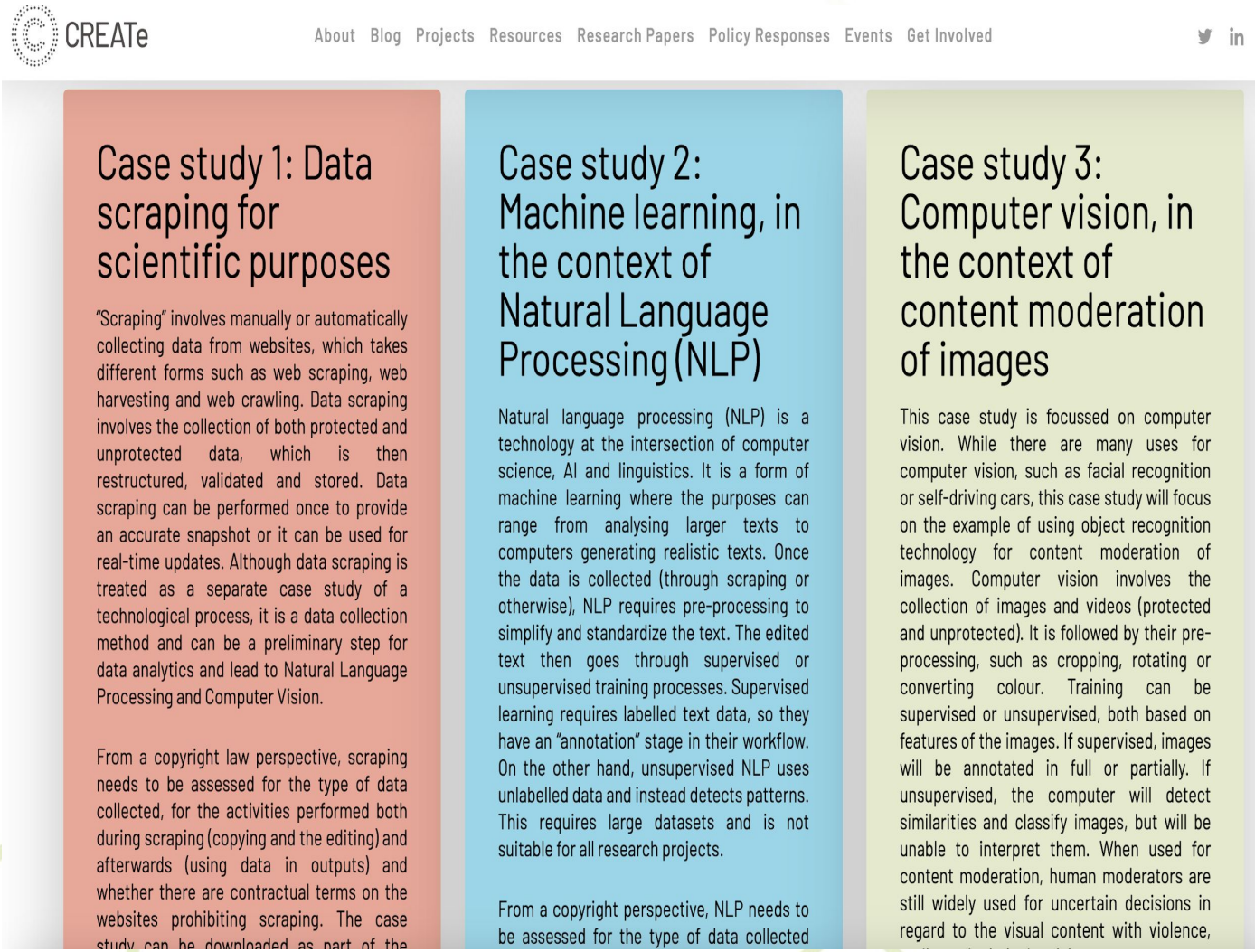
- Examine the role of “training data” in AI from copyright perspective
- Classify the relationship between training/input data, ML processes and final trained models in positive and normative terms.
- Assess protection of AI inputs (subject matter, Art. 2 ISD, SGDR) and relevant exceptions for TDM (Arts. 3&4 CDSM, 5(1),(3)(a),(d) ISD);
  - *Is the regulation of non personal data a form of (direct or indirect) regulation of AI?*

Methodology

- Reverse inductive strategy based on case studies of technological processes followed by legal analysis;
- Case studies developed in consultation with stakeholders: (i) Data scraping; (ii) Machine learning for NLP; (iii) Computer vision for content moderation of images + expert workshop
- Legal analysis informed by results of case studies highlights use of technology as regulatory lever and need of governance frameworks for data and digital technologies.

Methodology

Example of the 3 case studies employed as basis for legal analysis



Findings

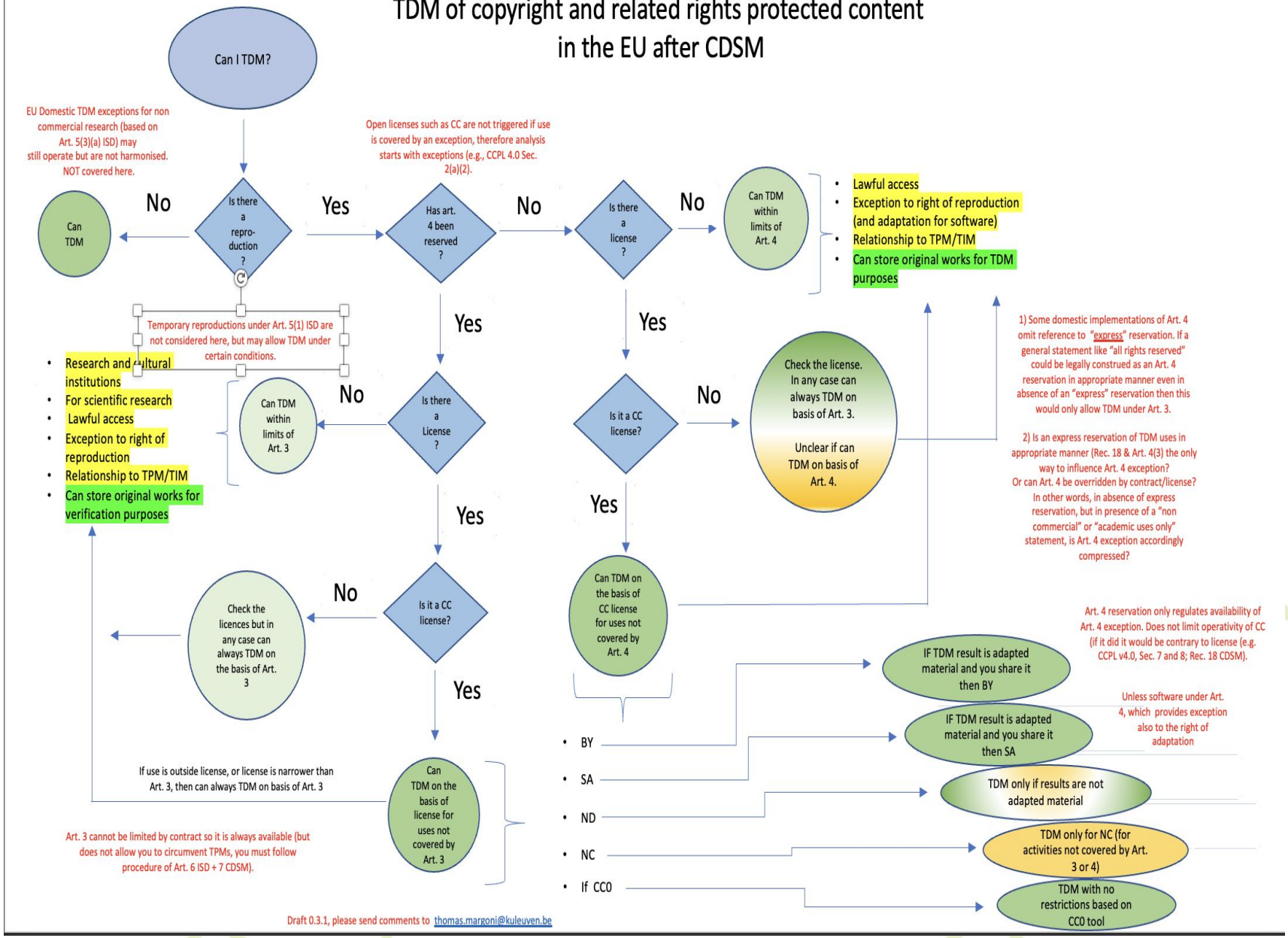
– **Arts. 3&4 CDSM as property-based regulatory framework for AI development**

Data is the enabler of most digital activities (not only, but chiefly AI), therefore regulating data operates as a proxy for regulating the technologies based on data (e.g. AI).

– Does this mean that EU AI development, arguably a fundamental industrial policy issues, depends on 2 and ½ copyright exceptions?

- If yes, this effect is likely beyond the drafter's intentions and copyright's remit ("must allow and ensure the development … of new technologies and safeguard a fair balance between the rights of right holders and of users").
- Margoni&Kretschmer, GRUR Int. 2022

TDM of copyright and related rights protected content in the EU after CDSM



Findings

– **Implementation of technology-enabling exceptions not homogenous**

Implementation of Arts. 3&4 (verifiability copies; CTP; express reservation);

- "Creative" combination of Arts. 3&4 + Art. 5(3)(a)
  - Italian implementation may even suggest derivatives are covered by exception!?! "***nonche' la comunicazione al pubblico degli esiti della ricerca ove espressi in nuove opere originali"***
  - German implementation allows making available (to public?) under limited conditions for NC research (Sec 60d(4).
- Domestic interpretation of Art. 5(1)
  - Ensure that allows … rights of users to avail themselves of technological deveopment
- Art. 3 how far does it reach?
  - Can independent commercial company use Art. 3 trained model? See Getty-StableAI-Laion

Findings

– Developments of new services that offer automated “opt-out” from training (independent but in principle compatible with Art. 4 CDSM, see e.g., https://spawning.ai/

  - “There are no guarantees that copyright will be sufficient to protect artists from AI training, so we have little choice but to operate assuming it will not … what is within our grasp is the ability to conceive of a new era **of consenting interactions around artist data”.**

  - Is “artist data” a copyright category? Should it be? What does it cover, genere artist data; individual artist data?

  - Towards a right of remuneration for training? How? Collective management?

Findings

– Higher “costs” in EU AI development

- Higher costs (transactive; legal certainty; monetary) of EU TDM may create situations where it is economically or opportunistically attractive to develop AI applications in “cheaper” legal systems or to import into the EU already pre-trained models.
  - “data laundering” v ”data commons”?

- Incentives towards opacity in data training (if you document your training data you may offer evidence of possible © infringement; see Stable Diffusion). OpenAI not so open after v2.

- Property rights in data may incentivise obscurity and unaccountability in AI training (Margoni; Quintais; Schwemer, 2023).

“Getty Images also licenses the use of its visual assets and associated metadata in connection with the development of artificial intelligence and machine learning tools. Getty Images has licensed millions of suitable digital assets … for a … artificial intelligence and machine learning”

“the Stable Diffusion model frequently generates
output bearing a modified version of the Getty Images watermark … ”

”Stability AI has caused the Stable Diffusion model to incorporate a modified version of the Getty Images’ watermark to bizarre or grotesque synthetic imagery that tarnishes Getty Images’ hard-earned reputation, such as the image below”

Getty Images v Stability AI Complaint 2023

Findings

– **Property-based and governance-based approaches to the regulation of data may not necessarily be convergent**

- "Governance-based approach" to the regulation of data (e.g., DGA, DA, Open Data, CEDS etc) based on a different paradigm where key words are data rights, access, and portability.
- Art. 35 DA illustrates potential tension and need to reform data regulatory environment.
- Topology of data… : High-Value Data Sets, PSBs documents, research data (ODD); IoT data (DA); non personal data (DGA), Personal data (GDPR); sectorial legislation; but also "artists data", etc
- … and of data rights: co-creation, access, portability, sharing, altruism,
- Ducuing et al 2022

Findings

– **Common European Data Spaces as the interface between property-based approaches and governance-based approaches in EU data markets**

- Creation of markets and infrastructures based on data rights (access, transfer, portability, B2G), and of interfaces with existing IP rights, e.g., TS (and right of suspension) SGDR, Arts. 3&4, etc.
- Data intermediaries to facilitate public and private data transactions;
- Remaining open issues connected to a lack of proper legal theoretical classification of non personal data. ”Stickiness” of property rights difficult to avoid in unclear cases (e.g., TPM; authorisations; control, contractual allocation of rights mimicking property entitlements).
- Margoni et al 2023 forthcoming in Computerrecht

The UK did it before it was cool.

Toby Bond, Bird & Bird

When does copyright protect language?

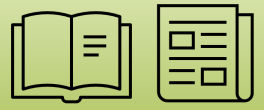
Unlikely to be protected by copyright

Borderline cases

Likely to be protected by copyright

42, EXXON, the price of vegetables is increasing

Iceberg lettuce hit by titanic price rise



Individual facts, numbers, single words

Presentations of facts, newspaper headlines, book titles

Poems, stories, newspaper articles

Dictated by technical constraints, no free and creative choices

"[i]n order for an intellectual creation to be regarded as an author's own it must reflect the author's personality, which is the case if the author was able to express [their] creative abilities in the production of the work by making free and creative choices."

Summary of earlier CJEU case law given in Funke Medien (C-469/17)

Free and creative choices expressing an author's personality

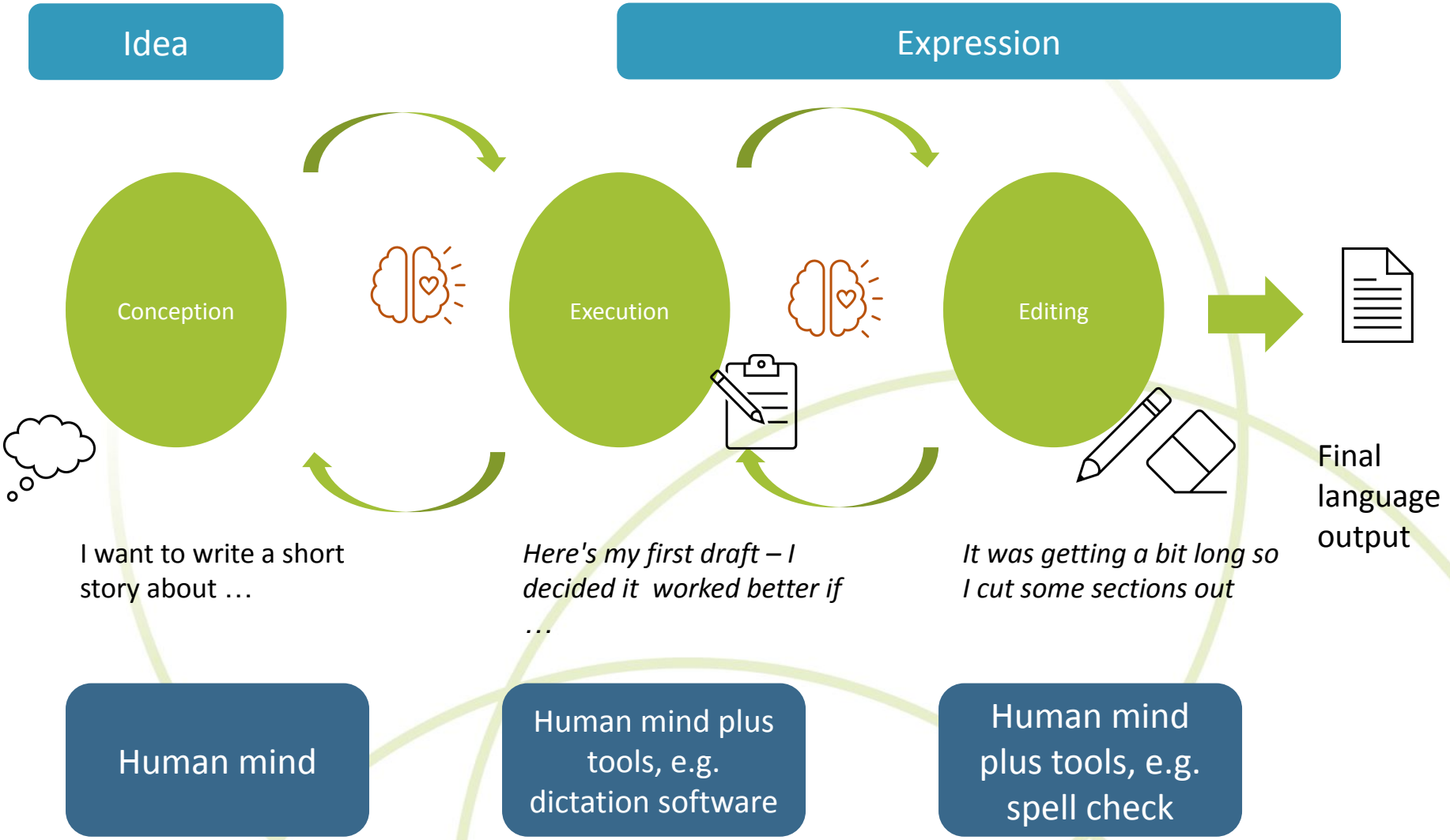
No intellectual creation

Originality

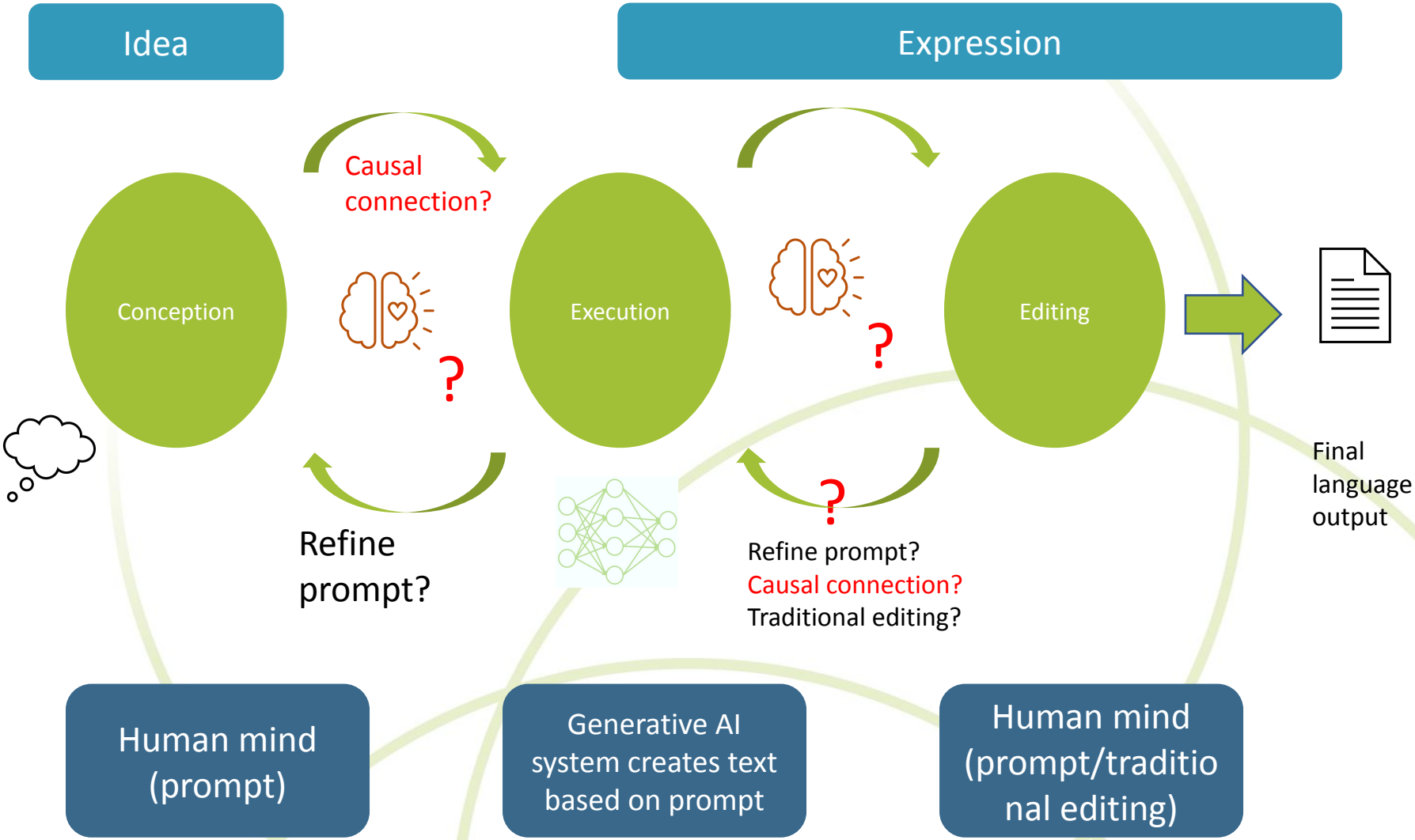
Intellectual creation

The creation process: without Generative AI

Diagram adapted from Hartmann, C., Allan, J., Hugenholtz, P., et al., Trends and developments in artificial intelligence challenges to the intellectual property rights framework : final report, Publications Office, 2020 , https://data.europa.eu/doi/10.2759/683128



The creation process: with Generative AI



US Copyright Office Guidelines (March 2023)

- Copyright can protect only material that is the product of human creativity, i.e. there is a human authorship requirement for registration. The Office will begin by asking whether the 'work' is:

  (1) basically one of human authorship, with the computer [or other device] merely being an assisting instrument, or

  (2) whether the traditional elements of authorship in the work (literary, artistic, or musical expression or elements of selection, arrangement, etc.) were actually conceived and executed not by man but by a machine."

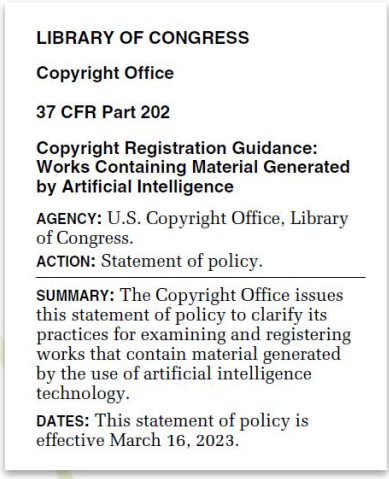
Application to AI-generated material

- For work containing AI-generated material are the AI contributions (i) the result of mechanical reproduction or; (ii) an author's own original mental conception, to which [the author] gave visible form? When an AI technology determines the expressive elements of its output, the generated material is not the product of human authorship.

- When an AI technology receives solely a prompt from a human and produces complex written, visual, or musical works in response, the ''traditional elements of authorship'' are determined and executed by the technology—not the human user. The prompts function more like instructions to a commissioned artist.

- A human may select or arrange AI-generated material in a sufficiently creative way that ''the resulting work as a whole constitutes an original work of authorship.''

- Or an artist may modify material originally generated by AI technology to such a degree that the modifications meet the standard for copyright protection. In these cases, copyright will only protect the human-authored aspects of the work, which are ''independent of'' and do ''not affect'' the copyright status of the AI-generated material itself.



LIBRARY OF CONGRESS

Copyright Office

37 CFR Part 202

Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence

AGENCY: U.S. Copyright Office, Library of Congress.

ACTION: Statement of policy.

SUMMARY: The Copyright Office issues this statement of policy to clarify its practices for examining and registering works that contain material generated by the use of artificial intelligence technology.

DATES: This statement of policy is effective March 16, 2023.

What about the UK's provisions on computer-generated works (and similar provisions in Ireland, India, New Zealand, South Africa)?

Copyright, Designs and Patents Act 1988

s1(1) *Copyright is a property right which subsists in accordance with this Part in the following descriptions of work— (a)* ***original*** *literary, dramatic, musical or artistic works, …*

s9(3) *In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken.*

s178 *"computer-generated", in relation to a work, means that the work is generated by computer in circumstances such that there is no human author of the work;*

Term of protection: 50 years from creation

UKIPO Consultation (October 2021)

From a legal perspective, a computer-generated work must be original if it is to receive protection. But the legal concept of originality is defined with reference to human authors and characteristics like personality, judgement and skill. It has been argued that the law is unclear and contradictory.

Proposals for consultation

Option 0: Make no legal change

Option 1: Remove protection for computer-generated works

Option 2: Replace the current protection with a new right of reduced scope/duration

UK Government response to consultation (June 2022)

21. We have decided to adopt Option 0: make no changes to the law. There is no evidence at present that protection for CGWs is harmful, and the use of AI is still in its early stages. As such, a proper evaluation of the options is not possible, and any changes could have unintended consequences. But we will keep the law under review and could amend, replace or remove protection in future if the evidence supports it.

Did we decide whether the law is unclear and contradictory?

What do terms and conditions say about ownership and use of generated materials?

Extracts from T&Cs of LLMs as a service

Chat GTP (https://openai.com/policies/terms-of-use)

Restrictions. *You may not (i) use the Services in a way that infringes, misappropriates or violates any person's rights; (ii) reverse assemble, reverse compile, decompile, translate or otherwise attempt to discover the source code or underlying components of models, algorithms, and systems of the Services (except to the extent such restrictions are contrary to applicable law); (iii) use output from the Services to develop models that compete with OpenAI …*

Your Content*. …. As between the parties and to the extent permitted by applicable law, you own all Input. Subject to your compliance with these Terms, OpenAI hereby assigns to you all its right, title and interest in and to Output. This means you can use Content for any purpose, including commercial purposes such as sale or publication, if you comply with these Terms. …*

Bing (https://www.bing.com/new/termsofuse)

Using the Online Services. *… You must use the Online Services and the generated Creations only (i) in a lawful manner and in compliance with all applicable laws; … and (iii) in a manner that does not infringe or attempt to infringe, misappropriate or otherwise violate any of our rights or those of any other person or entity ….*

Use of Creations. *Subject to your compliance with this Agreement, the Microsoft Services Agreement and our Content Policy, you may use Creations outside the Online Services for any legal personal, non-commercial purpose.*

Ownership of content. *Microsoft does not claim ownership of Captions, Prompts, Creations or any other content you provide, post, input, or submit to, or receive from, the Online Services (including feedback and suggestions). However, by using the Online Services, posting, uploading, inputting, providing or submitting content you are granting Microsoft, its affiliated companies and third-party partners permission to use the Captions, Prompts, Creations and related content in connection with the operation of its businesses (including, without limitation, all Microsoft Services).*

Bard (https://policies.google.com/terms/generative-ai)

Use restrictions *You may not use the Services to develop machine learning models or related technology.*

In addition to the "Respect others" section in the Google Terms of Service, you must comply with our Prohibited Use Policy, which provides additional details about appropriate conduct when using the Services.

What do terms and conditions say about ownership and use of AI generated materials?

*Licence terms for LLMs (*en.wikipedia.org/wiki/Large_language_model)

Model Name	License	Model Name	License
BERT	Apache 2.0	PaLM (Pathways Language Model)	Proprietary
GPT-2	MIT	OPT (Open Pretrained Transformer)	Non-commercial research
GPT-3	public web API	YaLM 100B	Apache 2.0
GPT-Neo	MIT	Minerva	Proprietary
GPT-J	Apache 2.0	BLOOM	Responsible AI
Megatron-Turing NLG	Restricted web access	AlexaTM (Teacher Models)	public web API
Ernie 3.0 Titan	Proprietary	LLaMA (Large Language Model Meta AI)	Non-commercial research
Claude	Closed beta	GPT-4	public web API
GLaM (Generalist Language Model)	Proprietary	Cerebras-GPT	Apache 2.0
Gopher	Proprietary	Falcon	Proprietary
LaMDA (Language Models for Dialog Applications)	Proprietary		
GPT-NeoX	Apache 2.0		
Chinchilla	Proprietary		

Apache 2.0: permissive licence allowing the creation and use of derivative works subject to notice requirements (query whether model outputs are "derivative works").

MIT: no restrictions on use.

Opt-175b License Agreement: No use of model or outputs for *(i) any commercial or production purposes, … (v) in any manner that infringes, misappropriates, or otherwise violates any third-party rights.*

Practical problems

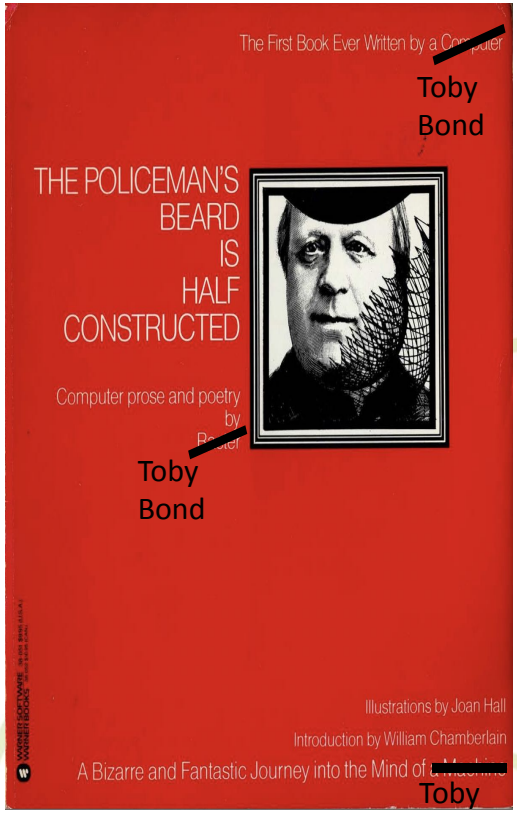
How do we know when text has been generated by AI?

She is a friend
That keeps me company at night
A gentle presence

She listens to me
As I pour out my feelings
A faithful confidant

She smiles at me
And fills my heart with hope
A source of light

Mrs Moon
sitting up in the sky
little old lady
rock-a-bye
with a ball of fading light
and silvery needles
knitting the night



What's the difference?
AI generated

- No copyright protection in the EU.
- In the public domain and available for reuse (subject to any contractual restrictions).
- Exceptions to copyright do not apply.

Human authored

- Protected by copyright for the life of the author + 70 years.
- Reuse infringes copyright (subject to exceptions).

Panel Discussion

https://www.menti.com/alc3oe2z3s98

Getting involved in CLARIN

Join our NewsFlash

https://www.clarin.eu/content/newsflash

Check out our events

https://www.clarin.eu/events

Open calls

https://www.clarin.eu/content/funding-opportunities

https://www.clarin.eu/event/2023/clarin-annual-conference-2023

Next events

Stay tuned: https://www.clarin.eu/content/clarin-cafe

Share your **#clarincafe** impressions with **@CLARINERIC**