DDM/:SME



Legal Access to Data:

Open Data & Open Data Licensing Text and Data Mining

Content – Legal Access to Data



- Open Data and Open Data Licensing
- Text and Data Mining (incl. for Al training)



Open Data and Open Data Licensing



Ideology



- The always mentioned quote
 - "The best thing to do with your data will be thought of by someone else" (Rufus Pollock)
 - Not only economic, but also "societal value" – citizen engagement

Ideology



Recital 8 ODD: "Documents produced by public sector bodies of the executive, legislature or judiciary constitute a vast, diverse and valuable pool of resources that can benefit society. Providing that information, which includes dynamic data, in a commonly used electronic format allows citizens and legal entities to find new ways to use them and create new, innovative products and services."

The Regulatory Landscape



- Access / Re-use Interrelated
- Access >>> National law
- Re-use >>> EU law

Access to Data – What? Issues



- Issues
 - Obliged subject Who?
 - Object of disclosure What?
 - Manner of disclosure How?
 - Claims to access How?
 - Remedies Well, what if not...

Access to Data – What? Law



- European Convention on Human Rights (1950), Art. 10 para. 1
 - Freedom of expression
 - freedom [...] to receive and impart information and ideas without interference by public authority [...]
- Decision of the Grand Chamber of European Court of Human Rights from 8. 11. 2016, No. 18030/11 (*Magyar Helsinki Bizottság v. Hungary*)
- ▶ However limits Art. 10 para. 2 ECHR
- Art. 19 of International Covenant on Civil and Political Rights
- Council of Europe: Convention on Access to Official Documents (CETS No 205; Tromsø, 18 June 2009)

Basic Principles



- Declaration of the Committee of Ministers of the Council of Europe on Freedom of Expression of Information of 29 April 1982
 - Right to receive information held by public authorities
 - No need to prove the reason
 - Principle of equality
 - Access within reasonable time
 - Reasoned withholding
 - Reviewable withholding

EU Law



- Charter of Fundamental Rights and Freedoms of the European Union
- Article 11: Freedom of expression
- Article 42: Right of access to documents of the Union's institutions, bodies and other entities
- Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents

Re-use of data – HOW?



- Issues
 - Possibilities of disposition
 - Limitations
 - Obligations
 - Ease of use (formats and legal conditions)

Re-use of data – HOW?



- ▶ 1989 Guidelines for improving the synergy between the public and the private sectors in the information market
 - in a reusable form through electronic information services
- Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information
 - Workable legal environment for the transparent and nondiscriminatory use of public sector information published for this purpose.
 - IF PROVIDED THEN IN THIS WAY
 - Both for voluntary as well as mandated access

Re-use of data – HOW?



- , Open Data Directive" Directive (EU) 2019/1024 of 20 June 2019 on open data and the re-use of public sector information
 - "establishes a set of minimum rules govern ing the re-use and the practical arrangements for facilitating the re-use" (A1 ODD)
 - non-commercial and commercial purposes

The Aim of ODD



- reduce market entry barriers (SMEs)
- reduce first-mover advantages
- promote transparency and accountability
- R 13, 14, 36, 51 ODD
- https://coraxfoundation.com/2021/10/12/the-open-data-directive-what-the-introduction-of-high-value-datasets-means-for-data-re-users/

Basic Principle A5



- "public sector bodies and public undertakings shall make their documents available in any pre-existing format or language and, where possible and appropriate, by electronic means, in formats that are open, machine-readable, accessible, findable and re-usable, together with their metadata.
- Both the format and the metadata shall, where possible, comply with formal open standards

ODD – Exceptions A1/2 ODD



- c) IPR
- d) Sensitive data (national security, statistical confidentiality, business secrets)
- h) personal data

Basic Terms



- A2/3 ODD Document "any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording)" or part of such content
 - held by public sector bodies (national, regional and local level), publicly funded libraries, research organisations and other cultural heritage bodies

Basic Terms



- Public undertakings mostly funded by or under the control of public bodies (dominant influence by funding and governing)
- Formal open standards
 - "standard which has been laid down in written form, detailing specifications for the requirements on how to ensure software interoperability"
 - Not official community-driven

Single Point of Access



- Practical arrangement
- ▶ A9 make the datasets findable and re-usable
- E.g., CZ: https://data.gov.cz/english/
- **EU-WIDE**

A6 Costs



- Basis: Free of charge
- ▶ A6/1 "the recovery of the marginal costs incurred for the reproduction, provision and dissemination of documents as well as for anonymisation of personal data and measures taken to protect commercially confidential information"
- + extensive search costs
- A6/2/b, A6/5, R38 Cultural institutions, including libraries, archives and museums: "charge above marginal costs in order not to hinder their normal running"

A10 ODD Research Data



- PUBLICLY FUNDED DATA!
- FAIR principles (Findability, Accessibility, Interoperability, and Reuse)
- Open is standard as closed as neccessary
- ▶ A10/1 ODD: "National policies and relevant actions aiming at making publicly funded research data openly available ('open access policies')"
- A2/9 ODD: not scientific publications (i.e., only underlying data)
- Open repositories https://about.zenodo.org/

High Value Data Sets



- A2,,documents the re-use of which is associated with important benefits for society, the environment and the economy"
- A14 free of charge, in machine-readable format and via APIs, and, where relevant, as a bulk download
- Six thematical areas:
 - geospatial, earth observation and environment, meteorological, statistics, companies and company ownership and mobility
 - could be expanded by Commission
 - Commission Implementing Regulation (EU) 2023/138 of 21 December 2022 laying down a list of specific high-value datasets and the arrangements for their publication and re-use
- https://data.europa.eu/data/datasets?is_hvd=true&locale=en

Open Data



- ▶ HOW? the data are provided
- A5/2 ODD: "open by design and by default"
- Open data >>> technologically efficient way of providing and re-using public sector information

Open Content Definition v2



- RETAIN
- **REUSE**
- REVISE
- REMIX
- REDISTRIBUTE

David Wiley, http://opencontent.org/definition/

Open Definition



- New: "Knowledge is open if anyone is free to access, use, modify, and share it subject, at most, to measures that preserve provenance and openness".
- Source: http://opendefinition.org/od/2.1/en/

No SGDR database



A1/6 ODD: "The right for the maker of a database provided for in Article 7(1) of Directive 96/9/EC shall not be exercised by public sector bodies in order to prevent the reuse of documents or to restrict re-use beyond the limits set by this Directive."



Open Data Licensing

DDM/:SME

Open Data Licensing



- Rule A8 ODD "re-use of documents shall not be subject to conditions, unless such conditions are objective, proportionate, non-discriminatory and justified on grounds of a public interest objective"
- Standard licences are encouraged
- ▶ A2/5 ODD standard licence "a set of predefined re-use conditions in a digital format, preferably compatible with standardised public licences available online"

Open Licences



Commission notice — Guidelines on recommended standard licences, datasets and charging for the reuse of documents (52014XC0724(01))

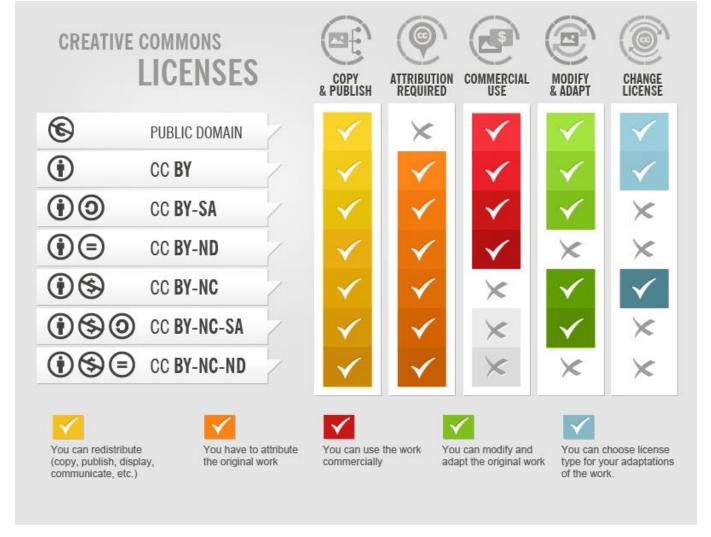
Public licences – characteristics



- Allow sharing (modification)
- Under specific conditions
- Always attribution
- Irrevocable
- Automatic termination upon breach
- Creative Commons CC0 waiver/public domain mark
- https://creativecommons.org/choose/
- LEGO-approach modular licensing terms
- Fully enforceable (German courts: CC are valid licensing terms no issue)

Creative Commons Licences





JoKalliauer; foter, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons

Data Governance Act



Filling the gaps of Open Data Directive

Basic idea



- Public sector bodies hold vast amounts of data protected by third-party rights (such as trade secrets, personal data or intellectual property) that cannot be used as open data but that could be reused under specific EU or national rules. Whenever such reuse is allowed, public sector bodies will have to comply with the reuse conditions laid down by the DGA. Notably, the reuse conditions should be non-discriminatory, transparent, proportionate, justified and made publicly available.
- https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32022R0868#keyterm_E0001

General principles



- Part of the European Strategy for Data
- No obligation to share data!
- Enabling the re-use and sharing of "protected data" by introducing safeguards (Articles 3-8 DGA)
 - "secure processing environment", "disclosure control"
- Data intermediaries (Articles 9-14 DGA)
- "Data altruism" (Articles 15-22 DGA)



Text and Data Mining (incl. for Al training)

DDM/:SME

The need of use to data



- Innovation
- Access v Control [Rec 8 DSM-Dir]
- Incetives & Investment > thwarted by simple access [e.g. rec. 7 Database-Dir]
 - (7) Whereas the making of databases requires the investment of considerable human, technical and financial resources while such databases can be copied or accessed at a fraction of the cost needed to design them independently;

Legal Control of Data



- Copyright = original authorial works
 - Even 11 words (*Infopaq*)
 - Reproduction
- Sui generis protection for databases = investment in obtaining, verification and presentation data (not data creation itself)
 - Extraction and re-utilization

Pro rightholders

Infringing

- Factor 1: The purpose of the use is commercial: companies like OpenAl operate for profit, offering paid versions of their systems
- Factor 1: Most systems have no proper safeguards against the production of output that infringes copyright
- Factor 1: The output of the systems is nothing more than derivate works of the works used in the training
- Factor 2: The works used to feed the systems are often of a highly creative nature: illustrations, photographs, books, music, etc.
- Factor 3: The systems are fed with huge numbers of protected works, and usually the works in their entirety

Pro Gen Al

Non-infringing



- Factor 1: The purpose of the use is transformative: protected works are used only as training data, with the aim of producing entirely new creations
- Factor 1: Gen Al systems are not designed to memorize existing works and reproduce them, but to generate new output
- Factor 2: The works used for training have already been published and are freely accessible on the Internet
- Factor 3: Copying the works in their entirety is necessary to achieve the transformative purpose
- Factor 3: Making full copies of the works during training is irrelevant if they are not made available to the public

Pro rightholders

Infringing

- Factor 4: Works are being used without remuneration, despite many rightholders offering licenses for Al training (e.g. Getty Images)
- Factor 4: Content produced by Gen AI competes in the same market as the human-created works with which these tools are fed, reducing incentives for human creativity
- Other: Most Gen AI systems have no opt-out mechanisms

Pro Gen Al

Non-infringing



- Factor 4: The use of works as training data does not compete in the market for works as expressive creations
- Other: The learning process of Gen Al systems is similar to that of humans when reading works
- Other: If the systems were trained only with works in the public domain, they would inevitably be biased (in favour of the dominant ideologies of the past)
- Other: One of the aims of fair use is to promote technological innovation



Legal Access to Data: TDM Art 2 (2) DSM-Dir





Any automated analytical technique

automated computational analysis [Rec 8 DSM-Dir] → technology-neutral



aimed at analysing text and data in digital form

such as text, sounds, images or data [Rec 8 DSM-Dir] → media-neutral



in order to **generate** information

including (but not limited to) patterns, trends and correlations
i.e., not only storing

EU: Art. 3 DSMD – TDM for scientific research



- I. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.
- ▶ 2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.
- 3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.
- 4. Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.

Art 3 DSM-Dir: TDM for Scientific Research



- Reproductions
 - Databases [Art 5 (a) and Art 7 (1) Database-Dir], Works, Performances, Sound recordings, Films [Art 2 Info-Dir], Press publications [Art 15 (1) DSM-Dir]
 - Computer programs not included
- Storing (appropriate level of security) and retaining for scientific research (incl. verification of results)
- Purpose = Scientific research
- Lawful access
- Further
 - Contract immunity [Art 7 (1) DSM-Dir]
 - Measures to ensure the security and integrity of networks/databases
 - TPM: claim for removal [Art 7 (2) DSM-Dir] > Art 6(4) Inf-D]

EU: Art. 4 DSMD - "unlimited" TDM



- ▶ 1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.
- 2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.
- > 3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

Art 4 DSM-Dir: TDM for Other Purposes



- Beneficiaries: Not restricted
 - Researchers not affiliated to research organisation
 - Commercial research, commercial enterprises, authorities etc.
 - → Data-driven management, Al
- Purpose: TDM
- Scope
 - Reproductions
 - Databases [Art 5 (a) and Art 7 (1) Database-Dir], Works, Performances, Sound recordings, Films [Art 2 Info-Dir], Press publications [Art 15 (1) DSM-Dir], Computer programs [Art 4(1)(a) and (b) Software-Dir]
 - Retaining as long as necessary for TDM
- Lawful access
- Restrictions
 - Expressly reserved in appropriate manner (e.g. machine-readable)
 - https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/
 - https://spawning.ai/
 - TPM: claim for removal

TDM: Lawful access requirement



- Art 3 + Art 4 DSM-Dir
- Lawful...
 - USE [Art 5 (1) Info-Dir]
 - Source [CJEU ACI Adam, CJEU Copydan]
 - Access
- Lawful Access [Rec 14 DSM-Dir]
 - Open access policy
 - Contractual arrangements (e.g. subscriptions) or
 - Other lawful means (freely available online)
- So obvious that in some countries (CZ) this requirement is fully left out, as it is already incorporated in the three-step test
- Works made available without consent of rightholders? [CJEU Film spelet]
 - NOPE

Further Exceptions for TDM



- Art 25 DSM-Dir → Info-Dir
- Art 5 (1) Info-Dir [Rec 9 DSM-Dir]
 - Temporary reproduction + transient or incidental still applies
 - Cache, RAM
 - Storing and deleting after TDM
 - Commercial TDM
 - Lawful use
 - Restriction
 - Explicit reservation according to Art 4 DSM-Dir [Rec 18 DSM-Dir]
- Art 5 (3) a Info-Dir
 - Non-commercial research
 - Researchers not affiliated to research organisation
 - Reproduction and making available to the public (if justified by research purpose)

The problems / Criticism



- (1) definition
- (2) beneficiaries
- (3) rights
- ▶ (4) technological overridability, and
- ▶ (5) access to original sources
- MARGONI, Thomas and KRETSCHMER, Martin, 2022. A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*. 1 August 2022. Vol. 71, no. 8, p. 685–701. DOI 10.1093/grurint/ikac054.
- https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/ OPTING OUT

Al Act



- Recital 105: TDM is Al training
- Recital (63) 'general-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market
- Transparency obligation
- Art. 53 para. 1 let. c): "put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790";

Al Act - Recital 107



- "it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model"
- "this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used"
- **▶** Template by AI Office

Al Act - Recital 106



,,Any provider placing a general-purpose Al model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place. This is necessary to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union."



A short glimpse to the USA

DDM/:SME

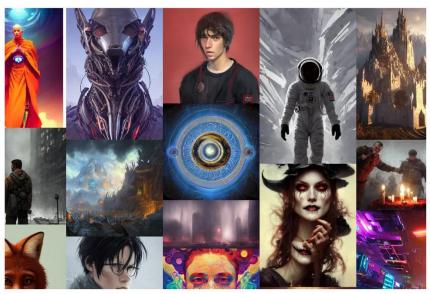
17 U.S. Code § 107 - Limitations on exclusive rights: | Tair use | LAW | Pair | Law | Law

- Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—
- (1) the purpose and character of the use, <u>including</u> whether such use is of a commercial nature or is for nonprofit educational purposes;
- (2) the nature of the copyrighted work;
- (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- (4) the effect of the use upon the potential market for or value of the copyrighted work.
- The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart



Al art tools Stable Diffusion and Midjourney targeted with copyright lawsuit



A collage of AI-generated images created using Stable Diffusion. Image: $\underline{\textit{The}}$ $\underline{\textit{Verge via Lexica}}$

The suit claims generative AI art tools violate copyright law by scraping artists' work from the web without their consent.

By JAMES VINCENT

Jan 16, 2023, 12:28 PM GMT+1 | 28 Comments / 28 New





https://stablediffusionlitigation.com/pdf/00201/1-1-stable-diffusion-complaint.pdf



https://githubcopilotlitigation.com/

/ Microsoft, GitHub, and OpenAl

reproducing open-source code using Al. But the suit could have a huge impact on the wider world

are being sued for allegedly violating copyright law by

of artificial intelligence.

USA: Class action against GitHub, Microsoft (owner of GitHub), and OpenAl



The Verge / Tech / Reviews / Science / Entertainment / More +

ARTIFICIAL INTELLIGENCE / TECH / LAW

The lawsuit that could rewrite the rules of Al copyright "software piracy on an unprecedented scale"



By JAMES VINCENT
Nov 8, 2022, 5:09 PM GMT+1 | □

The key question in the lawsuit is whether open-source code can be reproduced by AI without attached licenses. Credit: Getty Images

https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data https://www.theverge.com/2023/1/28/23575919/microsoft-openai-github-dismiss-copilot-ai-copyright-lawsuit



Getty Images sues Al art generator Stable Diffusion in the US for copyright infringement





/ Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its Al model 'without permission ... or compensation.'

Feb 6, 2023, 5:56 PM GMT+1 | 16 Comments / 16 New

By JAMES VINCENT

An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images





Copyright Lawsuits v. Al Companies

ND California



Kadrey v. Meta; Chabon v. Meta Huckabee v. Meta Judge Chhabria



In re OpenAl ChatGPT Litigation (Tremblay, Silverman, Chabon) Judge Araceli Martínez-Olguín



Nazemian v. NVIDIA Corp.

Dubus v. NVIDIA Corp.

Judge Tigar



O'Nan v. Databricks, Mosaic ML Makkai v. Databricks, Mosaic ML Judge Breyer



Bartz v. Anthropic
Judge Alsup



Concord Music. v. Anthropic
Judge Lee



<u>Doe 1 v. Github, Microsoft, OpenAl</u> Judge Tigar



Sarah Andersen v. Stability Al,
Deviant Art, Midjourney, Runway Al
Judge Orrick



Zhang v. Google Judge Lee

D. Mass.



<u>UMG Recordings v. Suno, Inc.</u> Chief Judge Saylor IV



Center for Inv. Rep. v OpenAl, Microsoft Judge Stein

Authors Guild v. OpenAl,

Also: Alter, Basbanes

Huckabee v. Bloomberg

New York Times v. OpenAl Daily News v. Microsoft



SDNY

Microsoft

Judge Stein

Judge Stein

Judge Garnett



UMG Recordings v. Uncharted
Labs d/b/a Udio
Judge Hellerstein

Delaware



Thomson Reuters v. ROSS Intell.
Judge Bibas



Getty Images v. Stability Al Judge Hall

Creative Commons Attribution-Noncommercial-NoDerivslicense (CC BY-NC-ND)

J.L. v. Alphabet, Google

Judge Lee

prepared by ChatGPTiseatingtheworld.com

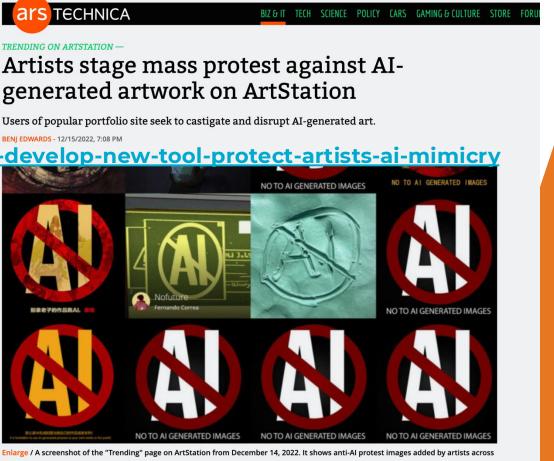
Reaction of the artistic community



- haveibeentrained.com
- spawning.ai
 - Al opt-out

https://news.uchicago.edu/story/uchicago-scientists-develop-new-tool-protect-artists-ai-mimicry

Cloaking



DDM/:SME



Thank you for your attention!









