OPEN International ACCESS WEEK

Will My Open Access Article Be Used To Train Artificial Intelligence? How Do I Protect Myself?

kiz, Ulm University

Charis Brem

Dr Jonas Mirbeth

Will My Open Access Article Be Used To Train Artificial Intelligence?

The short answer: Yes!

But it doesn't have to be to your disadvantage. Here‘s what you need to know.

Content of today‘s course

- Training Artifical Intelligence (on Open Access content)
- Unterstanding Open Access: The Creative Commons licences
- Open and strong: How publishing under CC BY protects authors‘ rights!

Content of today‘s course

- Training Artifical Intelligence (on Open Access content)
- Unterstanding Open Access: The Creative Commons licences
- Open and strong: How publishing under CC BY protects authors‘ rights!

How AI Is Trained!

How it‘s done:

- Over the years by feeding it large amounts of text, images, and videos.
- Crawling the internet, using datasets like Common Crawl.
- Edited text, such as books and journalistic articles, is favoured for AI training.

Rising challenges:

- Growing need for new data to continue training AI models.
- Access restrictions by website operators.
- Legal challenges (EU AI Act, The New York Times has sued OpenAI and its partner, Microsoft, for copyright infringement in December 2023).

Training AI - ‚Scale is all you need‘ (1)

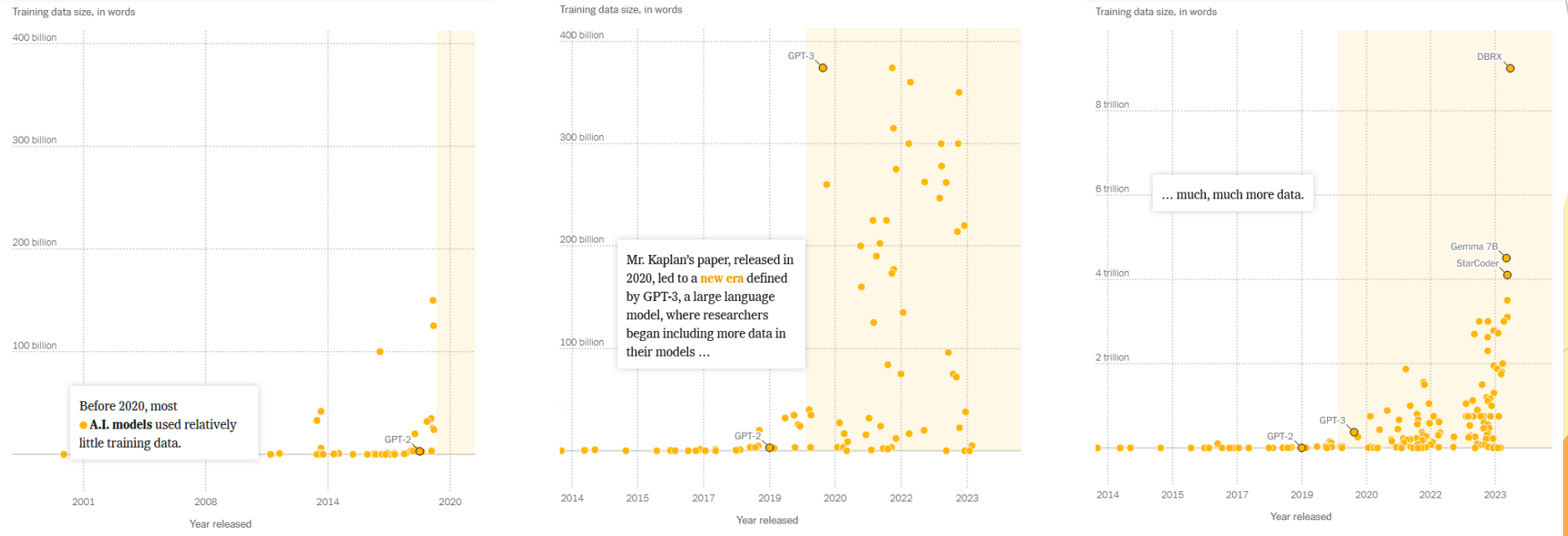
Jared Kaplan et al. (2020) were able to prove the connection between size and performance in generative data models. This correlation is the subject of relevant research by the New York Times in 2024:

> In January 2020, Jared Kaplan, a theoretical physicist at Johns Hopkins University, published a groundbreaking paper on A.I. that stoked the appetite for online data.
>
> His conclusion was unequivocal: The more data there was to train a large language model — the technology that drives online chatbots — the better it would perform. Just as a student learns more by reading more books, large language models can better pinpoint patterns in text and be more accurate with more information.
>
> “Everyone was very surprised that these trends — these scaling laws as we call them — were basically as precise as what you see in astronomy or physics,” said Dr. Kaplan, who published the paper with nine OpenAI researchers. (He now works at the A.I. start-up Anthropic.)
>
> “Scale is all you need” soon became a rallying cry for A.I.

Metz, Cade et al., How Tech Giants Cut Corners to Harvest Data for A.I., NYTimes, updated April 8, 2024, https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html. Kaplan, Jared et al., Scaling Laws for Neural Language Models, arXiv, https://doi.org/10.48550/arXiv.2001.08361.

Training AI - ‚Scale is all you need‘ (2)

The New York Times charts illustrate the exponential increase in training data sets from before 2020 up to and including 2023. There have long been bottlenecks in human-generated data, so AI companies are increasingly dependent on the use of 'synthetic', i.e. machine-generated data sets.



Metz, Cade et al., How Tech Giants Cut Corners to Harvest Data for A.I., NYTimes, updated April 8, 2024, https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

Berlin Declaration on Open Access

OPEN ACCESS
Initiativen der Max-Planck-Gesellschaft

ENGLISH Suche

BERLINER ERKLÄRUNG | BERLIN KONFERENZEN | POSITIONEN | AKTIVITÄTEN | NOTIZEN

Berliner Erklärung

Hinweise für die Unterzeichnung

Regierungen, Universitäten, Forschungseinrichtungen, Förderorganisationen, Stiftungen, Bibliotheken, Museen, Archive, Fachgesellschaften und Berufsverbände, die die in der Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen zum Ausdruck gebrachte Vision teilen, sind eingeladen, sich den Unterzeichnern anzuschließen, die die Erklärung bereits unterzeichnet haben.

Kontakt
Prof. Dr. Martin Stratmann
Präsident der Max-Planck-Gesellschaft
Hofgartenstraße 8
D-80539 München
Deutschland

Email: President oder Open Access Contact

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities

The Internet has fundamentally changed the practical and economic realities of distributing scientific knowledge and cultural heritage. For the first time ever, the Internet now offers the chance to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access.

Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen

Die Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen vom 22. Oktober 2003 wurde in englischer Sprache verfasst. Sie ist einer der Meilensteine der Open Access-Bewegung und liegt inzwischen in zahlreichen Übersetzungen vor. Der Wortlaut der englischen Version ist maßgebend:

https://openaccess.mpg.de/Berliner-Erklaerung

Definition of an Open Access Contribution

Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

1. Open access contributions must satisfy two conditions:The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship (community standards, will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use.
2. A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, inter operability, and long-term archiving.

https://openaccess.mpg.de/Berliner-Erklaerung

Content of today‘s course

- Training Artifical Intelligence (on Open Access content)
- Unterstanding Open Access: The Creative Commons licences
- Open and strong: How publishing under CC BY protects authors‘ rights!

What is CC (= Creative Commons)?

- Nonprofit organization
- Standardised open access licence agreements shall help authors to publish their content under legal protection
- Works are protected by copyright (no public domain!)
- However, subsequent users are granted more rights than are already provided for by copyright law
- Authors grant licences independently, CC does not act as a contractual partner
- licences are modular, authors decide according to their needs
- licences are irrevocable
- Individual agreements may be made as long as they do not restrict but allow more than the licence

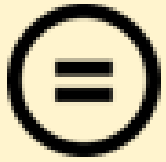
Advantages for the licensor

- CC-licenced content is protected content
- Standard contracts are easy to understand; no in-depth legal knowledge or advice is required
- Before the CC licences: either not protected or "all rights reserved", with CC: "some rights reserved“
- CC licences can be used worldwide and are valid for the duration of copyright protection
- Works under CC licences spread quickly as access is free and re-use is easier than under “traditional” licences
- As the licence conditions are very transparent, more content should be re-used and less content used illegally

Advantages for licensees

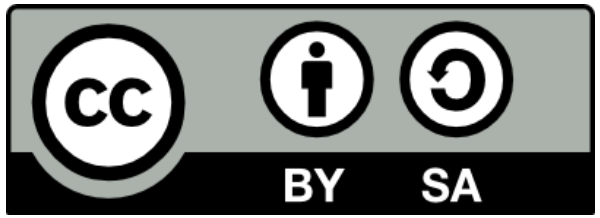
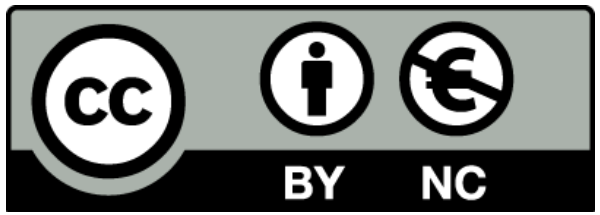
- Quick and easy access
- Subsequent use is clearly regulated
- Standard contracts show clearly what is permitted
- No need to obtain rights for re-use, only compliance with licence conditions (-> copyright notice) is obligatory
- Subsequent use options often go beyond what is permitted by copyright law, e.g. in the case of adaptations

Creative Commons licence types

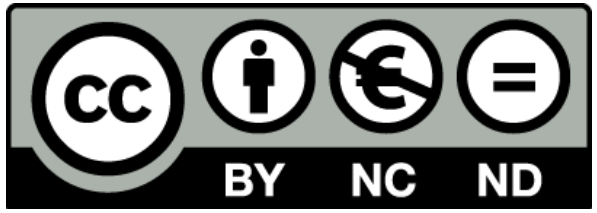
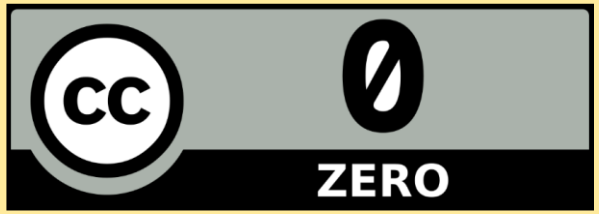
	BY	Attribution
	NC	Non-Commercial
	ND	No Derivatives
	SA	Share Alike

Quelle: https://www.creativecommons.org

The CC licences

	BY: credit must be given to the creator
	BY: credit must be given to the creator SA: adaptations must be shared under the same terms
	BY: credit must be given to the creator ND: no derivatives or adaptations of the work are permitted
	BY: credit must be given to the creator NC: only noncommercial uses of the work are permitted

The CC licences

	BY: credit must be given to the creator NC: only noncommercial uses of the work are permitted SA: adaptations must be shared under the same terms
	BY: credit must be given to the creator NC: only noncommercial uses of the work are permitted ND: no derivatives or adaptations of the work are permitted
	no copyright protection, public domain

Quelle: nach https://de.creativecommons.net/was-ist-cc

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution - You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions - You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation .

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

CC BY 4.0

ATTRIBUTION 4.0 INTERNATIONAL

Deed

Quelle: https://creativecommons.org/licenses/by/4.0

Content of today‘s course

- Training Artifical Intelligence (on Open Access content)
- Unterstanding Open Access: The Creative Commons licences
- Open and strong: How publishing under CC BY protects authors‘ rights!

Do NC/ND licences protect against AI training?

- No. There‘s a growing need for new data to continue training AI models and business practices that do so, many of which are facing legal changes.
- In May 2024, Informa (the parent company of Taylor & Francis) signed a $10 million data-access agreement with Microsoft, providing access to train AI.

How to use Open Access to counteract the formation of monopolies in AI:

- As loss of control over published content happens …
- Prevent the concentration of exclusive access to scientific information without extra costly licences required by publishers.
- Further material: MPDL, DEAL Praxis_Webinar-CCLizenzen, June 2024, Speaker: Dr. Till Kreutzer, in German.

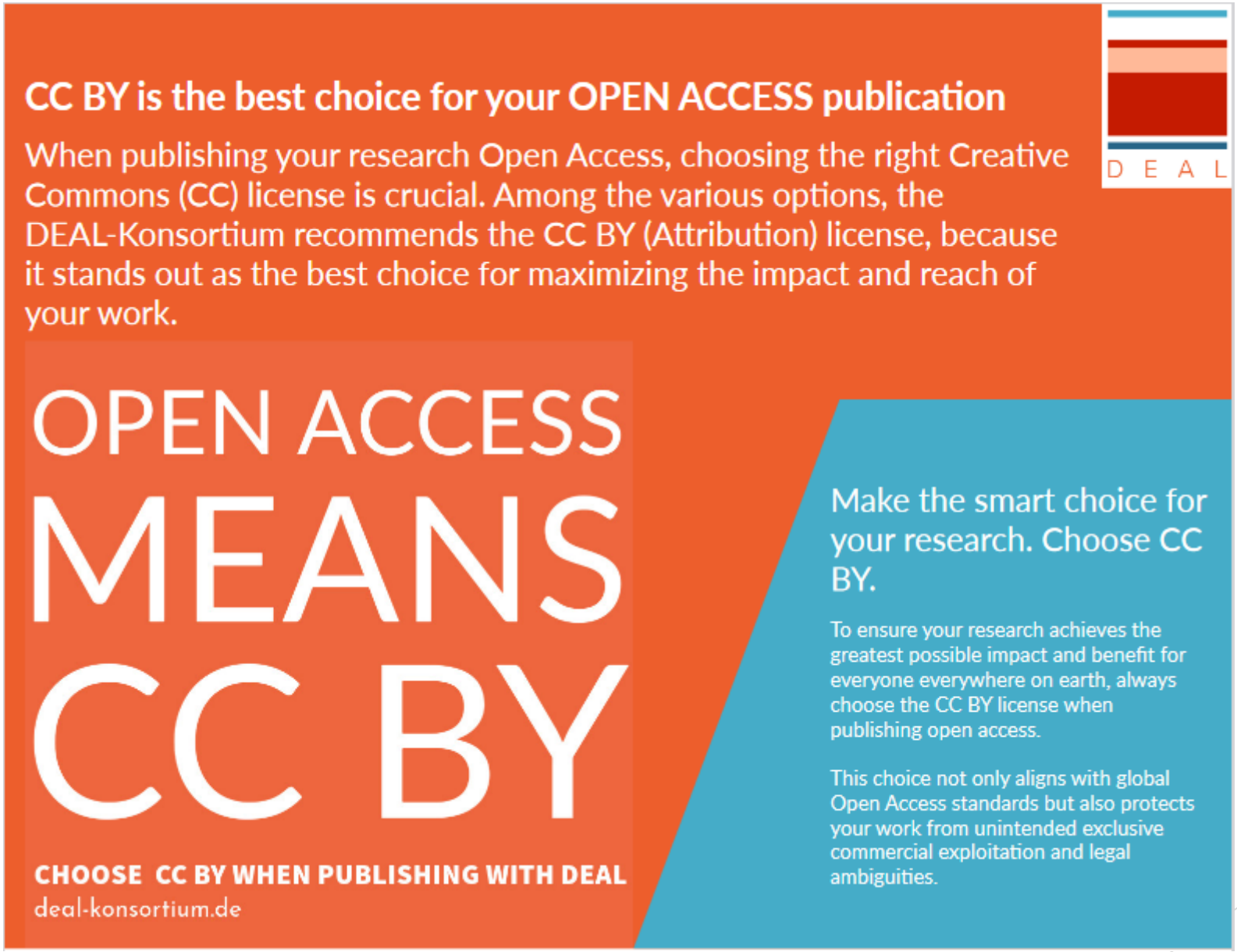
Publishing Open Access

Standard und Best Practice

- In the spirit of Ulm University‘s Open-Access-Resolution

In the context of DEAL contracts (Elsevier, Springer Nature, Wiley), we recommend the CC BY 4.0 license:

- Authors' interests remain protected
- Loss or transfer of exclusive rights of use to the publisher by selecting NC and ND license components (Elsevier/Wiley)



https://deal-konsortium.de/en/why-ccby

The Problem with "Non-Commercial" Licenses

Licenses featuring the "non-commercial" (NC) addition, such as CC BY-NC or CC BY-NC-ND, may seem appealing at first glance. However, they come with significant drawbacks:

- **Exclusive commercial rights to publishers:** While choosing a "non-commercial" license type excludes commercial uses, publishers usually require you to assign those reserved commercial rights to them. Unfortunately, many publishers typically claim these rights exclusively, limiting your control over your own work.
- **Commercial exploitation by publishers:** Once publishers hold (exclusive) commercial rights, they can commercialize your research, including licensing it to AI companies or other commercial entities (including for commercial use), without your consent and without any revenue sharing.
- **Legal uncertainty:** The definition of "non-commercial" is ambiguous under German law. This leads to considerable legal uncertainty as to whether the respective use is permitted. Very often, uses are excluded that the author does not actually want to prevent. For example, it is unclear whether and in which cases NC material can be used in collaborative projects between public and private research institutions. Use by freelance professionals such as doctors, lawyers, architects or even independent research by individuals is clearly not permitted if it serves commercial purposes. In light of these considerations, it becomes evident that NC licenses impede a multitude of desirable uses, thereby contradicting the fundamental tenet of open access.
- **Not compatible with the Open Access definitions:** NC licensed material is not “Open Access” per definition. The “Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities” requires Open Access works to be licensed “for any responsible purpose”. Commercial use of research is, obviously, a reasonable purpose in this regard.

https://deal-konsortium.de/en/why-ccby

The Advantages of CC BY

The CC BY license offers numerous benefits that make it the preferred choice for open access publishing:

- **Maximized reuse and dissemination:** CC BY allows others to distribute, remix, adapt, and build upon your work, even commercially, as long as they credit you. This maximizes the reach and impact of your research.
- **Equal commercial use for all:** Yes, CC BY does allow for commercial use, but it does so equally for everyone. And given that non-commercial is often interpreted extremely narrow, to the extent that posting a NC-licensed article on a website with advertising can be considered a breach of the NC license terms, it is important to allow it. While this might initially seem daunting, it actually serves as the best protection against exploitation by individual players. When everyone has the same rights to use your work commercially, it prevents any single entity from monopolizing or unfairly profiting from it—addressing current concerns, such as those related to AI.
- **Alignment with key Open Access statements:** CC BY is aligned with major Open Access declarations, such as the Berlin Declaration on Open Access. It is also the preferred license of many research funders and organizations worldwide.
- **Legal clarity:** CC BY provides clear and straightforward terms, reducing legal uncertainties and ensuring your work can be freely used and shared across various platforms and by diverse audiences.

https://deal-konsortium.de/en/why-ccby

Takeaways

1. As loss of control over published content happens …
2. Check if you transferred the rights to the publishing house or if they are still yours.
3. As rights holder, you do not have to adhere to the licence conditions.
4. Prevent the concentration of exclusive access to scientific information by choosing CC BY 4.0.

Your Questions - Q&A

Funding requests:

- oa@uni-ulm.de
- +49 (0)731 50 - 154 76

Licences & Green Open Access

- kiz.publikationsmanagement@uni-ulm.de
- +49 (0)731 50 - 314 28

All topics presented and further information can be found under the following links:

https://www.uni-ulm.de/open-access

https://www.uni-ulm.de/oa-foerderung

kiz Kommunikations- und Informationszentrum

universität ulm