Clarifying Copyright to enable AI Research in Africa

RIA, AI and Intellectual Property Brief 1

Executive summary

AI is an important general-purpose technology that can be used to analyse information in ways that has not been possible until now, and to automate information-intensive tasks. AI is used to refer to a wide range of related but different techniques for developing software. One of these techniques requires a large collection of input data to develop the software. Text, photos and software have been used to develop notable examples. Since most text, photos and software are under copyright, does the use of them to develop AI implicate copyright? Copyright law in Africa is unclear whether use to develop AI is infringing but could be interpreted to restrict use of inputs for training AI. Because Africa has little AI development capacity, the restriction could prove fatal to development of African AI capacity. Global AI developments have been shown to ignore African issues. African countries should individually and collectively create legal provisions that clarify when copyright permits use for training AI, fairly balancing competing interests while enabling AI development.

Introduction

Artificial Intelligence (AI) is the most recent general-purpose technology. It has been employed in many sectors due to its ability to accomplish tasks and generate results far more quickly than a human can. Creating an AI model involves input data, processing and output. Copyright questions arise when the pre-existing data used to train AI models are in a form that haven't been considered data until now

such as photos and texts (Craig, 2022). Does use of a copyright-protected text or an image in training an AI model infringe copyright in that text or image? If use of vast amounts of human creativity to train AI models is permitted, then should the results have the benefit of copyright? This brief answers the first question. The second question is discussed in a companion brief: Generative AI and sustainable creativity that explains why African countries shouldn't extend copyright to AI outputs.

Copyright restricts making copies and in many countries digital transmission, with a few exceptions. But copyright law was made to restrict use by humans. A person's enjoyment of a photograph is radically different to use as one of ten thousand images to produce software. It is not clear in many cases whether the processing of texts or images to train AI infringes these rights, although it involves copying. This is complicated by the different exceptions in each different jurisdiction to use copyright works for research. It is unclear if these apply to the use of copyright texts and images for the purpose of AI training. The lack of clarity surrounding the use of copyright-protected works in AI training has not deterred the transnational technology platforms that have most of the expertise in training AI. However the lack of legal clarity is likely to hinder the use of copyright texts and images by African research institutions and entrepreneurs which have fewer resources and are more risk averse. To enable African research, African countries should amend copyright laws to include a provision that permits the use of copyright works for training AI; however the provision should balance the different interests.

The technology

AI is a mutable term. A useful description from The Alan Turing Institute defines AI as "the design and study of machines that can perform tasks that would previously have required human (or biological) brainpower to accomplish"(*Data Science and AI Glossary*, 2023). But AI is more than computational techniques, it's a new dynamic affecting the environments, economies, cultures and politics. State of the art techniques were developed using energy intensive computation, rare minerals and huge troves of input data (Crawford, 2022). Africa appears only briefly in the process, at the beginning as supplier of rare minerals, and sometimes near the end as with ChatGPT for which the data was refined by badly compensated data labourers in Kenya (Perrigo, 2023).

Cutting-edge AI involves the development of algorithms via repetitive procedures. For inputs, training data is collected. This data is then preprocessed to remove any noise or irrelevant information. For some techniques the data may be labelled. The result is the training dataset. A layered neural net is used to filter the training dataset. For example, the training dataset could be photographs of trucks.

The neural net develops connections between the nodes, the nodes enable processing of further data. The process is iterated until the neural network can perform a task with a measure of accuracy determined by the developers. For example, the AI model might distinguish between an image of a truck and an image of another kind of vehicle in 60% of tested iterations. Similar processes could be used to generate photo-realistic images of trucks or with greater refinement to classify different types of trucks. The training dataset, to be large enough to train the model, would include a wide variety of photographs of trucks, perhaps a few photographs by professional photographers who sell images of trucks, or take them on commission; but the vast majority would be taken for other reasons, to sell the truck in question, or in the background of photographs of people. Many would be amongst the vast majority of photographs that are taken and shared online for non-commercial purposes. Use in an AI training dataset would be unrelated to, and would not interfere with the reason the photographs were taken.

Use of copyright material to train AI

The Berne Convention, binding on most African countries, requires national law to give copyright holders the exclusive rights to reproduction and adaptation of literary and artistic productions, while later treaties give rights holders the rights to control distribution of the work. Although these laws were created to reward the creators of novels, paintings and operas they apply automatically to the far more banal, and multitudinous emails, selfie photographs and videos created everyday by ordinary people and for which no market exists. Copyright was developed in response to new printing technologies which enabled the mass production of copies and is consequently not fitted for digital technology which involves the making of multiple copies, many of them ephemeral simply to communicate. As a consequence purely technical copying is not always infringing.

Most training processes appear to involve ‘scraping’; that is copying images or text already available on the Internet. The images in a training dataset might never be seen by the human developers, instead they are changed into formats suitable for training the neural net. They might be seen by data labourers who label them or remove offensive images. While technically copying, it is not clear whether the acts involved are infringing in copyright law, especially since much of the copying is temporary. The legal analysis is further complicated because the use may be permitted by an exception for research. However, given the massive number of images collected to train AI models it is likely that at least some collection for training AI is potentially infringing in many countries. But even then it is not clear whether rights holders would be eligible to collect compensation from the infringer since the use of a single item in a training set of thousands doesn’t have a clear economic value. It is

only in the aggregate that value emerges. Use in training does not undercut sale or royalties for other uses since it doesn't substitute for them. What is clear is that for large datasets obtaining permission to use the inputs is practically impossible. It is impossible to even ascertain who all the copyright holders are given the paucity of available metadata, and even less possible to obtain contact details for rights holders and to get their permission. It is not known how much potentially infringing copying is taking place since developers of many large AI models have not shared their datasets (Crawford, 2022), possibly since this would expose them to copyright claims, however speculative.

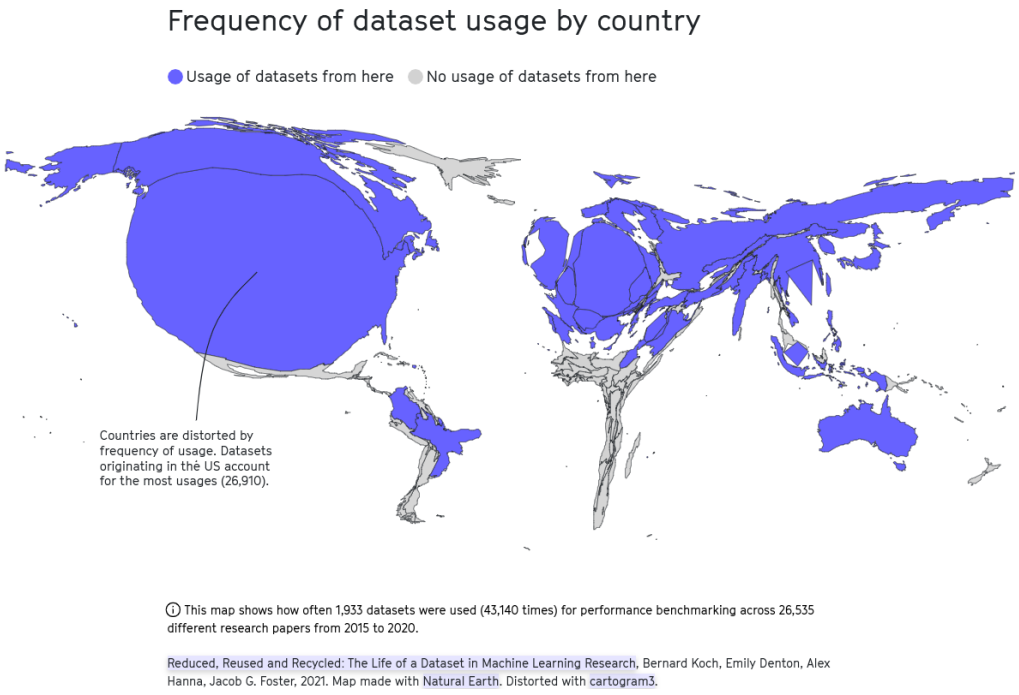
Copyright laws in Africa, even those that have been recently amended are not clear on when use of a copyright image or text to train an AI is an infringement of copyright or not. Choosing to require AI researchers to get permission for use for training is the same as prohibiting AI research since it is simply not possible at scale. At the same time use for training does not usually interfere with the commercial exploitation of the image or text, and that is for those few texts and images for which commercial exploitation is possible. While some AI models generate outputs that resemble the creativity on which they are trained, many other models do not, instead proving useful in other ways. Therefore it is important that contemporary copyright laws permit at least some uses of copyright-protected works for AI training.

Why is it important to clarify copyright to enable AI research?

There are already indications that African influence in generative AI is minimal. Most of the training data used to create generative AI products are based on European and English languages (Adams et al., 2020). Few AI models are trained on research that include Africa. A 2021 research review of the datasets used to benchmark AI models found that of 1 933 datasets only 12 were from Africa, and all of those from Egypt (Koch et al., 2021). Benchmark datasets are used to test and compare models. If African data does feature in training data but not benchmark data, results that represent Africa will be regarded as less accurate. As a result there are concerns that outputs from generative AI will be skewed towards the Global North, with little to no development on generative AI in the Global South (Komminoth, 2023). Additionally, products that have apparently been trained with multi-lingual datasets still present inaccurate results when used in relation to Africa (Abott et al., 2023).

Although generative AI products are not inherently designed for specific regions, the bias in the dataset tends to result in products tailored for the Global North rather than the Global South (Koch et al., 2021). Therefore, further excluding African inputs into AI through copyright will result in a failure to

ensure African AI growth and influence as well as perpetuate biases, and risk cultural colonialism (Yoon, 2023).



AI is a strategic technology. If African countries do not develop their own AI, they will be consumers of AI developed in other contexts that serve the agendas of those who have developed AI. Requiring African researchers to seek permission to use copyright works will leave African countries reliant on countries that don't require permission. The result would be that only a handful of large tech companies will create and control AI. Multinational tech platforms don't need copyright law to give them permission to train AI because they already have rights to use vast numbers of photos, videos and software under the contracts that users agree to when they use the platforms. African AI researchers don't have those permissions. If AI is not clearly enabled by African countries there will be little or no African AI. African policymakers should create a clear exception for use of copyright works to train AI.

How can law enable AI research

The European Union and the UK have narrow exceptions to copyright for computational analysis such as text and data mining that arguably extend to training AI, but permit only non-commercial use. The European Union exception, which permits only non-commercial use, has been found to be too restrictive (Geiger, 2021). In the United Kingdom the existing exception has proven so unsatisfactory that the Intellectual Property Office recommended allowing use for any purpose (*Artificial Intelligence and Intellectual Property*, 2022). An expert report on an "innovation friendly" regime stresses that "restriction of data access for training sets would be likely to put the UK at a disadvantage and impede domestic development of the technology" (Vallance, 2023) but the law has not been changed, instead the government is hoping a licensing solution will emerge (HM Government, 2023), even though those hopes have proven fruitless since the UK amended its copyright law in 2014 .

In stark contrast to those attempts, the country that produced the most AI breakthroughs, the United States, has not needed to make changes to its copyright laws to enable AI research. An existing provision of copyright law entitled "fair use'' permits some use of copyright text and images to train AI models (Kumar & Kumar, 2023). Other countries with similar provisions also flexibly permit use to train AI but not all uses. The use for training must pass a balancing test that takes into account relevant factors including "the effect of the use upon the potential market for, or value of, the copyrighted work". Not all uses for training would be permitted. Use of an artist's work to imitate their style would likely not be authorised (Office of Legal Counsel and Legislative Affairs of the Israel Ministry of Justice, 2022). If a model produces outputs that affect the market for the originals then use for training may not be justified by fair use (Henderson et al., 2023). These flexible exceptions that protect existing markets for copyright images, texts and software, where these exist, avoid the trap of effectively prohibiting use for AI research that could be used commercially even when it does not affect the business models of rights holders. Only a flexible provision can be adapted quickly to rapidly changing technologies.

While flexible, open-ended copyright exceptions are the most attractive solutions for countries there is also a need for cross-border coordination. Training sets are gathered from across the Internet although Africa seems to be underrepresented. AI powered services are accessible across borders. Experts recommend that an international institution such as the World Intellectual Property Organisation should negotiate binding copyright guidelines and norms of when use of copyright images, texts etc. may be used to train AI which have universal application (Flynn et al., 2020). But African policymakers should not wait to have rules set for them, nor pause continental integration to wait on the processes

of international organisations. Harmonisation across Africa of copyright guidelines can enable African AI development, while protecting African creative industries. But while it is necessary for the development of the technology to give clear legal permission to developers of AI it is not necessary to give copyright over the products of those AIs, as discussed in our companion brief.

Policy recommendations

- National copyright laws should be amended to include a flexible provision that enables use of copyright works in AI research that includes a balancing test to reduce harms to authors.
- Counties should include a requirement for a minimum balanced exception for training AI in the Intellectual Property Protocol of the African Continental Free Trade Agreement. The provision should treat flexible, open-ended copyright provisions as meeting the requirement.
- National copyright laws and continental treaties should clarify that copyright does not subsist in data.

Authors

This brief was prepared by Dr. Andrew Rens, Hanani Hlomani and Samantha Msipa for Research ICT Africa. Correspondence may be addressed to arens@researchictafrica.net.

Research ICT Africa
Workshop 17
V&A Waterfront
Cape Town, South Africa
https://researchictafrica.net/

References

Abott, J., Dossou, B., & Rooweither, M. (2023, February 9). Comparing Africa-centric Models to OpenAI's GPT3.5. *Lelapa*. https://lelapa.ai/comparing-africa-centric-models-to-openais-gpt3-5-2/

Adams, R., Fourie, W., Marivate, v., & Plantinga, P. (2020). *Introducing the series: Can AI and data support a more inclusive and equitable South Africa?* (11333/DOI 10/15280; Policy Action Network (PAN) Topical Guides: AI & Data Series 1). http://repository.hsrc.ac.za/handle/20.500.11910/15280

Artificial Intelligence and Intellectual Property: Copyright and patents: Government response to consultation. (2022). UK Intellectual Property Office. https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation

Craig, C. J. (2022). The AI-copyright challenge: Tech-neutrality, authorship, and the public interest. In R. Abbott (Ed.), *Research Handbook on Intellectual Property and Artificial Intelligence.* https://papers.ssrn.com/abstract=4014811

Crawford, K. (2022). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Data science and AI glossary. (2023). Alan Turing Institute. https://www.turing.ac.uk/news/data-science-and-ai-glossary

Flynn, S., Geiger, C., Quintais, J., Margoni, T., Sag, M., Guibault, L., & Carroll, M. W. (2020). Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3578819

Geiger, C. (2021). *The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive* (No. 66; PIJIP/TLS Research Paper Series 2021). Program on Information Justice and Intellectual Property. https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1069&context=research

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). *Foundation Models and Fair Use* (arXiv:2303.15715). arXiv. https://doi.org/10.48550/arXiv.2303.15715

HM Government. (2023). *HM Government Response to Sir Patrick Vallance's Pro-Innovation Regulation of Technologies Review*. HM Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1142798/HMG_response_to_SPV_Digital_Tech_final.pdf

Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *ArXiv:2112.01716*. https://arxiv.org/abs/2112.01716v1

Komminoth, L. (2023, January 27). Chat GPT and the future of African AI. *African Business*. https://african.business/2023/01/technology-information/chat-gtp-and-the-future-of-african-ai

Kumar, R., & Kumar, P. P. (2023). Training AI and Copyright Infringement: Where does the Law Stand? *Indian Journal of Integrated Research in Law*, *II*(I), 1303–1316.

Office of Legal Counsel and Legislative Affairs of the Israel Ministry of Justice. (2022). *Opinion on Machine Learning and Copyright*. Office of Legal Counsel and Legislative Affairs of

the Israel Ministry of Justice. https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/machine-learning.pdf

Perrigo, B. (2023, January 18). The $2 Per Hour Workers Who Made ChatGPT Safer. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Vallance, P. (2023). *Pro-innovation Regulation of Technologies Review*. Government Chief Scientific Adviser. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1142883/Pro-innovation_Regulation_of_Technologies_Review_-_Digital_Technologies_report.pdf

Yoon, S. (2023, April 25). Artificial generative intelligence risks a return to cultural colonialism. *VentureBeat*. https://venturebeat.com/games/artificial-generative-intelligence-risks-a-return-to-cultural-colonialism/