Q&A

UKSG Webinar: The Open Access – AI Conundrum: Does Free to Read Mean Free to Train?

Stephanie Decker | Caroline Ball

Q: Is there some way (e.g. some legal mechanism) that can mandate and enforce the propagation of attribution alongside use of content by AI?

SD: I am not a legal expert but I don't think so. The other problem is that law is made nationally, but AI is internationally available.

CB: Unfortunately not, other than the current litigation approaches regarding copyright infringement – since by not complying with the terms of the licence, that's basically what it is. Without enforced transparency around training data and the processes of these LLMs, enforcing the terms of CC licences is practically impossible.

Q: How AI will change OA? Will we see different OA models due to AI?

SD: That appears to be happening with the Creative Commons licenses, though I am unclear what this will mean. The OA organizations have not really addressed it, and in the UK, the mandated use of OA continues unchanged despite the AI harvesting of data.

Q: Is "AI" being used here to refer specifically to AI models that power chatbots, or to AI models more generally?

SD: AI here refers to LLMs – not all uses for LLMs are chatbots.

CB: We do need to be more careful about terminology, because 'AI' does cover many things, not all of them problematic or contentious. In this context, we are mostly talking about the broader class of AI that covers the large language models, the 'conversational' interfaces. That could well cover chatbots, it largely depends on the model behind the chatbot, and what data it's trained on. Some may be small, localized chatbots trained on in-house data, and that's fine.

Q: Why might institutions choose not to use their agency? What profit/political motives are at play?

SD: I am unclear on this, as the profits accrue primarily to the AI companies, and I assume you mean the HE sector and academic publishers. Why there is not more pressure on publishers to license their content to AI firms to share the profits with Universities, authors, and journal owners is a mystery to me. No other business would leave money on the table like that. More concerning was the Copyright Law consultation, which effectively proposed writing a blank cheque for foreign AI companies… Peter Kyle's motives (or that of his office) are unclear to me.

CB: There can be financial pressures, incentives around prestige and being seen as 'innovators' or the flip side, the fear of being left behind, or aversion to risk and its seen as safer to follow the crowd. There can be a lack of understanding of all the issues by the people making decisions, or too many people involved for clear lines of decision-making, or siloed decision-making. Plus, in my experience, it's a good idea never to underestimate institutional inertia!

Q: Does this point to CC-BY-NC-ND as a default for OA licensing?

SD: I don't think it is currently the default, and it does not stop AI ingestion.

CB: I hope not, because that's risking collateral damage, by restricting legitimate scholarly use, remixing, translation, adaptation, and those are activities that are particularly important outside our own, dominant, Anglophone contexts. And NC particularly is often quite ambiguous (universities charge fees and make profit/surplus, so are they 'commercial'? Much debate on that one!)

Q: Does CC-BY-NC (non-commercial) theoretically block use by for-profit AI models?

SD: Apparently not.

CB: In theory, CC-BY-NC could block use or allow for legal redress, either because the AI models' use of it is deemed to violate the NC element, or because they fail to uphold the BY element. That said, it's impossible to say right now because a lack of any precedents in case law. It will likely come down to whether courts decide training models in this way could be covered by a copyright exception. If they decide it's a legitimate activity covered by a copyright exception (fair dealing, for example, or data mining) then the licence may not apply at all! But there's no clear legal decisions on that yet, and even then it will vary depending on individual country's copyright law. And if such a decision upheld this kind of use, then it's a question of what happens next? Reform to copyright law to reflect the fact that the existing legislation just didn't anticipate this kind of use, and taking steps to find some way to address it? And what appetite might there be for that, and what kind of lobbying activity will kick into action? Watch this space!

Q: Are there any technical solutions available to block the harvesting of data with code?

SD: Robot.txts are used currently (but effectiveness does not appear to be great: https://www.theatlantic.com/technology/archive/2025/06/generative-ai-pirated-articles-books/683009/), and I understand that Cloudfare is working on something. I am also increasingly being accused of being a robot when I try to search on Google Scholar. Presumably this is the kind of disruption that Nightshade provides for image scraping, but I am not very familiar with this area.

CB: Nothing robust enough to manage something of this scale. Paywalls can deter harvesting/scraping, but there's always ways around them. There have been some attempts at 'data poisoning' (like Nightshade) or inserting watermarks or flags into data that can be tracked/identified, but again the scale is the issue. And any attempts to block AI bots etc will probably have knock-on effects for general users/readers, making the web less accessible and user-friendly for everyone. The only real solution is legal clarity first, and then legislation to enforce governance, accountability, reciprocation etc, and penalties for lack of compliance.

Q: For Caroline, are there any initial actions you would like to see from Institutions?

CB: I'd like to see institutions developing very clear position statements on AI training and licensing, both external-facing and for internal use, rather than leaving decisions to individual authors and researchers – and alongside that, really investing in AI literacy for staff (all staff, not just academics!) and students, making sure they understand how these LLMs work, what they are and what they are not, what the issues and flaws are, what the ethical considerations are. I'd like to see them demanding transparency and accountability from suppliers (and this applies to all uni/library suppliers, not just AI – what data are they collecting/using, how are they using it internally and externally, what mitigation strategies do they have for the harm being caused, what reciprocation are they providing to weigh against their extraction). I'd like to see collective bargaining and collaboration in the same way we have for R&P deals etc (though there are issues with any collective bargaining approaches, and the smaller institutions often pay the price here). I'd like to see VCs using their weight and influence to advocate for policy reform to protect research and the commons, and to impel these companies to start giving back and investing in the future of the research they're profiting so much from.

Q: If a paper is published under a CC-BY-NC licence couldn't the authors seek legal redress if it's used to generate profit for an AI company?

SD: My understanding is that even where it is copyrighted material, AI ingestion constitutes fair use according to current interpretations (I find this mindboggling). Successful cases against AI companies are based on the fact that pirated copies were used, not that these books had copyright.

Q: Is it the case that the articles behind paywalls are not also feeding LLMs? Is it only an OA issue?

SD: Now they do where publishers have struck deals. Otherwise, it may have been Author Accepted Manuscripts, but it is hard to tell, as the origin of the material (the provenance) cannot be established. Much like with books, we know that a pirated database hosted in Russia has been used by some companies to train their Ais – the Atlantic reported on this: https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ . All of my pre-OA articles are certainly in one of those pirated databases, and both of my books are, too.

CB: It's not only an OA issue, we're seeing publishers making deals with AI companies for licensing of their backlists and holdings, sometimes with mechanisms for opt-out and/or compensation for authors, but not always! Unfortunately if authors have signed away their copyright, there's not much they can do about it – another reason why not signing away copyright is important, and for institutions to have Rights Retention policies. And of course, we've heard of AI companies also harvesting data from illegal sources like Z-Library, Sci Hub etc.

Q: How can we balance the messaging around AI anxieties vs promoting Open Research?

SD: Good question, I have no idea.

CB: We need to be open and flexible, it's a complex issue, and it isn't as a simple as black/white, use/don't use. You can be against the extractive, expoitative AI business models, whilst also recognizing the technological possibilities, and also champion the ethical case for open scholarship. For me, a really key element is transparency and accountability, and that is very much lacking at the moment with these big AI companies. The problem is not that knowledge is open, and we need to react by trying to lock things back down. The problem is that the use of it by these companies is concentrated and unreciprocated.

Q: Could a link to Stephanie's paper on Citation Laundering be shared please?

SD: Thank you for your interest - it's not a paper but a blog post: https://scholarlykitchen.sspnet.org/2025/04/15/guest-post-the-open-access-ai-conundrum-does-free-to-read-mean-free-to-train/

Q: A question for Stephanie: is AI training not a ""problem"" for all academic research, not just open access? As long as the material is obtained legally a couple of cases have judged that AI training is 'fair use' in the USA. My point is this isn't really about OA or CC licences specifically. It's about the copyright framework and the exceptions and limitations of copyright.

SD: That is a good point, though the number of litigations about whether it does indeed constitute "fair use" or not suggests that this is still under a degree of legal consideration. Also, the copyrighted materials were usually obtained illegally, and copyrighted material is rarely available for free online. The law clearly lags behind technological innovation, that is normal, and let's not forget, the law can be changed. AI ingestion as a use case did not exist when copyright laws were designed, nor when CC licenses were considered.

Q: I see a lot of authors inclined to select NC/ND licences, often unwittingly in breach of funder/employer policy. I suspect more researchers than we realise are not wholly sold on "open", and would happily limit readership if permitted to. Could AI trigger a backlash against CC licences in the academy?

SD: Well, I always chose them because I actually feel quite ambiguous about the OA mandates. Of course, I like that more people can read my work, and I always share private copies of any non-OA material when asked. But OA, when it is mandated, this also means I cannot opt out, and if I did, I am at a disadvantage if everyone else's work is freely available. But at a more fundamental level, I do not understand the use case – I suspect only a few academics in the social sciences, and even fewer in the humanities, understand it. Our institutions are paying publishers a lot to provide OA publishing a lot of money, and not necessarily less than they paid for subscriptions previously. All the "free" labour is provided by academics, who are also paid by universities.

In the OA movement, the argument is that taxpayers pay for it – but why does it go to the publishers? Larger because the actual mechanics of OA publishing are poorly designed and enable massive returns to scale. So while the rest of the publishing world is seeing a significant shift towards self- or indie publishing, we are dependent on a small number of academic publishers, even though technically, anyone can publish anything anywhere. As far as design flaws go, this is staggering…

And, as I said in the presentation, in the UK, the taxpayer is not paying for it. I don't just mean student loans, it's also that the entire system is subsidised by business schools recruiting international students (hence my Chinese parents comment). And working at a business school, where the screws get turned and turned year on year, we are really at the thin edge of that wedge. At the same time, we teach about value generation and capture and IP and other types of business models – frankly, from the perspective of a business school prof, OA sounds pretty crazy.

Q: Do we know, or is it even possible to know the extent to which AI is utilising OA academic material without due credit?

SD: On one level, it is impossible to know. But we know the intensity with which bots are grazing repositories, so it is clear that they are. As all my copyrighted material was ingested by several AIs as training data, I can confidently assume all my OA stuff was as well – and everybody else's, too.

CB: I suspect not! And a lot of that is down to the lack of transparency of how these LLMs operate, what data they're being trained, where and how that data is being collected. Given what we hear about bot traffic, about the strain on internet infrastructure, about site bandwidth, I'd say it's safe to assume if it's openly available out there somewhere, they've harvested it, legally or not.

Q: Is the danger that individual (particularly high status/wealthy) academic institutions do licensing deals unilaterally, what collective bargaining and action fora are currently standing up for the agency of the commons?

SD: In the US, perhaps, but really it is the publishers that are licensing right now, and I am surprised that there is not more of a movement to ask for a share of these revenues.

CB: That's always a risk – you hear people talk about the HE sector as though it was uniform and united, and whether we're talking about collective negotiations for R&P deals, or collective action regarding AI, the HE sector is very much not united! There are the haves and the have-nots, and the way negotiations proceed don't always reflect the fact that 'one size fits all' deals don't fit all institutions! But your Oxbridge and Russell Group unis definitely don't face the same challenges, or have the same priorities and imperatives as your post-92 unis, for example, so maybe there needs to be less national collectivization and more localized? I'm not sure I have a solution for that (I'd be much in demand, if I did!)

Q: As a publisher, I'm seeing authors increasingly choose CC-BY-NC/ ND licenses (CC-BY is our default for OA but we allow other CC licenses). Authors are clear they are doing this as a signal about AI usage. Yet if training on content is fair use, the CC license distinction is meaningless. How to advise authors at this time of completely unsettled signals?

SD: As academic authors, we have no choice, and I agree with you, this is a clear signal. To be fair, I did it before AI, as I was always uncomfortable with the standard CC-BY license and tried to restrict as much as I was allowed within the mandate.

Q: What do you think a fair/ethical open access + AI landscape looks like?

SD: That is a nice question – I don't know but I am wondering how we would even get there. It should really start with clear understanding between all types of knowledge producers and AI companies that mass harvesting needs to be accompanied by a giving back – and I don't mean us paying as extra to consume AI models. Probably the best solution would be a large-scale investment in knowledge-creation and exchange infrastructure. But for as long as this remains a volatile and dynamic space (Chinese AI firms now harvest statistical weights from the US companies) I doubt anything will happen. All technology adoption follows S-curves, and until we reach the plateau and the dust settles, it is difficult to know.

CB: There has to be transparency about training data – where the data came from, how was it gathered, was it used with consent etc. There needs to be mechanisms for proper, meaningful attribution/acknowledgement. There needs to be reciprocity built into the business model, in recognition that the data may be 'free' (either because it's theoretically open access, or because they're infringing copyright) but the labour that produced has a cost. There needs to be more safeguards built in, and more transparency, about how equity is considered, how bias is accounted for and addressed. There needs to be mitigation strategies for countering the environmental harms and reducing them. And all of this needs to start happening now, when these companies are just starting out, and we potentially have the power to make them – because very few companies in history have voluntarily chosen to act ethically or surrender power, and the more entrenched they get, the less leverage we have to force them to do so.

Q: In terms of data, what would be best licence? ND ? and what truly would be consent for data to be use for ai?

SD: Great question – consent presupposes understanding, so it would need to start with a clear understanding of how the AI model uses it and it would be divulged in a recognisable form.

Q: How can limits to AI scraping be squared with academic rights to content mining? Murray-Rust, P., Molloy, J. and Cabell, D. (2014) 'Open Content Mining', in S. Moore (ed.) Issues in Open Research Data. London: Ubiquity Press. Available at: https://doi.org/10.5334/ban.b.

SD: There are people more expert than me, but I would think the intent behind usage is commonly a factor: research vs. commercial use. Licenses for databases usually have differential pricing, too.

Q: Martin Eve suggests that, at least in the US, mining/training might be considered fair use and thus legal despite a no-derivatives (ND) license. Is attempting to stop scraping futile? https://blogs.lse.ac.uk/impactofsocialsciences/2025/11/24/creative-commons-licenses-and-copyright-may-not-stop-academic-work-being-used-to-train-ai/

SD: Yes, that is my understanding – I don't think any type of CC license prevents scraping.