VELOCITY OF content
A Copyright Clearance Center Podcast

CCC At Frankfurt Book Fair 2023

with
Dr. Hong Zhou, Wiley
Dr. Namrata Singh, Turacoz Group
Carlo Scollo Lavizzari

Recorded Thursday, October 20, 2023
For podcast release Monday, October 30, 2023

KENNEALLY: Scientists are given to understatement. In their paper describing the double-helix structure of DNA, published in *Nature* in 1953, James Watson and Francis Crick mildly declared, "this structure has novel features which are of considerable biological interest." Sir Alexander Fleming even had less to say about one of the great medical discoveries of the 20th century, penicillin. Fleming, who won the Nobel Prize in 1945 for his work, noted about the nature of research, "one sometimes finds what one is not looking for."

And on November 30th, 2022, in a tweet that afternoon, @SamA wrote, "today we launched ChatGPT. Try talking to it here." In a brief thread, Sam Altman predicted, "language interfaces are going to be a big deal, I think. Talk to the computer, voice or text, and get what you want for increasingly complex definitions of want."

Well, going to be a big deal? Oh, yeah. Definitely. The large language models that are fundamental to generative AI solutions, including ChatGPT from OpenAI, rely on machine-readable content available on the web as books, scholarly journals, or other curated publications. Much of that is protected by copyright. Already, licensing programs are emerging to allow such uses of copyrighted content. In the meantime, publishers, academics, and researchers are recognizing and responding to the opportunities that AI presents to them. These communities all share concerns that machine-enabled solutions must incorporate essential human rights to equity and security, authorship and authenticity.

My name is Chris Kenneally. I'm with Copyright Clearance Center, CCC. I'm very happy to welcome you today to this discussion. And I'm very happy to share this platform with a panel of experts who are going to share their insights on these issues.

I have a wonderful panel, and I want to briefly introduce them now. Immediately to my right is Carlo Scollo Lavizzari. Carlo, welcome. He's an internationally recognized specialist in intellectual property law and policy, with 20 years of experience working in Africa, Europe, and the US. He is the author of a short paper, "A Snapshot of the

Relationship between AI, Copyright, and Licensing," which is available today at this event.

To Carlo's right is Dr. Namrata Singh. Dr. Namrata, welcome. Dr. Namrata is founder and director at Turacoz Group, a medical communications company working with pharmaceutical, biotech, medical device, and diagnostic firms, as well as academic institutions, to support research and publications globally. Dr. Namrata is a pediatrician and a founding member of the AI Working Group at the European Medical Writers Association.

At the far end to my right is Dr. Hong Zhou. Dr. Hong Zhou, welcome. Dr. Hong Zhou is director of intelligent services and head of AI R&D for John Wiley and Sons and leads the intelligent services group in Wiley Partner Solutions. Dr. Zhou holds a PhD in 3D modeling with artificial intelligence algorithms, and he's a chef and contributor to the Scholarly Kitchen blog published by the Society for Scholarly Publishing.

I think, Dr. Zhou, it's important to start with you and to learn more about the publisher's perspective here and how these tools are going to become part of your workflow. Tell us about the role that AI solutions will play in this intended evolution at Wiley from being a content provider to being a knowledge provider.

ZHOU: That's a good question. Basically, Wiley is shifting from a content provider to now a knowledge provider. So we started thinking about and talking about – when we talk about knowledge, what does it mean, knowledge? The people that consume or digest or absorb the knowledge and understand applied knowledge. And the knowledge is distributed to the people. So there's these clear three key elements – people, knowledge, and how you interact, distribute it out, and understand. So we apply AI to support these three key elements.

For example, when we talk about knowledge, first, we need to create the knowledge. How? We apply the AI to automatically extract the key metadata and to present the knowledge for a human-friendly, human-readable, and machine-readable. Also, we extract the hidden knowledge – the valuable concepts, entities, and relationship between the entities from this unstructured text content. This is knowledge creation. Then, in order to better serve the people, we need to understand who they are and what are their interests so that we can distribute relevant articles and knowledge to them – to humans. So we apply the AI to understand, perhaps, our publisher partners to better understand their audience – what they want, what's their interests, what's their expertise, so we can make sure we can make more personalized the information distribution to them.

Lastly, we want to know, how can we help the researchers, users, to discover this knowledge? Because today, there's all the information overloading. It's too much to read. So we need to apply the AI to simplify, to generate some summarization to help. For example, Wiley Partner Solutions has a research exchange summation, so we automatically extract the metadata to make the summation work much easier. Also, we have Literatum, the largest digital publishing platform for scholarly content in the world. Then we can make all this knowledge searchable and readable over the internet. We also validate the knowledge, because this is very important. Integrity as a publisher is a key responsibility to make sure this knowledge is good quality, good standards. So we also apply AI to detect paper mills, image manipulations, etc.

KENNEALLY: It seems to me that in order to understand where publishing will be going using these AI tools, it's important to understand how people are using them. And it sounds to me like it's changing the relationship – two relationships, really – the relationship of the researcher to the content, but it's also changing the relationship of the researcher, the authors, to the publisher.

ZHOU: Exactly. This changes the whole research experience with this content, especially in the generative AI. Now, for example, normally we consider the four main roles of research – the authors, researchers, and also editors and reviewers. This is normally the roles. Then, you have all this AI, so actually there's not only the content, but also the AI-powered solutions to change these roles' experiences.

For example, for authors, the generative AI now we can use to help polish and help them draft the manuscript and polish the writing to increase their writing qualities, etc. It's especially useful for those whose native language is not English. Actually, this is the most popular generative AI application in the world now. And for researchers, we can apply the AI to identify the ways how to improve the manuscripts and also for the reviewers – identify the ways to improve the reviewer's quality and also help the reviewers to quickly see if there's any support or contradicting the literature in this. Also, for editors, we can apply the AI tools and identify if there's any relevant reviewers who can review this and identify any emerging journal opportunities – new opportunities.

For researchers, lastly, I believe that every researcher in the field should really have the personal research assistant to help them not only discover information, apply the information, and remember this information, knowledge, but also this AI – especially the AI agents can help them to plan, execute, and analyze experiments. Definitely, this will speed up the research outcome.

KENNEALLY: As the content grows – content is going to grow for a lot of reasons. There are going to be more authors who are enabled to submit to the journals. But also, AI content is

going to be part of the growth as well. What kind of concerns does that raise for you at Wiley about copyright issues and copyrighted content? I associate the copyrighted content with quality content, because it's content that's been curated, peer-reviewed, and there's an element of trust that's part of it.

ZHOU: Yes, indeed. So right now, the current status situation is that the AI governance is far behind the AI capabilities, which is dangerous. Actually, it's impacted the research and also the publishing, because it's very hard for the people to manage all these AI capabilities. That's why we need to create the legal framework to catch up to these technologies to have the response.

I do have several concerns about this. The first concern, as everyone knows, is copyright infringement. Today, generative AI generates content which infringes on copyright without permission. This is a problem. This already is a lot of the suit cases over the internet.

Even worse, recently, OpenAI and Google – they released a web crawler to automatically grab information over the internet to train the model – to improve the model. Although they allow their users, website publishers to disable, to block this, but I think as AI's capabilities expand, these things become much more complicated. That's one concern.

Another concern, actually, is that AI can generate content which is similar to the original content, but is not enough to be considered as copyright infringement. This is one scenario. Another scenario is it generates some content which infringes the copyright, but it's hard to detect. In both cases for the copyright holders, it's very difficult for them to enforce the rights – in both cases.

KENNEALLY: Well, Dr. Hong Zhou with Wiley, thank you very much. I want to turn to Dr. Namrata, because you have an interesting perspective. You're creating content, and you're working with many of the research institutions as well as the pharmaceutical companies and others who are going to be submitting content to journals at Wiley and elsewhere. As you do all of that work, what's your message for publishers about AI and content creation?

SINGH: I think what has happened as, as you mentioned, when ChatGPT was launched, that was the honeymoon phase in the initial part, where everybody was – that it's going to make life easier for especially content creators. But what I have seen and what I have realized in the last, say, eight to 10 months – it has put an additional responsibility on the medical writers. It's just not the reduction of the time which is important here. It is not only the efficiency which we have to focus on, but also the checking, all the legal aspects, the integrity part of it, the copyright, as we mentioned. It's not a very magic bullet kind of

solution, but it is something which is there. We cannot ignore it. And the publishers also will have to acknowledge.

I did come across some instances where people spoke about that if it is AI-generated content, then should it not be accepted? How do you differentiate? I think that is where more focus of the publishers should be – that even if it is AI-generated content, is there a human being who was behind it and who approved it, reviewed it, checked the authenticity of their content? So the human part is going to be there. In fact, even if the efficiency becomes better, the responsibility is becoming higher here.

So the balance in times to come with all these new developments happening and new technologies coming up – it's going to get tougher going forward. It's not something which is going to maybe make life very easy for the medical writers, but it's going to be – you know, you have to wade through that ocean of tools which are out there and the additional responsibility on you, because I've read about all these class-actions and lawsuits. The writers may come in the center of all this, because they're the ones who created the content. When you're writing it or when you're working with your clients, sometimes you are pushed to do things which you feel that you should not be doing. We're not the ones who own the content.

So that's an additional responsibility on us to educate the client, to tell especially the sponsors, the pharmaceutical companies, look, it's not so simple, even if it's an AI tool. You are not going to get a paper in, say, hours out of it. There's going to be some human interactions behind it. And similar for the publishers also.

KENNEALLY: I think that gets back to the point, Dr. Namrata, that Dr. Zhou was making about the importance of trust in content here. The challenge with a lot of the generative AI tools is the lack of transparency, the lack of identifying the sources. That must be really troubling for writers, because they do, as you say, have this real responsibility.

SINGH: Yes. A lot of guidelines and recommendations did come out over the last couple of months. We had an ICMJE guideline came out which mentioned about the responsibility of authors – additional responsibility. If you have used an AI tool, then you mention that in your methods section. You mention the name of the tool. You mention the version if it is there or the whole technology part behind it. This is where, I guess, the transparency works. But ultimately, the responsibility is on the author. And as we write on behalf of the author, then the responsibility comes back to us. But guidelines and recommendations do help us just to know what is right and what is wrong and what we can do and what we cannot do.

KENNEALLY: Well, I think what's coming through for me is the real responsibility, the sense of principle that is involved here. And I want to pull back a second and let people know about your background. You're a pediatrician.

SINGH: I'm a pediatrician.

KENNEALLY: Right. So you have dealt with patients and you have worked with people who are at one of the most important parts of their lives. And many of the writers that you have are women particularly trained in medicine who are trying to balance family and work and so forth. This is deeply a part of what you do. It's not just the business part. It's your professional concern.

SINGH: Yeah, it is. And another concern here is we have people working from various locations. So the technology part and the security part – somewhere, the integration with the IT and technology – all this is where everybody, whether it's a small company or a large company, I guess – that's where everything has to converge. We cannot have it in isolation. We cannot have a tool and just ask the team to implement and start working on it. So we do a small experiment on various tools, but that is only for an experiment's sake, where we are not working on the client projects or confidential data on that. We navigate these tools. We see what are the advantages, disadvantages, how much time it is reducing, or what is the productivity? And we have certain parameters which we evaluate on this. That's completely a research activity which we are doing internally. And there are very strict instructions to all the writers, and we have the whole IT kind of infrastructure also where they're not allowed to work on client projects yet.

KENNEALLY: I was a journalist in my previous life, and we always said consider the source. It sounds to me like that's equally important in your profession – consider the source.

SINGH: Yeah. And there's one more very, very important concern which is here. The tools now which are there – they have been trained on manually generated data until now. Now, what's going to happen in the next couple of years? Maybe more AI-generated content is coming, and then the training will be on those AI-generated content. Eventually, maybe the quality of the content overall – because everything is connected. We are not working in silos here. We are connected to so many various stakeholders. Eventually, if we don't have these rules and regulations in place, then it might have a very serious impact on the research integrity itself. Then what papers we are getting afterwards – can we trust them? Can we quote them as our references? That might happen.

KENNEALLY: Dr. Namrata, thank you for raising those important points. I want to turn finally to Carlo Lavizzari. Carlo, welcome again. You have contributed to CCC what I'll call a white paper here – a brief paper on "The Snapshot of the Relationship Between Artificial

Intelligence, Copyright, and Licensing." We have copies of this available for everyone here. I guess I want to ask you a little bit – because we assume a lot here when we hear about AI. Definitions are sometimes crucial to this. What do we mean by this AI or that AI? Which technologies particularly are using copyrighted materials as the inputs for their solutions?

SCOLLO LAVIZZARI: The models that have been catapulted in everyone's consciousness – and thanks for having me today – are the large language models, LLM, which is maybe one instance of a so-called foundation model, which is a broad, general-application AI tool. Typically, those are trained on either structured or unstructured inputs of enormous proportions, some really fantastical numbers, and are trained to be of general application. They are then sometimes enhanced by what I would call specific libraries to make them more suited to perhaps medical writing or fire regulation of buildings in architecture or whatever. So then you sort of specialize the general foundation model into the area you're interested in.

KENNEALLY: What are the rights of copyright holders in relation to all of those inputs and then to the outputs as well?

SCOLLO LAVIZZARI: Copyright – there are two questions, the input side and the output side, predominantly. On the output side, I think it resolves along the line of similarity. Are the works similar, or is there a new right that maybe creative people need to protect their style? But there's also freedom of expression that should put a limit to what you can protect. If you want to emulate a famous writing genre, it would be a pity if this was somehow inhibited by overprotecting this kind of similarity of writing style.

On the input side, definitely in order to make any valuable artificially intelligent tool, you will need entire reproductions of works that are scraped from the open internet or they are procured through special purchases, I guess – licensing of materials to use. I think in the previous panel, we heard that open access titles are frequently used. Frontiers is doing that with Google. So those are the source materials. They will still need to be processed, normalized in some fashion, because in order for the machines to work well, they do need either at the input side structured data, or then they need a – how shall I say – a calibration phase, where whatever they learn from unstructured data is tested against labeled data. I don't know if that answers.

KENNEALLY: It's very helpful. And Frankfurt Book Fair is a global event, but it's important to remind the audience that copyright is a national issue. Can you briefly tell us about some of the responses in various jurisdictions to these questions that have been raised? They're very new questions, so not everything is fully cooked at this point. But tell us where we're at.

SCOLLO LAVIZZARI: Absolutely. It's quite spotty. That's the very short answer. You have copyright law being territorial, so it applies in whatever country legislates. There have been a few countries that have made specific rules or are in the process of doing so.

In the EU, there are extensive rules on so-called text and data mining for non-commercial and commercial purposes. Text and data mining isn't identical to AI, but there is a significant overlap, so that many people engaged in AI will at least deploy text and data mining as well. There are exceptions and opt-out mechanisms that allow you in some circumstances to use text as input in a non-commercial sphere, especially, even though software, for instance, isn't included there, only literary works, let's say – articles, books. On the commercial side, there is an opt-out possibility for materials available on the open internet in the EU. That's a very specific rule.

In the meantime, the EU is busy legislating a so-called AI Act and is addressing what was said earlier, the transparency requirement, especially for foundation and generative AI models. The AI Act that is expected to go through, but who knows, in December would include an obligation to disclose what copyright-protected works have been used to train these generative models. I do think it is an important marker to force greater transparency into these models.

China has gone – if I may add this also – has gone out of its way to create 15 principles, I think around the 15th of September.

ZHOU: 18th.

SCOLLO LAVIZZARI: 18th of September. This has come into law – OK, another three days. They are principles, but they are pretty mandatory. As part of the principles is, in fact, transparency as to labeling of data and also what data has respect for intellectual property rights.

KENNEALLY: I've heard people compare the AI Act that's in the European Parliament right now to kind of a GDPR for AI, that if it were to come into effect in the EU, it would have an influence globally.

SCOLLO LAVIZZARI: That is predicted, in the sense that any significant AI giant or business or in publishing, people really publish for the world, or these AI entities are developed for the world. So to ignore a segment of 400 or 500 million people in Europe is going to be hard. Similarly, 1.3 billion people in China – it's going to be hard to ignore and say, oh, we're just not going to deploy our tools in those jurisdictions. I don't think many companies would consider that an option.

KENNEALLY: And as we've heard from Dr. Zhou and Dr. Namrata, there's going to be a great demand not only to use the tools, but to have materials to keep feeding the tools. That is a real challenge. What kinds of licensing would meet that demand? Can you tell us?

SCOLLO LAVIZZARI: I think as it was said earlier, there's an IBM short video that says AI is big on productivity and performance, but it's low on trust and transparency. These licenses can either be from segments of publishing, perhaps, that have large content that they can license, or it could be a collective license, where an agency such as CCC is used to deal into linking many-to-many situations. You have many writers, many publishers on the one side. You have many pieces of content on the other side used by different AI tools. So that is one mechanism is collective licensing.

It was said in a US submission to the Copyright Office – the key of copyright is credit – attribution – consent, and compensation. These type of licenses are well suited to handle consent, credit, and compensation.

KENNEALLY: There's one more point I'd like to ask you about, and maybe the others can help us with this, too. We talk about content all the time. I'm a writer, so I think content means words. But in scholarly publishing, content means images. It means data. It means all of these things. Is that what we're talking about? All of that comes under the umbrella of copyright, of course.

SCOLLO LAVIZZARI: A large segment thereof is in-copyright materials. Many images, photos, illustrations, and scientific articles are subject to copyright. There will also always be raw data that is not subject. In fact, the EU again has legislated in a Data Governance Act and has specifically excluded sensory data from any protection. When your fridge says to the supermarket in the future you're out of milk, that type of data is not going to be suddenly engulfed in this discussion about how to relate in-copyright content to the AI endeavor.

ZHOU: Maybe I can just add one more point to build on this. I think that AI is going to move from the more analytic AI to the generative AI, from the single modality to the multi-modality. Also, the research outcomes, like images, videos, and text is positively correlated, and AI can be used to generate this rich research output – images, AI can be used to generate videos, generate text, and even virtual reality in the future. But on the other hand, these rich research outcomes can be used to feed back to the AI to build an even better model. This is kind of the loop.

KENNEALLY: I'll leave with some advice. *It's always important to use your own judgment and common sense to verify information from multiple sources before making any important decisions or taking any actions.* It's great advice, and you could take it – it came from ChatGPT. All right?

I want to thank our panel – attorney Carlo Scollo Lavizzari, Dr. Namrata Singh, founder and director of Turacoz Group, and Dr. Hong Zhou, director of intelligent services and head of AI R&D for Wiley. Thank you all very much indeed. My name's Chris Kenneally. Thank you.

SINGH: Thank you so much.

(applause)

END OF FILE