STM Advancing trusted research

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands

TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

STM SUBMISSION TO UNITED KINGDOM INTELLECTUAL PROPERTY OFFICE (UK IPO)

Open consultation:

Artificial intelligence and intellectual property

Call for views: copyright and related rights

November 30, 2010

TO: AIcallforviews@ipo.gov.uk

The International Association of Scientific, Technical and Medical ('STM') publishers welcomes the opportunity given by the United Kingdom Intellectual Property Office to contribute to its views on AI and IP.

STM's members are at the forefront of digital innovation, providing stored and organized information, tagging and enriching content and creating ontologies. All of these advancements, together with the accuracy of the scientific record maintained by science and academic publishers, help to ensure that machine learning is endowed with both depth and accuracy. Publishers are already meeting the needs of the AI era, by developing tools, services and platforms that support and enhance machine learning. STM's members will offer an ever-increasing number of products and services to make their content available, so that high quality and accurate content sets can be used as training data in machine learning under various licensing schemes. The availability and accessibility of high-quality training data is vital for empowering AI developers with the licensed materials required to achieve both deep learning and to unlock the great potential of AI. Likewise, the wide array of licenses offered by publishers ensures that there are ample, accessible materials available for the continued training of both people and machines. This brief will under (A) provide input to Questions 1-9 in the section on copyright and related rights (B) make additional observations that STM believes are relevant and as requested.

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

A. STM answers to Questions of Copyright and Related Rights Section

Questions

1. Do you agree with the above description of how AI may use copyright works and databases, when infringement takes place and which exceptions apply? Are there other technical and legal aspects that need to be considered?

In STM's view, the tools and technologies currently available in the market and still subject of intense innovation that tend to be described under the umbrella "*Artificial Intelligence*" are, strictly, neither involving anything "artificial", nor anything comparable or approaching human "intelligence"[1]: rather, behind every AI entity stands one or more humans that make the arrangements to bring together computing power, algorithms and, ideally, relevant high quality "big data" in ways that have only recently become possible to be combined.

The above description of how copyright-protected content and databases and datasets are used is largely correct, but assumes that random content can be used by AI, a bit like a vacuum cleaner would hoover up dust and then, magically, the content gathered could be transformed into artificial knowledge rivalling human ingenuity. This is not so: machine learning for the most part is highly human supervised "learning" (machines "learn" differently from humans and the process of using copyright-content is central to engage in machine learning, but should not be confused with obtaining a human education).

In reality, the age-old principle "garbage-in, garbage-out" continues to apply to machine learning and AI. The quality and provenance of the input is critical to outcomes. Clearly high-quality input, curated, organized, enriched with metadata to give it context, normalized across a vast number of resources, will greatly improve the chance for a machine learning process to focus on what matters, rather than developing unforeseen bias or direction or forming correlations that for the past were good predictors of outcomes, but do not hold stability for the future, or are unable to cope with adversarial examples.

The above, STM's view demonstrates that far from being an "incidental" use, the targeted use of copyright-content, which tends to have all the attributes that make machine learning more successful in terms of human-centred outputs, the re-use of copyright content is intrinsic to AI, a part and parcel that must become part of the cost of doing business, in order to ensure a sustainable and stable flow of vital high-quality provenance-assured inputs for successful human-centred and ethical AI.

[1] In this regard, it may be useful to compare the terms used in other languages for the technologies. For instance, the Chinese term for Artificial Intelligence translates more literally into English as "man-made capacity for knowledge" (人工智能 – Ren Gong Zhi Neng), emphasizing that it is humans who make these technologies and that their capacity for understanding has no "super-human" attributes.

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

2. Is there a need for greater clarity about who is liable when an AI infringes copyright?

Like the internet does not provide information, AI does not infringe copyright; humans make information available on the internet and humans infringe copyright using, employing or perhaps setting in motion AI applications.

There are potentially very complex fact patterns that may emerge, but courts have traditionally been able to cope and interpret the infringement provisions of intellectual property statutes consistently and predictably. Using the an economic analysis of the law, may in future also sometimes help to allocate benefits as well as allocate risk and to internalize negative externalities of AI machine use. The UK has been leading the way in that early on the CDPA defines the author of computer-generated works as 'the person by whom the arrangements necessary for the creation of the work are undertaken'. Whilst this rule is one of assigning or allocating authorship ownership of copyright and related rights, it is submitted that the rule may also extend to assessing who is at least co-responsible for an infringement, ie directly or indirectly liable.

3. Is there a need to clarify existing exceptions, to create new ones, or to promote licensing, in order to support the use of copyright works by AI systems? Please provide any evidence to justify this.

In answering this question, we invite you to consider STM's additional observations under B. Under B. (I) below, you will find an illustration of AI-related services and products STM publishers are actively involved in. The list illustrates that the current legislative fame work does not in fact inhibit innovation in the field of AI or make it impossible for publishers to innovate. STM is therefore of the view that any new regulation of the AI space should be undertaken with great care and should also consider existing and future incentives and ecosystems which, in STM's view, are likely more collaborative, sometimes entirely private sector driven but sometimes entailing also public-private partnerships, a mix of standardization, best practices, self and co-regulation and licensing or jointly owning of IP assets.

In STM's view the UK has already legislated by introducing a number of exceptions as part of the Hargreaves Reform which legislation is an adequate basis. STM is aware of many hubs and clusters in the UK that engage in text and data mining analytics and other technologies. Nothing would indicate that the balance of copyright "rights" vs exceptions needs to be recalibrated. STM is of the view that the UK should let the EU DSM Directive play out and see to what extent the framework in the EU holds lessons when applied in practice, rather than starting to regulate the space of AI right now, tilting copyright licensing or tilting towards a broader set of exceptions from copyright infringement.

4. Is there a need to provide additional protection for copyright or database owners whose works are used by AI systems? Please provide any evidence to justify this.

In STM's view data is the new frontier and perhaps a consultation on rights management data infrastructure would be a useful follow up to this consultation. The difficulty in responding to this question is in part driven by the amorphous concept of "data". Data does not exist in a vacuum, especially not data that is able to be re-used. What is needed are standards and incentives to gather, collect, enhance and curate as well as curate and preserve data with context. This is the field of

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

metadata and many STM members see tremendous opportunities to contribute value to the AI-value chain or ecosystem, as publishers are extremely well positioned to organize data and metadata. For AI-technology related examples we refer to part B of this submission, but we stress that STM publishers are engaged in numerous standard-setting initiatives as well as in actively building a data and metadata infrastructure that will lead to new forms of knowledge entities, provisionally sometimes referred to as "knowledge stacks" – where human-written contributions, such as books, chapters or journal articles themselves can be regarded as metadata pointing and citing datasets and raw data usefully organized and rendered comparable across laboratories, even countries or scientific disciplines. In order to engage in this activity investment in infrastructure and standardization of thesaurus nomenclature generation is needed. In STM's view it would be ill-timed to either detract or add to layers of protection on data at this time, as the rate and pace of innovation and collaboration is such that the environment has not stabilized and any regulation, if simple is likely to have complex and unintended effects.

5. Should content generated by AI be eligible for protection by copyright or related rights?

Some AI-generated content is undoubtedly capable of copyright protection or protection in the form of related rights. However, in STM's view it is premature to add or detract from the level of protection at this juncture. The UK and its courts in some sense are leading the way thanks to the provision of computer-generated works being recognized as a category, but also through the simple fact that UK and English courts in particular are the chosen forum for many important controversies in the field of IP.

6. If so, what form should this protection take, who should benefit from it, and how long should it last?

In STM's view it is premature to add a layer of protection to AI-generated works. The present copyright and related rights framework offers already adequate and effective mechanisms to protect rights in IP associated with AI inventions or generated works.

7. Do other issues need to be considered in relation to content produced by AI systems?

In STM's view at this stage there are numerous ethical issues to be considered and we refer to our additional observations under (B) (I)-(III) below.

One of the areas that ethical issues may raise are transparency-related: AI will increasingly suffer from a lack of "interpretability" and "reproducibility". It may be critical to demonstrate and preserve how AI-decisions are made or have been made and to devise a scheme of escrow agency, similar to arrangements that are voluntary in the field of escrow agents for software source code. The risk is that AI-generating rightsholders will either through combination, divestiture or bankruptcy leave the industry and some AI entities will be hard to reconstruct or may lose their efficacy without a steward or owner. Similar to the stewards of open source and machine readable code, there is a need to decide who should maintain the integrity of training data and preserve training and other data actually used in the configuration, calibration or operation of an AI entity; under what considerations should access to such data be permitted. This may implicate rights of copyright and also the law of trade secrets.

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

Turning to the question of "infringement by AI", given that AI systems use substantial volumes of copyright materials both in the 'training' process and as part of any commercial deployment, transparency obligations may be necessary to enable rightsholders to trace infringing uses of copyright content ingested by AI systems. Similar to the remedy of "delivery up" and "provision of an accounting" for sales of infringing articles, analogous remedies should perhaps be considered to supplement the rights of copyright holders.

8. Does copyright provide adequate protection for software which implements AI?

STM does not have any specific comments other than to denote that the pace of innovation in AI is massive and there appears no shortage of protection of software in general. The phenomenon of copyright content and software "as a service" offers additional protection.

9. Does copyright or copyright licensing create any unreasonable obstacles to the use of AI software?

Copyright and database protection is a feature not a bug of the framework for innovation in STM's view. Using high-quality copyright-protected inputs for machine learning is simply a cost of doing business. There are ample ways to license content. AI-entities are also the new readers of books and scientific content. Many publishers will specifically publish with machine learning in mind; the use will become – to the extent it is not now – part of the normal exploitation of a copyrighted-work. The process is comparable, albeit not to be regarded as the same, as re-use of music in a record player was once upon a time a rarity and then became the norm and now is largely obsolete due to streaming. The same process can be expected in relation to information, knowledge, publications and machine learning.

B. STM Additional Observations

(I) Examples of actual deployment of AI technology in STM publishing

STM has organized a working group that brings together the expertise of its members on issues of publishing production, standards and technology, IP policy and protection, open science and open data policy, as well as public and government affairs as they relate to AI.

We are happy to be a resource to the UK IPO and others in the UK government on these issues as AI policy development continues.

In the interest of providing context for our engagement on these issues, STM provides the following non-exhaustive list of specific examples of areas of application of AI technology in STM publishing:

Recommend articles to readers (e.g., "more-like-this" or "you might also be interested in…")**:**

● Some recommendation engines are popular and trusted, including some that suggest articles from all publishers based on recent user interactions.

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

Improve search and browse with dynamic classification by topic:

● Taxonomies and classifications provide uniform subject terms across all articles in certain scientific domains.

● Some such taxonomies are semi-automatically learned, maintained and applied with AI tools.

Offer enhanced search & browse based on entity search:

● Deployed on some publisher database offerings. Match author-to-journal, paper-to-editor, or paper-to-peer reviewer

● Some publishers use peer review engines to all or a subset of data from Web of Science and other indexes.

Enrich citations:

● Publishing tools identify different structural components of textual references, used in accepted manuscripts during production of final version of record of a published article. Integrate primary publications with information in secondary databases and vice-versa

● Use of AI tools to match ID of research funders against a registry and improve details about funder official name, ancestors/descendants, country, etc.

Detect and prevent fraud:

● Some publishers use Cross Check plagiarism-detection service on every submitted manuscript.

● Some publishers use external fraud detection services that in turn deploy AI tools.

Journal suggestions:

● During manuscript submission process, authors are may be automatically given the suggestion of submitting to another or a set of alternative journals.

Machine-generated books:

● In 2019, Springer Nature published the first machine-generated book about the latest research on lithium-ion batteries.

Author (name) disambiguation:

● Used by some publishers. Experimental data checking

• Some publishers offer tools that have been trained by machine learning to so that authors and editors can detect assignment errors before publication.

ML-powered research workflow tools:

• Wiley have developed ChemPlanner, a tool that integrates computer-aided synthesis design capabilities with reaction mining approaches to offer productivity gains in synthesis planning.

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

- The Chemical Abstracts Service SciFinder[n] now includes the capability for researchers to perform a full retrosynthetic analysis for new or known molecules. This capability was developed in part with the ChemPlanner technology.

Research reliability:

• Scite.ai uses ML and NLP to determine whether a paper is supported or contradicted by the papers that cite it.

*****

The above list is by no means comprehensive or a reflection of the full potential the STM sector sees in applying AI in an ethical, trustworthy, human-centred and value enhancing way.

(II) The need to monitor convergence and possibly develop global consensus on industry or sector-specific frameworks for policy and ethical guidelines

STM is aware of a great number of efforts in the policy space to define, categorize and devise sensible objectives and policies related to AI. For convenience, these efforts may be classified as efforts of public bodies, corporate perspectives and public perspectives. In STM's view it is important to see to what extent these different perspectives converge and where they emphasize different attributes of AI. By way of example, but without endorsing any particular framework, we refer to the EU Commission framework "Ethics Guidelines for Trustworthy Artificial Intelligence" (April 2019), the framework of IEEE's "Global Initiative on Ethics of Autonomous and Intelligent Systems" (2nd edition 2017) and the Public Voice's "Universal Guidelines for Artificial Intelligence" (October 2018).

(III) The need for broad consultation on these issues; Any state intervention, government policy or regulation should be preceded by a definition of broadly consulted and then defined desired outcomes and to choose an evidenced-based approach in order to drive sound policy formulation and regulatory impact assessments. Finally, STM is grateful for this consultation not least because it demonstrates the USPTO's commitment to an evidence-based, open and transparent and inclusive consultative approach to approaching AI for the benefit of humanity, societal and planetary welfare.

STM Contacts

[redacted]

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

About STM

At STM we support our members in their mission to advance research worldwide. Our over 140 members based in over 20 countries around the world collectively publish 66% of all journal articles and tens of thousands of monographs and reference works. As academic and professional publishers, learned societies, university presses, start-ups and established players we work together to serve society by developing standards and technology to ensure research is of high quality, trustworthy and easy to access. We promote the contribution that publishers make to innovation, openness and the sharing of knowledge and embrace change to support the growth and sustainability of the research ecosystem. As a common good, we provide data and analysis for all involved in the global activity of research.