CMSS AI/ML TASK FORCE

ANALYSIS OF AI RISK IDENTIFICATION AND MITIGATION CHART

The CMSS AI/ML Task Force asked Work Group 2 to explore the risks presented to medical specialty societies by artificial intelligence ("AI"), and especially large language models ("LLMs") such as OpenAI's ChatGPT or Google's Gemini. Relatedly, the Task Force asked Work Group 2 ("WG2") to develop proposed safeguards for the use of AI tools by specialty societies. WG2 was not asked to examine the benefits and opportunities presented by AI and LLMs, but rather, while keeping such opportunities in mind, identify risk mitigation strategies.

Over the course of several meetings, WG2 developed a chart that identified three primary categories of risks, examples of each, and mitigation strategies for each. The AI Risk Identification and Mitigation Chart is attached as Exhibit A. This document further explicates each of the three risk areas, discusses some examples, and proposes risk mitigation strategies.

The three main risk areas are broadly categorized as (1) intellectual property protection, both avoiding infringement and ensuring copyrightability of content generated with LLMs; (2) reputation and integrity to ensure accuracy and avoid bias; and (3) IT security and privacy. As a cross-cutting approach to address all three risks, societies should draft and adopt acceptable AI use policies for staff.

1) Intellectual Property Protection

a) Protecting copyrighted content

One of the principal legal risks presented by LLMs and faced by all copyright-holding entities, whether for-profit, non-profit, or individuals, is the inputting of such content into LLMs. Such copying-and-pasting, without authorization, is presumptively copyright infringement.[1] Based on the experience of WG2 members, such infringement is happening in two major ways.

First, specialty society ***members*** are themselves taking society content (either publicly available or behind the membership paywall), and inputting it into LLMs for research purposes. This research is typically aimed at evaluating how quickly and accurately LLMs can learn and apply a given area of medicine. In the experience of WG2 members, the researchers using such content are usually unaware of copyright limitations, and such infringement is not intentional.

Societies have successfully communicated with their members about such research, and the members are frequently apologetic and aim to cooperate with their society. Some societies

[1] Litigation over the scope and application of the fair use doctrine is already unfolding, but it is too early to predict how caselaw and rules in this area will develop.

have been able to provide *ex post* authorization to permit the research to proceed, under certain conditions. If such conditions can be agreed upon, they are doubly useful as they both promote innovative research in the society's field of medicine and help foster cooperative relationships from the perspective of member management.

The conditions typically require: (1) the researcher inform the relevant LLM company that copyrighted material was inadvertently input into their LLM, request deletion of any content retained by the LLM, and agree that no further "training" occur using such content;[2] (2) limit the scope of society-copyrighted material to an acceptably small amount of content, and usually "lower value" content;[3] (3) expressly require agreement that no further research using society content will take place without express written consent; and (4) require the researcher to edit the paper to include a statement that permission was obtained from the society to use the content and also recognize the society in the paper's acknowledgements.

Where such conditions cannot be met, especially if large amounts of society content have been input into an LLM and a paper has already been published, societies may consider sending cease-and-desist letters to the researchers and/or informing the publisher that the society views the paper as infringing their copyright and demanding retraction..

In addition, societies can consider putting permanent banners in conspicuous but convenient parts of ***all*** society webpages, informing their members that all content is copyrighted and the society will consider member requests to use such content in AI-based research on a case-by-case basis, and noting such requests should occur prior to the inputting of society material into an LLM. Members of WG2 anecdotally report these approaches have been successful in decreasing this type of infringement.

The second major way society content is infringed is by web trawling and scraping. This infringement is more difficult to prevent and redress, as it is almost always conducted by automated bots. Societies are currently exploring two major options here: (1) anti-piracy vendors that are marketing new and innovative services to identify LLMs that have wrongfully ingested copyrighted content; and (2) metadata tags embedded into all webpages stating that the webpage contains copyrighted material that must not be ingested into an LLM without express written authorization[4]. Both approaches are very new, so their efficacy is currently unknown. Moreover, these approaches are intended to inform would-be infringers that the societies expressly claim copyright, which can lead to higher recovery of damages in any possible lawsuit. In other words, while these emerging approaches may not directly prevent

---

[2] As a practical and technical matter, it is currently unclear whether LLMs can actually "delete" or "unlearn" content input into an LLM. It is thus especially important that the researcher agree to only use LLMs that, on a prospective basis, expressly permit the user to opt-out of any training or retention.

[3] Individual societies must make their own determinations about what constitutes "lower value" content, but this will typically include publicly available content.

[4] An alternative approach would be the use of tools such as Robots.tx.

ingestion, they can warn would-be infringers that societies would intend to aggressively prosecute copyright infringement.

b) Ensuring copyrightability

Copyright law requires human authorship for a creation to be eligible for copyright protection. Thus, any materials generated wholly or primarily by LLMs, even if reviewed by a human being, will not be eligible for copyright protection. However, the technology is so new that the quantum of human authorship necessary to make a work copyrightable is presently unknown. Attorneys working in this area believe that merely having LLMs review and edit a document, or contribute some part of a document, will not void copyrightability. There is not yet caselaw to establish the bounds of this principle, so even the foregoing sentence cannot be known with certainty.

Nonetheless, LLMs have tremendous potential to maximize efficiency in the generation of written or visual work product. A categorical rule that would preclude all use of LLMs based solely on copyright concerns is therefore likely imbalanced to the operational needs of societies.

Instead, WG2 members believe societies would be well-advised to stratify their written work product by value. Documents that create significant value for societies, such as scientific materials, clinical guidelines, journals and magazines, or any other written products that generate revenue for societies, should be held to the highest standard to ensure copyrightability. This means that LLMs should be used sparingly – if at all – in connection with these documents. This approach should maximize copyrightability, especially in these early days when the exact bounds of permissible use of LLMs is unknown.

Lower value documents, such as documents strictly for internal use by societies, or external facing documents or communications that have lower value to societies, can be more readily generated by LLMs, where the society does not believe that copyright protection is especially important. Of course, as discussed in the next section, ***all*** material generated, edited, or reviewed by LLMs must be reviewed for accuracy and integrity.

2) <u>Reputation and Integrity to Ensure Accuracy and Limit Bias</u>

The second primary risk presented by AI and LLMs to societies is the possible generation of content that is inaccurate or biased. LLMs are inherently limited by the content that is input into them and on which they train. This problem is at least two-fold: inaccuracies and bias. Specialty societies are hugely dependent on their reputation for their content to be both accurate and unbiased, and LLMs present risks on both of fronts.

As to inaccuracies, LLMs can only generate content based on their inputs. If that content is inaccurate, outputted content will also be inaccurate. But LLMs face another risk – so-called hallucinations. LLMs do not truly "think" by themselves – instead, they generate material based on statistical models of associated content. Hallucinations occur where LLMs believe that certain words or phrases should be associated with each other, but in fact the generated

content, even if not facially nonsensical, may have no relationship to truth or reality, especially in highly technical and scientific areas like medicine. WG2 members have seen LLMs generate suggested treatments that are clinically inappropriate for the diagnosis.

Regarding bias, LLMs are also limited to the content of the inputs upon which they were trained. If any such inputs reflect intentional or unconscious bias, the outputs will also reflect such biases including but not limited to racial or ethnic and gender bias.

While LLM companies are working to refine the technology on both of these fronts, societies must review all content generated, reviewed, or edited by LLMs for accuracy and lack of bias. This risk may be especially pronounced for functions where LLMs are used expressly for their speed and ability to stay abreast in real-time, such as news updates or social media content. Human review is thus essential for ***all*** documents. Societies should promulgate acceptable AI use policies for staff, which should require human review of all content generated, reviewed, or edited by LLMs.

A specific risk of bias arises in the context of human resources and hiring. If an AI-based screening tool is trained on certain model resumes or CVs for a given position (which may favor certain backgrounds), the risk of the tool incorporating implicit bias is particularly pronounced. Societies would be well-advised to limit the use of AI for resume review. On the other hand, AI may be useful in generating job descriptions or qualifications, though such documents would still require human review.

3) <u>IT Security and Privacy</u>

The final category of risk presented by AI and LLMs is in the field of IT security and data privacy. LLMs, especially those available as "freeware," can present substantial risks to IT security and/or data breaches. Given the newness of this technology, well-meaning staff may seek out multiple LLMs, inadvertently exposing societies to IT and privacy risks. Furthermore, many AI tools and LLMs have document and data retention policies that may shift frequently, be difficult to understand, or by which the AI company fails to abide. Thus, society documents and data may enter the cloud and become "forever documents," inconsistent with best practices regarding data and document retention, or with society privacy policies.

Multiple approaches can mitigate this risk. First, societies should take proactive steps to inform and train staff about appropriate use of AI tools and LLMs. This should include a list of AI tools approved by IT and Legal departments. Some societies maintain this list as a hyperlink within the AI acceptable use policy, as the list can develop and change in real-time. Societies are also well-advised to pay for licenses for AI tools, as opposed to relying on free versions, as the subscriptions frequently include terms of service (i.e., contractual provisions) that forbid training or document and data retention. Considering the risk posed by "forever documents," the cost of the licenses is well worth the investment. To those ends, any subscription contracts with AI/LLM companies should be carefully reviewed by IT and Legal for compliance and

alignment with society data and document retention policies. This is especially important for AI tools that record, transcribe, or summarize staff meetings.

Finally, WG2 recommends a presumptive prohibition on the inputting of highly sensitive data (such as membership, PHI, or PII) into any LLM, even on a subscription basis. HIPAA, CAN-SPAM, state privacy laws, and general member management issues present a very high risk to a specialty society should any breach occur. Should a society staff member believe that an AI or LLM tool is sufficiently valuable for inputting of such data, the request should be reviewed by IT and Legal, and whichever tool is proposed must have robust technical protections and terms of service that adequately protect such data. Under such circumstances, a business associate agreement or data processing agreement may be warranted.

AI and LLMs, and the laws governing them are evolving rapidly and in real-time. This document may require concomitant revisions to stay up-to-date with technological and legal developments.