General Questions

1. AI systems do not create *ab nihilo* as in human creation; they only create remixes of their library of "training" material. In the same way as music producers must pay royalties for hooks they borrow, users of generative AI should pay royalties to everyone whose work has been (presumably) stolen in order to generate a given piece. This tech is likely to further kick the bottom out of the copywriting, music writing, and artistic industries, lowering the cost of business for the unscrupulous and reducing opportunities to earn for the ethical.
2. The wholesale theft of written text for AI systems is a huge problem for the writing and translating industries (the latter already undermined by Computer Aided Translation (CAT) programs). In the same way as CAT, this is going to lead to the amalgamation of the market into the control of a few major players, who provide bargain-bucket services based on the huge library of information their generators work from.
3. I am not a formal researcher, my apologies.
4. International consistency is only important in terms of where servers are primarily based; synchronicity between the US and the EU will cover the majority of international legal disputes. I believe the EU's approach of heavily taxing AI-created works is the correct way forwards, since said taxes can pay the unemployment of the creatives thrown out of work by these remixing machines.
5. Yes. Legislation is required on several fronts. Transparency over the works used in the generation of a given piece. Requirement to pay content creators for the use of any works still in copyright; perhaps using existing library rates for each content piece created from a library including that work. Requirement to enforce the minimum of existing legal ethics in production of content (cf. bomb-making advice; racist job descriptions; false validation of white nationalism etc). Due to the potential scale of this industry, costs should be based on percentage of income, not flat rates (as per European GDPR legislation). If AI content is to be considered copyrightable, copyright should default to the engineers in the project, not the company employing them.

Training

6. All and any written material online, whether protected from crawling or not. This includes fiction, non-fiction, and insane essays by lunatics and fringe theorists. Similarly, music hosted on streaming or torrent sites, and photos and text from all public-facing (and often private or semi-private) social media. These are generally seized through illegitimate or back-handed means that can most charitably be compared to tuna-fishing without dolphin-friendly nets. This misuse of owned content without remuneration is the foundational sin of this generation of generative AI.
    1. Online, through data-scraping. The volume required means that 3rd parties are usually also scrapers and data thieves rather than academic researchers. Perhaps the only benefit of the widespread piracy of material by AI companies is that research until now locked behind expensive paywalls might make its way out to the public.
    2. Licensing is generally absent. I think it should be based on Creative Commons, with strong enforcement of "no remix" and other subsidiary caveats.

3. Non-copyrighted material is used by anyone who can access it. To foster ethical small-industry AI models, it might behoove the Library of Congress and similar publically-funded institutions to make pre-fabricated libraries of different content types (text, paintings, photographs, etc) to foster experimentation & competition.
4. Not qualified to comment. However, for visual & audio recordings, storage will quickly spiral into unmanageable heights, to an extent where user bandwidth will not be able to cope with queries to the database.

7. My knowledge of how AI models are trained comes from off-record conversations with a troubleshooter who consults for Amazon on their AI products.
    1. Training materials are examined, and sometimes reproduced in the model's database. This reproduction can range from caching to a full duplication. Any use (even without duplication) of copyrighted material to turn a profit – especially when part of an unsupervised digital mélange – implicates the rights of copyright owners to be remunerated. I would argue this applies per use to generate a work, not per access to originally create a data set.
    2. Unclear.
    3. As well as retraining, models can simply be given directives on which returns to favour or disallow. Arguing over economic feasibility puts the cart before the horse. If a training model presents unethical or illegal returns, the model is ontologically not worthwhile, and does not *deserve* to make money if retraining is prohibitively expensive. Gangsters cannot be allowed to excuse their crimes on the grounds of *fait accompli* – this only incentivizes future criminals to go large and get "too big to fail".
    4. As mentioned in 5, I believe all generative AI must be forced into transparency over their dataset, and therefore what material produced a result.
8. The only case in which I believe unauthorized use of copyrighted works could *possibly* constitute fair use is if it was part of a student or purely academic process. Any commercialisation or profit from the use of copyrighted works would be morally equivalent to remixing music and therefore require the payment of royalties.
    1. Purpose and character is absolutely key to the use of copyrighted works in training AI. If a distinction is to be drawn – and I think it would be specious, like asking whether one should pay for only the basecoat or final coat of paint on a house – I would say fine-tuning is even less worthy of consideration as fair use. While pre-training could be considered equivalent to a human's simple immersion in the culture their art arises from, fine-tuning more closely apes a forger's training.
    2. Entities that distribute copyrighted material without paying the producers are pirates, and should be treated as such.
    3. Fair use applies only to the original non-commercial or research purpose. As an example, look at all the drugs created with the assistance of government funds that are later turned into cash cows by pharmaceutical companies. A model that profits from copyrighted works must include remuneration for the producers of said works, regardless of the model's original provenance. Otherwise, for-profit developers will

simply bribe university departments to start their models for them and then take over the project (s) once profitably large.

4. The quantity of training materials is vast, but so is the streaming library listened to by most musicians. The creators of those songs get royalties (even if only a pittance) each time their song is played, and creators of copyrighted works used for training materials should also be paid. The fact that generative AI requires piracy on a mass scale never before imagined is not a 'get out of jail free' card for the fact it requires piracy on a mass scale never before imagined. Imagine the insanity of arguing that Napster wasn't infringing copyright simply because it infringed quite so many copyrights.
5. The potential effect is limitless because of the exponential knock-on effect of cheap AI works pushing human creators out of business, in the same way that consolidated supply chains in other industries have destroyed many small third-party businesses through "tyrannies" of scale. The inquiry should be whether it affects the market for that general class of works, because the audience for a specific tome or musician over another cannot be quantified in an objective sense (the mood one is in at any moment being entirely subjective).

9. Creators of AI training materials should get the affirmative consent of copyright holders for the use of their works.
    1. This should apply to commercial uses, with an option to also opt-out of research uses too.
    2. Metadata like robots.txt or even simply tags in the css of eBooks could serve as flags denying use of content for AI training uses.
    3. Very few, if the willingness is there. Again, I need to remind you that the scale of crime does not disqualify it from punishment: you wouldn't drop a Grand Theft Auto charge just because of the *volume* of cars stolen.
    4. As per European GDPR legislation, penalties for content infractions should be based on percentages of company *turnover*, since most IT companies make negligible profits. A commissioner's office may need to be established.
    5. No. But the company owning the copyright should absolutely be encouraged to pursue AI generators using their materials.
10. When submitting content for registration (copyright, provision of copy to Library of Congress &c), willingness to licence it can be marked at the same time. Then, libraries of record can quickly and easily provide libraries of licensed material to interested parties.
    1. Not really, no. The world is global now.
    2. Again, national organisations of record, the MPIAA, BMPA etc. could provide libraries of licensed content. The law should not be weakened further to facilitate AI training, that way madness lies.
    3. The regime should be compulsory on the AI end, but content creators should have the right to opt out of licensing their creations. Royalty rates could easily be set, allocated, reported, and distributed in the same way as library rates are currently sent out.

4. Not really, no.
5. If you're talking about rates, of course. How can you compare a page of a novel to a 1000 word essay to moments of a song? If you mean should it be easier to pirate books than movies? No.

11. As in European GDPR legislation, each company (and third-party provider) should have a data officer who is responsible for the company's compliance with data laws. There should be a clear chain of custody & responsibility for datasets from content creator to AI generator. The main issue obtaining licenses is simply *volonté*, because it's easier to "move fast & break stuff" than it is to be an ethical practitioner.
12. The degree is, to an extent, irrelevant. How much does stolen bicarbonate of soda contribute to a cake? It's still stolen, and it's still a part of the finished work that can't be isolated from the whole. Every item in the database contributes to its returns, and thus all should be compensated.
13. It would grossly depress the predatory interest of hedge funds, and slow development to a pace where proper rails and safeguards can be developed to be installed in any new models from the ground up.
14. Nope, pretty much covered it all. The important thing is to stick with the GDPR/Finnish model of fines being percentages of turnover, because 'tech bros' will soak defined sums as the cost of doing business, especially with rising inflation making specific sums increasingly meaningless over time.

Transparency & Recordkeeping

15. As mentioned since the earliest questions, yes, transparency in both training models and training datasets is vital.
    1. As with code, reference to libraries rather than component data is fine, as long as what said components are is also easily available.
    2. Disclosure should be readily available. Compare the Open Gaming Licence used by 3rd party providers of material for Dungeons & Dragons (issued by Wizards of the Coast). AI works could very simply provide URL links to the copyrighted items used in their creation.
    3. The obligation for clear and transparent supply chains, as per modern slavery acts.
    4. Initial outlay would be significant, in terms of data entry, at least.
16. Notification should be issued to the library or other content lister of record, and quarterly reports returned to copyright owners as with royalties and sales under the current music industry model.
17. Modern slavery laws and others regarding transparency in the supply chain could be argued to be relevant, particularly since engineers receive no credit for the works their AI models create.

Generative AI Outputs

If your comment applies only to a particular subset of generative AI technologies, please make that clear.

Copyrightability

18. I think that music once again provides a precedent. Sampling a track does not eliminate the new song's creator's copyright rights. Similarly, a human using generative AI *could* be argued to be the author of the piece. However, I think significant input in both selecting fine-tuning material and inputting prompts is required; iteration is almost certainly necessary. Otherwise the result should be considered a Creative Commons remix at best.

19. I think the opportunity should be taken to incorporate 'copyleft' community standards into the copyright regime – many of these would be suitable for AI-generated works.

20. No. No. Yes.

    1. See previous comments ref. 'copyleft' protections. Copyright covers the *ab nihilo* creative energy of human effort, and AI-generated material, being simply remixes of existing works, is a different category – in terms of sheer potential quantity over time, if nothing else.

21. Promotion of AI-generated material is the opposite of the promotion of the progress of science and useful arts. It is a direct and massive attempt to sideline and minimise the human contribution to the advance of knowledge and understanding.

Infringement

22. Yes, in any case in which the work in question is used to train or inform the generated output.

23. Substantial similarity has to be extended, in my opinion, to cover the *sui generis* of a body of work rather than specific single *oeuvres*.

24. I think that *Armory v Delamirie* applies here; if the developer refuses to maintain records, it should be assumed that they are concealing malfeasance.

25. The primary liability resides with whoever presents the AI-generated material to the public. However, it is important to stress the ability to pass this cost down the supply chain, and for said supply chain to be transparent enough for creators to know *where* the infringing material entered the system.

    1. Nope. They should be treated like research models until they are turned to commercial purposes, at which point primary liability resides with whoever presents the AI-generated material to the public. The foundation (or github registry, or whatever) that maintains the registry should be free to make "don't blame us" a key licence condition, in order that commercial considerations not unduly depress the options of good-faith non-commercial users.

26. All information under 17 USC 1202b applies to the treatment of that information. Outputs should include URL links to all relevant copyright information.

27. Tech companies will do their darnedest to avoid any responsibility (cf Uber, Facebook, Deliveroo, WeWork, AirBnB, *ad nauseam*), so *Armory v Delamirie* has to be central to all legislation regarding copyright liability in AI-generated output.

Labeling or Identification

28. Yes. Consider requiring a URL or QR code to the materials used be embedded into the piece as standard.
    1. Whoever presents it to the wider public.
    2. Only the willingness to do so.
    3. Fines, as a percentage of turnover (not profit) as per European GDPR legislation.
29. Tineye and similar reverse image searches, the programs used to train self-driving vehicles, and the human instinct used to sex chickens. The limitation is the sheer quantity of variables in the sheer number of works that can be generated using AI by bad faith actors.

Additional Questions About Issues Related to Copyright

30. To the AI-generated material? None. To the person whose name, likeness, or vocal likeness has been stolen? The same rights as they currently have over the public projection of their image.
31. Federal law should always set a floor for state protections, including the right to the immediate removal of right-breaching content (which could, while we're at it, help with revenge porn, deepfakes, etc).
32. Ah, now. This is difficult. I can paint something "like" a van Gogh that includes no sunflowers or stars, and I haven't impacted VG. I could write a novel in the style of Stephen Fry, or a song in the vein of *every gorram Ed Sheeran song*, but I wouldn't be considered to be stepping on their toes. This is the only area where I even slightly sympathise with the AI generators. Protection should likely take the same form as current "substantial similarity" rules to avoid setting a creeping precedent that stifles *human* creativity.
33. Similar to deepfakes, it's important that copyright law defend a person – legal person or flesh and blood – to their own image and public presence. Given the extreme depths of possibility presented by AI, it may be worthwhile connecting this not only to copyright, but to libel and slander laws too.
34. Pretty much everything, I think. Thank you for taking the time to read this work.

Kind regards,

Mr Mathieu Mansur al-Haqq Antoine Moran, B.A. (Hons)