To: Hon Mark Dreyfus KC, MP
Attorney-General

From: Nik Samoylov
President, Campaign for AI Safety
nik.samoylov@campaignforaisafety.org

15 July 2023

Copyright matters in relation to AI models

Dear Attorney-General,

I write to you on the matter of application of copyright law to training and use of artificial intelligence (AI) models, particularly generative AI models such as GPT and Stable Diffusion.

This letter is in the context of the ongoing policy discussions regarding the regulation of AI in Australia and the Ministerial Roundtable of Copyright, which I understand will be turning its attention to AI in the coming months.

Our campaign is concerned with avoiding catastrophic risks of AI. We advocate for an indefinite moratorium on training dangerous AI models. We view respect for copyright as a crucial element of avoiding AI risks in general and promoting a culture of responsible AI.

In this letter I (1) outline our campaign's recommendations on the application and direction of copyright law in Australia, (2) provide two case studies, (3) explain the importance of copyright for AI safety, and (4) give further reasons for our campaign's recommendations.

1. Our policy recommendations

Generative AI arrived largely unanticipated by the public, creative professionals, publishers and relevant stakeholders. The vast majority of creators never imagined that their works would be used for machine learning, while copyright law was not drafted to address the ease and speed at which content can now be generated with AI. This has created a climate of uncertainty for Australian creative professionals as well as businesses using AI-generated content.

Therefore, our campaign recommends that the copyright regime is clarified and reinforced in relation to training of AI models on copyrighted materials. Specifically:

1. Third parties who wish to use copyrighted materials in training AI models must obtain **consent** from the owners of materials.

2. Third parties **must not be allowed to coerce or trick copyright holders** to provide such consent. For example:

a. Consent must not be part of terms and conditions of unrelated services (such as video or gaming distribution platforms). This way social media companies will not be able to deny services or features of their platforms to those who do not consent to such use of copyrighted materials.

b. Specificity of consent means that it needs to be a separate agreement, ideally a contract with consideration, which copyright holders have an opportunity to take time to consider, review, and negotiate. It must not be a checkbox underneath a registration form on a website.

3. **Consent must be granular**. For example, copyright owners should be able to specify if they allow third parties to have technical means to generate works "in the style of" the author of the materials. This would be similar to granular consent for processing data under UK GDPR[1].

4. Copyright holders must have effective means of **negotiating and receiving fair compensation** for the use of their copyrighted materials (such as opt-in collective bargaining mechanisms via established organisations).

5. **Data used in training AI models must be fully referenced**. References to the works used in training must be made publicly available.

6. Training models using "synthetic data" generated by other models trained on copyrighted materials must be considered equivalent to training on those copyrighted materials.

7. Legal liability in case of infringement of copyright must apply both to parties that train models and parties that use those models to generate content.

8. "Fair dealing" provisions relating to research and study must not apply to AI model training. Practices such as "data laundering"[2] need to be outlawed.

These clarifications require prompt and urgent legislation. This will provide certainty to the AI industry, the Australian business community, copyright holders (including artists, writers and publishers), and the wider public.

The terms I used above have the following meanings:

- **AI models** refer to generative AI models and systems (including constellations of models), including large language models (e.g. OpenA's GPT-3), multimodal models (e.g. Google Deepmind Gemini), large diffusion models (e.g. Stable Diffusion).

- **Training** includes fine-tuning as well as use in the context window in prompting[3].

---

[1] What is valid consent?, UK Information Commissioner's Office
[2] AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability, Andy Baio, Waxy, 30 September 2022.
[3] Inserting materials into the context window (prompt) of AI models may serve as an alternative to training and therefore must be treated similarly. Example long context window research: "[2307.02486] LongNet: Scaling Transformers to 1,000,000,000 Tokens"

2.1. Case study of Chat-GPT powered by GPT

Chat-GPT is a famous AI-powered chatbot that uses a variety of models in the GPT series (including GPT-3.5 and GPT-4) to generate answers (or "completions") to user inputs ("prompts"). GPT models were trained on various undisclosed sources of information.

It has been alleged that many of those sources of data are copyrighted or privacy-protected[4]. Thanks to the training data, Chat-GPT is able to reproduce (albeit with some error) portions of copyrighted works and summarise them[5].

This creates two big issues for copyright holders. Firstly, they may face diminished demand for their works when imperfect reproductions and regurgitations are accessible from OpenAI and Microsoft. Secondly, because of the secrecy around the training data used by OpenAI[6], copyright holders will struggle to determine if and when their rights have been violated.

Without intervention, AI companies like OpenAI stand to gain unfairly through the use of copyrighted work, where the authors did not have AI in mind, do not consent to such use, are not compensated, and do not enjoy attribution or moral rights.

2.2. Case study of Google Deepmind's Gemini

Google Deepmind is working on a system called Gemini, which they claim will have a number of advanced capabilities[7]. It has been reported[8] that Gemini may be trained using materials uploaded onto Google-owned YouTube platform.

While it is not clear what data is being used for training, if Google indeed is using videos hosted on the YouTube platform, that would be a violation of users' trust. Even if the terms and conditions allow Google to use footage in this, such a use would not have been in the minds of users at the time they agreed to these terms. Therefore legislation must be drafted to prevent a huge and permanent appropriation of a user's work with no compensation.

---

[4] "ChatGPT Language Model Litigation", Joseph Saviery Law Firm
[5] "OpenAI's ChatGPT and GPT-4 'memorized' these books", The Register, 30 May 2023
[6] Only cryptic names for datasets are given in OpenAI's GPT-3 paper, such as "Books1", "Books2", etc. "[2005.14165] Language Models are Few-Shot Learners"
[7] "Google DeepMind CEO Demis Hassabis Says Its Next Algorithm Will Eclipse ChatGPT", Wired, 26 June 2023
[8] "Why YouTube Could Give Google an Edge in AI", Jon Victor, The Information, 14 June 2023

3. Why copyright matters for AI safety

There are several reasons why enforcement of copyright as described in our recommendations matters for AI safety:

1. Enforcement of copyright in regards to training data has potential to **cool down** competitive race dynamics in the AI industry, which are widely seen as detrimental to ensuring robustness and safety of AI[9].

2. One of the dangers of advanced AI is mass unemployment. Artists and content creators are some of the first industries to experience this AI risk. It is important that the Australian Government sets a precedent that **protects workers at risk of unemployment** because of AI; especially when the relevant AI relies on misappropriation of their work. Strengthening the copyright regime is a step towards fair treatment of people affected by automation.

3. AI labs often make reference to advancing AI that "benefits all of humanity" and avoiding "uses of AI or AGI that harm humanity or unduly concentrate power"[10] OpenAI is a not-for-profit company whose "primary fiduciary duty is to humanity". Another AI lab, Anthropic, is a public benefit corporation[11].

   Futuristically, these companies have set their sights on developing "AGI" (AI that is defined to have human-level capabilities in many domains), which they want to "align" with humanity's values. So that when/if control is ceded to intelligent machines, these computers will not harm people and will abide by humans' wishes.

   At the same time, AI companies do not currently compensate artists and other copyright holders — contrary to the law and to public opinion[12]. Such companies cannot be the ones to tell powerful machines what is right and what is wrong.

   Even as the concept of AGI is in the future, the Australian Government needs to set **a precedent that "AI", "AGI", references to "humanity", etc. cannot be used as an excuse** to disregard the laws of the land or the will of the people.

4. Our recommendation #5 is "Data used in training AI models must be fully referenced. References to the works used in training must be made publicly available". This should apply to all generative AI commercially available in Australia.

   The idea of datasheets of datasets for AI models was proposed more than five years ago[13] and its value is widely acknowledged, but it is not widely adopted in the AI

[9] "AI Is Not an Arms Race", Katja Grace, Time, 31 May 2023
[10] OpenAI Charter
[11] Company \ Anthropic
[12] According to our campaign's recent poll, 66% of American adults agree that "artists and content creators should be paid if their work is used in training artificial intelligence (AI) models", "USA AI x-risk perception tracker", Campaign for AI Safety, 1 July 2023
[13] "[1803.09010] Datasheets for Datasets", Timnit Gebru, et. al, 2018

industry at present[14] because doing so would expose AI companies to more copyright lawsuits as more people find out that their materials were used in training AI models.

In the context of AI safety, **transparency of data sources is important in aiding safety research**. For example, there is ongoing debate on the existence and the extent of "emergent abilities"[15]. These are unforeseen abilities hypothesised to emerge from large language models as they become more complex, including abstract thinking and dangerous outputs (such as ability to generate blueprints for poisonous chemicals[16] and bioweapons[17]). This unpredictability makes such abilities extremely difficult to prevent, detect, limit or moderate. If AI can indeed develop these abilities, which appears disturbingly likely, we must approach the development and study of new AI models with a far greater level of caution.

Alternative explanations of these abilities have been proposed, such as training data size[18]. But because the lists of references to training data of frontier models have not been released to the public, researchers face difficulty resolving this important issue.

This creates an "unknown unknown" factor in AI safety that muddles policy discussions. But this factor could be easily removed if companies were required to publish their datasheets of datasets.

[14] "Generative AI's secret sauce — data scraping— comes under attack", Sharon Goldman, VentureBeat, 6 July 2023
[15] Ability is defined as "emergent if it is not present in smaller models but is present in larger models", "[2206.07682] Emergent Abilities of Large Language Models", Jason Wei, et al.
[16] "[2304.05376] ChemCrow: Augmenting large-language models with chemistry tools", Andres M Bran
[17] "Dual use of artificial-intelligence-powered drug discovery", Fabio Urbina, et. al., Nature Machine Intelligence (2022)
[18] "[2304.15004] Are Emergent Abilities of Large Language Models a Mirage?", Rylan Schaeffer, et al.

4. Other considerations

I list other considerations why Australia needs a policy of strong copyright protections for creators of data used in training AI models.

1. **National interest.** Australia has a negligible AI sector and there are no frontier AI labs domestically. At the same time, Australia has a thriving arts and entertainment sector, which contributes $14.7 billion per year[19] in value added, with another (broader) estimate putting the value to $112 billion per year[20]. Protecting the interests of our local, existing, thriving, and proven industry must be a priority over the interests of an unproven, speculative, overwhelmingly foreign industry.

2. **Protection for suppliers in an oligopsonistic market.** Only a few businesses have the financial and computational resources to train powerful AI models. The same businesses or their affiliates happen to own dominant social media and search engines, which puts them in positions of extreme market power. They may attempt to include waiver of copyright into their general terms and conditions of service[21]. Individuals will not have the resources or capability to fight against such overreach.

3. **Ensuring competitiveness of the local AI industry through compliance with strict law in other countries.** If Australia enforces a strong copyright regime, local AI players will have the advantage of training their models in an ethical and compliant manner by default. Copyright in training AI models is set to remain a contentious issue overseas, with some countries choosing to enforce strict compliance. If Australia has strong copyright protections, Australian-trained AI models may enjoy a regulatory advantage in countries with strict compliance (similar how as a result of EU GDPR Adequacy Decisions, Canadian firms have an advantage over Australian firms in entering European markets).

[19] Background Brief: Economic Importance of the Arts and Entertainment Sector, The Australia Institute (2020)
[20] "The Value of the Arts", Sculpting a National Cultural Plan, Parliament of Australia (2021)
[21] "Google Says It'll Scrape Everything You Post Online for AI", Thomas Germain, Gizmodo, 3 July 2023

5. Next steps

I request that the considerations of AI safety are duly reflected in policy deliberations on this matter. I will be happy to be of assistance in case they need to be represented during the Roundtable or if you want to hear about the topics of AI safety and ethics further.

I see that the ideal end goal of these discussions should be legislation that reinforces and clarifies existing legal provisions on copyright and introduces further protections for copyright holders.

6. About the Campaign for AI Safety

The Campaign for AI Safety is a not-for-profit association with members from around the world. We are an association of people who are concerned about the dangers AI poses to humanity and we advocate for both a stop on the advancement of AI capabilities and stronger regulation that prioritises safe and responsible AI. We are not affiliated with any political group. For more information, and to read our policy recommendations, please visit www.campaignforaisafety.org.

I thank A/Prof Simon Goldstein and Mr Raphael Royo Reece for the useful comments in drafting this letter.

Yours faithfully,
Nik Samoylov