NLIU-Ikigai Policy Case Competition 2024

*ch*AI *pe charcha*
Harmonizing AI Innovation and Copyright for Digital India

Team Name: Artificial Intelligence meets Natural Stupidity

Team Code: TC19

I. STATEMENT OF PROBLEM

Nowadays, the conversation of the town is GenAi and it's potentials. A critical aspect of this conversation is the prevalent practice among GenAI companies, where copyrighted data is employed for AI training without obtaining consent from intellectual property (IP) holders or providing due compensation.

Essentially, the issue at hand can be distilled into two key questions: *First*, does the utilization of copyrighted data by GenAI companies for training Language Models (LLMs) constitute an infringement of copyright? *Second*, if it does, can such actions find refuge within statutory exemptions carved out for 'fair dealing'?

The challenge at hand is to create a delicate balance among the varied interests of stakeholders, addressing the legal and regulatory void, and establishing an environment that fosters innovation, economic growth, and scientific and artistic progress. Achieving this balance is crucial for navigating the evolving landscape of GenAI while upholding the principles of fairness, legality, and respect for intellectual property rights.

1. Comparative View

Policy-making in the digital realm, including GenAI regulation, generally is of one of three types: driven by market forces, or state interests, or by individual rights.[1] Historically, the U.S. favoured market-driven regulations with significant global influence. In contrast, recent trends, especially in the EU, lean towards prioritizing individual rights. Meanwhile, China's state-centric approach presents unique risks and challenges. As for AI training, the UK, initially in line with countries like Singapore and Japan in allowing the use of copyrighted material, is now shifting towards a more balanced approach. This change mirrors the evolving legal landscape in AI, marked by global legal uncertainties and ongoing debates, as seen in the U.S. and Canada, to determine the most effective regulatory strategies.

Jurisdiction	Policy	Key Points	Additional Notes
European Union	Mixed	EU Copyright Directive: Permits exceptions for text and data mining for non-commercial research purposes.[2]	Unclear whether exceptions extend to commercial GenAI training. Ongoing litigation challenging the boundaries of fair use for research-driven AI.
United States	Fair Use	Fair Use doctrine allows limited use of copyrighted material for purposes such as criticism, commentary, and research and learning.[3]	Arguments being made that Fair Use should extend to GenAI training for research purposes. Lack of clear legal precedent creates uncertainty for commercial applications.

[1] Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology* (Oxford University Press 2023)

[2] 'Generative AI, Copyright and the AI Act - Kluwer Copyright Blog' <https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/> accessed 4 February 2024.

[3] 'Fair Learning' (*Texas Law Review*, 20 March 2021) <https://texaslawreview.org/fair-learning/> accessed 4 February 2024.

China	Not allowed[4]	No specific provisions in copyright law regarding AI training.	Government drafts prohibit using copyrighted material. Models which have more than 5 percent of training data as copyrighted is to be blocked.[5]
Japan	Limited exceptions	Copyright law permits exceptions.	Copyright law continuously monitored, debated and updated.
Canada	Fair Dealing	Fair Dealing doctrine similar to US Fair Use, but with greater emphasis on education and research.	Applicability to GenAI training unclear and subject to ongoing debate.

2. Interests and Concerns of Stakeholder

Stakeholder	Key Interests	Concerns/Wants	Potential Conflicts
GenAI Companies	Economical access to varied training data. For creation of competitive AI solutions.[6]	Restricted or expensive access to data due to copyright laws. Ambiguity in legal and ethical responsibilities.	With IP owners regarding data rights increasing their cost of operation.
IP Holders (Artists, Authors, Creators)	Safeguarding intellectual and moral property rights. Fair compensation for using their works in AI training. Guarantee against AI diminishing or substituting their creations.	Reduced control over usage and modification of their works. Unfair distribution of economic gains from AI applications. Risk of AI models competing with or supplanting their creative work.	With GenAI firms over affordable access to extensive datasets. Compensation requirements and concerns about AI distorting original creations.
Users	Cost-effective, easily accessible AI applications and services. Optimal use and advantage from AI tools and platforms. Protection against privacy breaches and	Limited clarity and understanding in AI's decision processes. Bias in AI leading to unfair and adverse consequences.[8]	Demand for ethical regulations and standards in AI creation and implementation.

[4] 'China Proposes Stricter Curbs on Training Data and Models Used to Build Generative AI Services in Bid to Tighten Security | South China Morning Post' <https://www.scmp.com/tech/article/3237873/china-proposes-stricter-curbs-training-data-and-models-used-build-generative-ai-services-bid-tighten> accessed 4 February 2024.

[5] 'China Proposes Tougher Curbs on Generative AI Training Data and Models' (*South China Morning Post*, 14 October 2023) <https://www.scmp.com/tech/article/3237873/china-proposes-stricter-curbs-training-data-and-models-used-build-generative-ai-services-bid-tighten> accessed 4 February 2024.

[6] 'AI Companies Have All Kinds of Arguments against Paying for Copyrighted Content - The Verge' <https://www.theverge.com/2023/11/4/23946353/generative-ai-copyright-training-data-openai-microsoft-google-meta-stabilityai> accessed 4 February 2024.

	AI-based discrimination.[7]		
Governments & Policymakers	Encouraging innovation and economic progress via AI advancement. Safeguarding national security and public welfare against AI misuse. Advocating for ethical and beneficial societal use of AI.	Balancing diverse stakeholder interests. Minimizing risks and averting damage from AI misuse. Tackling the social and economic impacts of AI integration.	Difficulty in balancing the interests of IP holders and GenAI firms. Danger of excessive regulation impeding innovation.

3. Analysing anti-copyright arguments

Argument	Description	Potential Drawbacks
Knowledge as a Commons:	Information and ideas are inherent to humanity and should be freely shared, not artificially restricted by copyright.[9]	Challenges the financial incentive model for creative industries. Potential loss of income for content creators. - Difficulty in attributing and rewarding creators for their work.
Innovation and Progress:	Strict copyright can stifle creativity and limit the building upon or remixing of existing works.	Potential for plagiarism and unauthorized appropriation of original works. Difficulty in ensuring fair compensation for original creators. Concerns about loss of control over creative expression.
Fair Use & Public Interest:	Copyright limitations like fair use should be expanded to allow greater access and usage of copyrighted materials for non-commercial purposes.	Ambiguity in fair use definitions can lead to legal uncertainty and litigation. Potential for misuse of copyrighted material under the guise of fair use. Difficulty in balancing the interests of creators and users.

4. Justifying regulatory response

Concern	Evidence[10]
Loss of Control & Moral Rights	Artists and creators may lose control over how their works are used, modified, or commercialized within GenAI models, potentially violating their moral rights.
Economic Impact	Extensive use of copyrighted works in training data without fair compensation

[8] 'What Do We Do About the Biases in AI?' <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai> accessed 4 February 2024.

[7] 'AI Ethics in Focus: Addressing Bias, Privacy, and Transparency Challenges – Human Made' <https://humanmade.com/ai/ethics-in-ai/> accessed 4 February 2024.

[9] Dennis WK Khong and SU MON, 'ARTIFICIAL INTELLIGENCE AS A COMMON HERITAGE OF MANKIND' (2023) 14 UUM Journal of Legal Studies 113.

[10] Michael M Grynbaum and Ryan Mac, 'The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work' *The New York Times* (27 December 2023) <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html> accessed 29 January 2024; Emilia David, 'Getty Lawsuit against Stability AI to Go to Trial in the UK' (*The Verge*, 4 December 2023) <https://www.theverge.com/2023/12/4/23988403/getty-lawsuit-stability-ai-copyright-infringement> accessed 29 January 2024.

& Unfair Compensation	could harm existing creative industries and limit financial incentives for content creation.
Derivative Works & Plagiarism	AI models trained on copyrighted works may generate outputs that are too similar to the originals, blurring the lines between derivative works and plagiarism.
Market Competition & Disruption	GenAI models trained on copyrighted works could be used to create competing commercial products or services, threatening existing markets and industries.
Filter Bubbles & Cultural Depletion	Reliance on existing copyrighted works for training data may lead to GenAI models reinforcing existing biases and neglecting diverse or marginalized voices, contributing to cultural homogenization.

5. Broad principles on which new IP ecosystem should be based:

    1. Promoting innovation of AI models, especially Indian models by allowing easier access to copyright data and legal certainty.
    2. Promoting use of Indian publishers' data in training GenAI models to reduce AI bias.
    3. Ensuring creator's profits and incentives to create are not negatively impacted.
    4. Promoting economic growth by promoting businesses and innovation in India.
    5. Transparency and accountability by promoting disclosure requirements and three tierd dispute resolution.

II. POLICY RECOMMENDATIONS

Our policy framework, designed to balance the interests of all stakeholders in AI development, is structured into three distinct *yet* adaptable parts. Each part stands alone and meets different needs, but ideally, they function best when implemented together. However, recognizing the dynamic nature of AI technology, we've ensured that these components can be flexibly combined or modified to suit evolving needs and advancements in the field. This approach ensures our framework remains relevant, effective, and responsive to the rapidly evolving AI landscape. The three parts of the framework are:

1. Equitable Exemption

This intervention is based on the touchstone of principles of copyright law and jurisprudence laid down on it.

The following conditions must be met for the GenAI model to be allowed under this provision:

(a) the output must be of transformative nature
(b) it must develop and use output filtering mechanisms to prevent the creation of content that substantially replicates or derives from significant portions of the copyrighted work used for training;
(c) the avoidance of direct competition in the same market as the original copyrighted;
(d) the good faith by the user in not using Generative Artificial Intelligence to infringe upon existing copyrights; and

In light of the above, the following amendments to the Copyright Act 1957 are proposed:

(1) In sub-section (1)(a) of Section 52, after the clause (iii), the following clause shall be inserted:

> *"(iv) for the purpose of training or developing Generative Artificial Intelligence, provided such use adheres to the conditions specified in sub-section (1A)."*

(2) After sub-section (1) of Section 52, the following sub-section shall be inserted:

> *"(1A) Notwithstanding anything contained in sub-section (1), the fair dealing for the purposes of training or developing Generative Artificial Intelligence shall be subject to the following conditions: (a) transformative nature of output, significantly altering the original copyrighted work to create new and original content; (b) the development and employment of effective output filtering mechanisms to prevent the creation of content that directly replicates or extracts substantial portions of the copyrighted work used for training; (c) does not primarily result in Generative Artificial Intelligence outputs that directly compete with the original copyrighted work in its market or audience; (d) the good faith by the user."*

Further research required:

(i) It is difficult to draw a boundary between works which are or are not 'transformative.'

(ii) How to ensure that GenAI does not compete with original copyright work.

(iii) EE is a very high standard to reach. The practical significance of a standard like this is untested and therefore its effectiveness is highly uncertain.

2. Voluntary Licensing (VL) Framework

The VL Framework facilitates the use of copyrighted data for GenAI development through free and fair agreements between the parties through two distinct approaches:

2.1 Direct Licensing Between Developers and IP Holders:

This approach encourages direct agreements between GenAI developers and IP holders for accessing training data. The government's role in facilitating these direct licensing agreements is crucial. It can standardize agreements and provide neutral mediation services, offering support particularly for unintentional violations by innovators in GenAI. This approach aims to protect good-faith efforts and promote a culture of responsible innovation. To ensure balanced and responsible usage, it is proposed that the government must publish a sample dispute resolution clause and recommend it to be included in every licensing agreement.

A Sample Dispute Resolution Clause would include:

- Notice and Negotiation: In case of a license agreement violation, the Licensor issues a written notice to the Licensee for corrective action, with both parties engaging in good faith negotiations within forty-five (45) days.
- Grace Period for Model Correction: If a violation is related to the AI model's output, the Licensee gets a grace period of sixty (60) days to modify the model and implement effective filtering mechanisms.
- Termination: The Licensor may suspend or terminate the agreement if the Licensee fails to resolve the issue within the agreed timeframe.
- Arbitration: Unresolved disputes are settled through binding arbitration under the Arbitration and Conciliation Act, 1996.

More research is required to make an GenAI model un-learn a particular data source when the Licensor terminates the agreement. However, some research to facilitate that is already appearing.[11]

2.2 National Data Marketplace (NDM)

The NDM provides a streamlined platform for AI developers to access a wide array of data, facilitating innovation in AI technology. It simplifies the process of data acquisition and licensing by reducing administrative barriers. Simultaneously, it ensures fair compensation for IP holders, thereby encouraging them to share their data. The NDM's regulated environment ensures transparency and adherence to intellectual property laws, thus supporting both technological advancement and the protection of creators' rights.

1. **For IP holders:** IP holders can register their data (text, images, videos) on the NDM. Data is listed at creator-determined rates, and made easily discoverable through a searchable index. The NDM manages all licensing and transactions, ensuring secure and timely payment to creators.

2. **For Developers:** AI developers can browse and access a diverse range of data sets. Standard Licensing Agreements would be automatically executed between the parties when the AI developer pays for the data set, thereby enabling them to innovate at a rapid pace. A tiered pricing structure may be employed to ensure smaller entities and startups have equal access to data, preventing domination by larger companies.

3. **License Terms**: The standard agreement would have public policy considerations and restrictions against illegal use of data. It would also include a dispute resolution clause which would involve setting up a dispute resolution mechanism within the NDM, through an arbitration panel to settle disputes arising from NDM transactions.
4. **Regular monitoring:** Will be conducted every six months to adapt to technological advancements and market needs.
5. **Further research required:**
    - The NDM needs a mechanism to ensure the quality and reliability of the data being offered. This could involve standardizing data formats and establishing a vetting process for data before it's listed on the platform.
    - More data is needed before implementing a tiered pricing structure. Eventually with the growth of AI companies and licensing in India, it is expected that there will be enough data to devise a tiered pricing structure.

3. Compulsory Licensing (CL) Framework

This section presents a plan for mandatory licensing of copyrighted material used in GenAI training in India. This policy tool makes important training data more accessible, ensuring that smaller AI companies or startups have equitable access to essential data, which might otherwise be restricted due to high licensing costs or unavailability. In turn, the IP holders are also fairly compensated for their IP. The government will oversee this process, making it more uniform and less legally complicated.

1. Eligibility Criteria for AI Companies to be granted CL: AI companies must be legally registered entities in India. They must (i) not have turnover greater than 250 crores, (i)

[11] Ronen Eldan and Mark Russinovich, 'Who's Harry Potter? Approximate Unlearning in LLMs' (arXiv, 4 October 2023) <http://arxiv.org/abs/2310.02238> accessed 29 January 2024.

demonstrate a clear and ethical purpose for using the copyrighted material for AI development, and (iii) not have a history of copyright infringement or unethical use of data.

2. Application Process: AI companies would apply for compulsory licenses through an online portal managed by the Copyright Board of India (CBI). The application must detail (i) the specific copyrighted material needed, (ii) its intended use in AI training, and (iii) the expected outcome of the AI project.

3. CBI would assess the applications on the grounds of: (i) the potential societal or technological benefits of the AI project, (ii) the inability to obtain a voluntary license, and (iii) the impact on the copyright holder.

4. Compensation Mechanism: A standardized compensation formula would be established, considering factors including but not limited to (i) the nature of the copyrighted material, (ii) the scale of its use, and (iii) the potential commercial benefit. Payments could be managed through a centralized fund, ensuring transparency and timely remittance to copyright holders.

5. Rights of Copyright Holders:

    (i) Copyright holders would have the right to review applications concerning their material and raise objections if necessary.

    (ii) A clear and fair process for objections and appeals should be established, with the final decision made by the CBI.

The framework should be reviewed every six (6) months to check its effectiveness.

Further research required:

1. In figuring out the compensation mechanism. It could either be a flat fee or percentage based. Since there are very few instances of licensing agreements between AI developers and IP holders, it is unclear what compensation model would be agreeable to both the parties.
2. Protection of licensed data. It is hoped that protection of such data would be covered by the upcoming Digital Bharat Act.

III. CONCLUSION

Our proposed policy solution seeks to establish a harmonious framework that facilitates the training of Generative AI models using copyrighted data, striking a balance between fostering innovation and safeguarding creators' rights. The key components of this framework include the introduction of an Equitable Exemption tailored, the promotion of Voluntary Licensing through mutual agreements, and the implementation of Compulsory Licensing as a last resort, ensuring fair compensation for creators. To address ambiguity, we advocate for legal amendments that clearly define the boundaries of use. Additionally, our proposal calls for the establishment of a National Data Marketplace, fostering transparency in data transactions. This comprehensive approach aims to align technological advancements with intellectual property laws, creating a symbiotic environment that benefits all stakeholders in the dynamic realms of digital and creative economies.