Legal Playbook For Natural Language Processing Researchers

Last Updated: May 2022

Purpose of this Playbook

This playbook is a legal research resource for various activities related to data gathering, data governance, and disposition of an AI model available as a public resource. It aims to benefit academic and government researchers including those in New York State who wish to understand how best to use AI models to provide natural language processing ("NLP") as public infrastructure, but who do not have legal resources. The playbook aims to be a general informational resource to public organizations, including cross national organizations focused on non-commercial open science in NLP and promotion of the human rights to equal access to scientific advancement under UDHR Art. 27.

With this playbook, we strive to assist researchers who have less resources to help them guide their communities and their research, including low income communities who may not have access to legal resources. In particular, this playbook is cross jurisdictional, and hopefully will be relevant to NLP and data researchers in underserved language communities whose data will be processed (e.g., minority dual-language speakers) and those who wish to participate and have a stake in AI.

This playbook was drafted as part of the year-long BigScience workshop (https://bigscience.huggingface.co) and is not legal advice. This playbook is made available under the CC-BY license (https://creativecommons.org/licenses/by/4.0/). This playbook is not legal advice and is only intended as research of the legal landscape at or around the "Last Updated" date.

Organization

This playbook is organized by jurisdictions. In each jurisdiction, we provide an overview of the legal system, and then provide legal research for common questions that NLP researchers may have relating to intellectual property, licensing, privacy, text data mining and prohibited or restricted content and technologies.

Authorship & Contributions

The authors of the various playbooks and their executive summaries are named in their sections. The following are editors, reviewers and other contributors to this playbook: Huu Nguyen, Jen Dumas, Somaieh Nikpoor, Tara Davis, Danda Zhao, David Lansky, Stella Biderman, Jessie Dodge, and Yacine Jarnite.

Executive Summaries

Brazil Executive Summary

Under Brazilian law, training data sets have limited intellectual property protection in the form of copyright, covering only the expression of the structure or design of that dataset, and provided

that the content selection, organization and arrangement meets the requirements of an original and creative work in itself. To the extent that they are copyrightable, they are licensable. Unauthorized inclusion of copyrighted content in datasets is illegal.

Machine learning training models have no copyright or patent protection. However, software implementing training models, either in the form of source code or executable code, has copyright protection. What this means for NLP researchers is that while there may be some protections for the training data they create and their software, models may not be protectable, and thus control of the models after distributions may be challenging.

Brazilian laws include provisions that would prevent certain types of data from being text data mined, stored, distributed or generated, based or their content:

(i) Secret or top-secret information (state secret) and information protected from disclosure by legislation against anticompetitive practices may be prohibited from being text mined or distributed.

(ii) Child pornography, scene of rape and scene of sex, nudity or pornography that were made public without the consent of depicted person raise little concern regarding data mining, storage, distribution and generation of plain text data. The retrieval of textual metadata identifying the actual location of this type of content (e.g., URLs), or the inclusion of such metadata in the AI generated, may amount to a crime.

(iii) Conveyance of hate speech, defamation, promotion of terrorism and incitement or induction to suicide require special attention. Text data mining and storage of such content is not unlawful in itself, but distribution and generation can be.

Furthermore, national security reasons may impose restrictions on the distribution of data produced or held by the government and its agencies.

Colombia Executive Summary

Colombia has a civil law legal system that is governed, hierarchically, by the Constitution of Colombia, the laws enacted by the Congress of the Republic, the regulations issued by the government to implement and elaborate the content of the laws enacted by the Congress, and the courts' decisions issued by the judicial branch. In Colombia, artificial intelligence ("AI") and the use of data are regulated by civil, data protection, Intellectual Property ("IP"), and privacy laws and regulations, as well as other governmental policies and guidelines that allow companies and other organizations to legally develop and implement AI projects in the country.

Colombian data protection and privacy laws allow companies to collect, process, use and share personal data when the owner of such data has authorized such usage. Those laws and regulations require companies to implement a set of policies and procedures to guarantee the constitutional right of *habeas data* and privacy, and to avoid and protect data bases against data breaches. Additionally, Colombian laws prohibit the use of minor children's personal data, confidential information, reserved information, as well as the usage of any information accessed or used for criminal purposes or committing a crime.

Additionally, Colombian IP laws may be applicable to AI and ML projects and developments. When such works have enough creative expression to qualify as copyrighted work, they are likely to be protected under Colombian copyright laws as software. Similarly, when the creation meets the requirements to be considered an innovation under industrial property laws, such rules may apply. Furthermore, when the project or development is subject to trade secret, the IP laws governing such secrecy may apply. In any case, the use of any product or material that is protected under IP laws in Colombia requires authorization by the owner to be legally used, unless an exception applies.

Although Colombia does not have a specific regime applicable to AI and ML, the Colombian government has put in place several strategies aimed at promoting and encouraging the development of such projects in the country. First, the Colombian government has issued several policies regarding the use of bigdata and the application of privacy by design in AI projects in Colombia. Within these policies, the government has briefly included metadata as part of the governance policies that must be put in place to make use of big data in the country. Furthermore, the government has encouraged the creation and implementation of AI and ML projects in the country through its agencies. Particularly, the Superintendency of Finance and the Superintendency of Commerce and Industry have created spaces to develop AI and ML projects under the supervision and assistance of the regulator of the financial and commerce industries. These spaces are 'Innova SFC' and the *Sandbox on privacy by design and by default in Artificial Intelligence projects*, which consist of spaces for companies or organizations to develop AI or ML projects under the supervision and assistance of the governmental agency that oversees the activity.

This document explores the regulation of AI and ML under Colombian law by answering specific questions regarding IP, licensing, text data mining and fair use, and prohibited content.

Colombia Executive Summary in Spanish

Colombia tiene un ordenamiento jurídico de derecho civil que se rige, jerárquicamente, por las siguientes normas: (1) la Constitución Política de Colombia; (2) las leyes promulgadas por el Congreso de la República; (3) los decretos emitidos por el gobierno para implementar y desarrollar el contenido de las leyes promulgadas por el Congreso; y (4) las decisiones judiciales.

En Colombia, la inteligencia artificial ("IA") y los datos están regulados por diversas áreas del derecho y normas de diferente jerarquía. Las reglas que rigen la IA involucran la aplicación de leyes de protección de datos, Propiedad Intelectual ("PI") y de otros principios y políticas que rigen específicamente el mercado de la IA en Colombia. Todas estas normas se encuentran contenidas en distintas fuentes del derecho, pero principalmente en la Constitución Política de Colombia, las leyes específicas que rigen dichas ramas del derecho y las respectivas disposiciones reglamentarias emitidas por el gobierno para aplicar dichas leyes. Este documento identifica la ley aplicable en materia de IA en Colombia y responde preguntas específicas relacionadas con la aplicación de dichas leyes en casos particulares.

Primero, la regulación de la IA implica la aplicación de las leyes colombianas de privacidad. En Colombia, el hábeas data y la privacidad se rigen por los artículos 15 y 20 de la Constitución, la

Ley 1266 de 2008, la Ley 1581 de 2012, el Decreto 1377 de 2013 y el Decreto 1074 de 2015. Dichas normas establecen el marco general aplicable a la recolección y tratamiento de datos dentro de la jurisdicción colombiana. Asimismo, la sentencia C-748/11 emitida por la Corte Constitucional de Colombia constituye el caso hito en materia de derecho a la protección de datos en el país y, por lo tanto, se debe considerar en su aplicación.

Adicionalmente, la implementación de sistemas de IA en Colombia exige el cumplimiento de las leyes de PI, particularmente en relación con la protección del software por las leyes de derechos de autor, así como la celebración legal de contratos que tengan como objeto el producto de IA. La Ley 23 de 1982, la Decisión de la Comunidad Andina No. 351 de 1993, el Decreto 1360 de 1989 y la Decisión de la Comunidad Andina No. 486 de 2000 rigen el desarrollo de la IA en Colombia en el campo de la PI y establecen los requisitos legales para proteger AI bajo esta área del derecho.

Además, el gobierno colombiano ha establecido un conjunto de políticas aplicables a la IA y también ha suscrito acuerdos internacionales comprometiéndose a aplicar ciertos principios en materia de IA. Desde abril de 2018 se aprobó una política nacional de big data como parte de la cuarta revolución industrial, para promover el crecimiento económico y cumplir los objetivos digitales del Foro Económico Mundial, (2016). Esta política (CONPES 3920) se desarrolló con base en leyes y políticas anteriores sobre datos personales, datos abiertos y conjuntos de datos, así como también como parte de un esfuerzo de 'transparencia' por parte del gobierno para dar acceso abierto a todos los datos recopilados por sus entidades.

En esa misma línea, en noviembre de 2019, el Consejo Nacional de Política Económica y Social de la República de Colombia emitió un documento (CONPES 3975) que establece la política gubernamental aplicable a la innovación digital e IA en el país. Además, en 2019, Colombia adoptó los Principios de la OCDE sobre Inteligencia Artificial, que establecen una serie de principios para la administración responsable de una IA confiable y políticas nacionales y cooperación internacional para una IA confiable. Si bien estos documentos sólo incluyen políticas y principios generales, son herramientas útiles para comprender e interpretar ciertas definiciones y conceptos sobre la IA en Colombia.

Canada Executive Summary

Copyright can subsist in training datasets because they are either/both computer programs/software and compilations; however, the author(s) of the datasets do not acquire copyright in the works that comprise the datasets. Moreover, should any portion of the datasets comprise copyrighted works, the copyright owners' right to reproduce their respective works may be infringed and the dataset author(s) liable for that infringement. That being said, the defense of fair dealing may be raised to escape liability.

Whether publishing a language model trained on a subset of a published dataset will engender copyright infringement liability is a novel legal issue that does not have much in the way of supporting jurisprudence. That said, the *Copyright Act*'s provision regarding copyright owners' exclusive rights can be construed as suggesting that the publishing of a language model constitutes publishing a translation of the subset of the dataset on which the model was trained,

and that the actual use of the model constitutes performance of the dataset. To the extent that the foregoing proves to be the correct interpretation, the language model's publication may constitute copyright infringement in the dataset. However, the fair dealing defense may be available to escape liability.

Pursuant to the *Copyright Act*, both datasets and models are likely licensable. Accordingly, the license/terms of use for a publicly available dataset may impose restrictions and/or conditions on how the dataset may be used. It may, for instance, prohibit a dataset from being used to create commercial AI software. Therefore, it is crucial for the NLP researchers and AI engineers to devise and employ a systematic approach to verify a publicly available dataset's provenance, lineage and the rights/obligations (as set out in its license) to confirm that the dataset can in fact be used in the way intended.

Works that are not protected by copyright can be subject to text data mining, freely and without restrictions. For copyright-protected works, NLP researchers should look for text data mining policies, terms of use, etc. which explicitly grant the right to use the particular work for the purpose of text data mining to ensure that such usage of the work is permitted. Pursuant to such terms of agreement, NLP researchers are legally authorized by the copyright owners to text data mine the latters' works. However, absent such grant of authorization, NLP researchers may proceed to use the copyrighted works for the purposes of text data mining and then raise a fair dealing defense, should a copyright infringement claim be brought against them by the copyright owners.

Data gathered by text data mining can be released by means of a license; however, it is advisable that the terms of the dataset license be consistent with - or otherwise not contradict - those set out in the licenses of data sources in order to avoid potentially invalidating the dataset license. As for a model trained on the collected data and the training dataset, it can also likely be released via a license, provided that the terms of the license for the model is consistent with - or otherwise not contradict - the data source licenses and the dataset license.

It is difficult to assess whether an NLP researcher is subject to Canadian privacy laws (and any succeeding/amending legislation), as NLP researchers/AI engineers would not directly be soliciting personal information from individuals. Rather, they primarily intend to access and use web content or publicly available datasets that contain personal information that has been collected by a third party or that has been provided online voluntarily by the individuals to whom the personal information pertains. Moreover, the applicability of Canada's privacy statutes to an NLP researcher is contingent on the latter's use of personal information being deemed "in course of a commercial activity." Whether its use/work constitutes "commercial activity" is not clear.

China Executive Summary

Intellectual Property

Original data are protected by copyright law. Similarly, as a compilation of data, the original data training set will be protected by copyright law. However, the data training set can be protected as a trade secret. On the contrary, the data training model can be protected by copyright. The attribution of data-based rights should be discussed categorically: (1) personal data, (2) government data, and (3) corporate data.

Researchers may infringe the following intellectual property rights during data processing: first, publishing data sets containing HTML tags or document structures may involve violations of the right to disseminate works to the public through the information network, especially when these HTML tags insert pictures or hyperlinks in the text. Second, C4 or Oscar can be understood as computer programs. Therefore, when publishing a dataset that references a location in another dataset, even if the availability is limited, attention should be paid to the potential violation of the protected rights of computer software. Third, direct access to information from people may also infringe intellectual property rights. Depending on the different levels of input from the interviewer and the interview, the interview manuscript may be owned by the interviewee or jointly owned by the interviewer and the interviewee. The best way for the data collector to avoid any disputes in the future is to obtain the consent of the interviewee. Fourth, borrowed books are protected by copyright, so researchers using borrowed books to train language models may infringe intellectual property rights. Finally, China's existing copyright system cannot provide immunity for text data mining, so text data mining is likely to infringe copyright and other related property rights.

Licensing

Data licenses are agreements where the data owner authorizes others to use the data within a certain period and scope without changing the ownership. Since data ownership has not been settled down in law, the owner can be data subject or data processor according to the agreement between the data subject and the processor. Both datasets and models are licensable through contracts, and the restriction can be set as terms such as purpose of use, scope of use and exclusivity. Licenses are contracts, which is less binding than law. Thus, they may be invalid under the void or voidable conditions stated in laws. The relationship between licenses and terms of use is a little tricky. In circumstances where the data subject signed the terms of use with platforms which stipulates that the data can’t be shared, and then licensed third-party to use the data, the platform can claim breach of contract against the data subject. Generally speaking, NLP researchers can use licensed data to train their language model, and they also enjoy the copyright of the model as long as they state clearly on the license the terms related to use. Furthermore, licenses can be royalty free according to agreements.

Text Data Mining and Fair Use

Currently, there is no specific provision regulating data mining and data scraping. The recently enacted Data Security Law of PRC requires NLP researchers to comply with laws and regulations of China, respect social norms and ethics, and observe business and professional ethics when

collecting data. As the policy concern lies in intellectual property violation and the potential risk of intruding personal privacy, researchers should be aware of any relative violation in crawling data and abide by the IP laws and Personal Information Protection Law. In legal practice, courts generally apply Article 2 of Anti-Unfair Competition Law to regulate unlawful data mining behavior when it harms the legitimate rights and interests of other operators or that of consumers, when it violates the principle of good faith and recognized business ethics and when such behavior destroys the open and fair market competition order in the Internet environment. As social media platforms are another important stakeholder in this context, NLP researchers should note that most platforms, in its Terms of Use, do not allow data scraping of social media content without the consent of the operators even when the user gives his permit already. Moreover, by reading into the courts' opinion in Sina Weibo v. Maimai, courts tend to require triple authorization in this context. Generally, researchers will be insured in a safety belt if it obtains authorization of use from stakeholders before starting the data mining activities.

Privacy

The legal concern around accessing web content and datasets that have PII are mainly provided in the PRC Personal Information Protection Law ("PIPL"). The PIPL does not distinct data controller or data processor. Rather, it uses the term of "Personal Information Handler", which has the same meaning as "Data Controller" under the GDPR. In general, when handling personal information, individual consent in the primary legal basis. In case consents cannot be obtained, researchers may consider another legal basis specified in Article 13.6 of the PIPL, which permits processing of "*the personal information that has been disclosed by the individuals themselves, or other personal information that has been legally disclosed within a reasonable scope in accordance with this Law*." Unlike GDPR, legitimate interest is not a valid legal basis under the PIPL.

Providing PII to other processors or making PII public by any means requires the handler to notify the data subject and obtain the separate consent from the individual. Such separate consent shall be given by individuals under the precondition of full knowledge, and in a voluntary and explicit statement.

According to Article 6 of the PIPL, personal information processing shall have a clear and reasonable purpose, and shall be directly related to the processing purpose, using a method with the smallest influence on individual rights and interests. The collection of personal information shall be limited to the smallest scope for realizing the processing purpose, and excessive personal information collection is prohibited.

The application scope of PIPL is extra-territorial. Apart from all the processing activities happening in China, the activities happening outside China are subject to the PIPL if it provides products or services to people within China or analyzes their behaviors. Similar to GDPR, the offshore controller shall establish special institutions or designate representatives within China to handle affairs relating to personal information protection, and submit the names of relevant institutions or the names and contact information of representatives to the regulatory authorities. (Article 53).

For special protection of children's personal information, apart from the general protection offered in other laws and regulations, such as the Civil Code and the Personal Information Protection Law, there is a special regulation called the Provisions on the Cyber Protection of Children's Personal Information, which states that a network operator collecting, using, transferring or disclosing any child's personal information shall notify the child's guardian in a conspicuous and clear manner, and obtain verified consent from the child's guardian for the collection, use, transfer or disclosure of personal information of the child.

The applicable remedy to infringement of personal information rights include administrative punishment, civil liability and criminal liability. Maximum fine can go up to 50 million yuan or 5% of revenue.

Prohibited Content

Although the legislation in data security is a relatively new area in China and the legal framework is just established with a series of newly adopted laws and regulations in recent years, there are some restrictions on handling data (including collection, storage, distribution of data, data mining, etc.) that need to be addressed. First, data handling shall not violate laws and regulations, social morality and ethics, business ethics, and professional ethics. Second, data exportation is subject to the State's control and shall follow certain measures formulated by the State, especially information related to personal identity, national security and critical infrastructure. Third, the laws provide more protection to minors by imposing more requirements of handling minors-related data. Fourth, terms and conditions of the license contract are the key to the control of licensing use of data sets and models. Last, the newly issued Ethics Specifications of the New Generation of AI sets out some ethical requirements on conducts related to AI.

South Korea Executive Summary

Training data and models can be protected by copyrighted works of the Copyright Act in South Korea; data training sets and models express human thoughts and emotions and have a creative nature. Moreover, if text data mining does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, it would fall under the 'Fair Use' of Copyright Act in South Korea.

Importantly for NLP researchers, the current Copyright Act allows non-contractual use when it falls under the fair use doctrine, but there are arguments that the Copyright Act should be revised to make a separate provision for data mining that can be possible without the consent of the right holder.

Even so, to access web content that has PII (Personal Identifiable Information) legally, it is required to comply with the Personal Information Protection Act of South Korea. Of note, NLP researchers should be aware that pseudonymization is one of the best options to prevent personal information infringement.

If distributing models causes providing PII to third parties, a consent of the data subject is required in South Korea under Personal Information Protection Act in principle.

Circulation of data which contain obscene content, defame other persons, harmful to youths under the Youth Protection Act, and speculative activities etc. may be prohibited. Also, distributing, or providing child or youth sexual exploitation materials for commercial purposes, or possessing, transporting child or youth sexual exploitation materials is prohibited.

Of particular interest to NLP researchers, South Korea has one of the few cases where an NLP application and company has been sanctioned by the government. On April 28, 2021, the Personal Information Protection Commission of South Korea("PIPC") judged that the Scatter Lab Corporation's act of using KakaoTalk conversations training data during the development and operation of an AI chatbot service "Iruda" violated the Personal Information Protection Act. PIPC imposed penalty surcharges of 55,500,000 won and administrative fines of 47,800,000 won on Scatter Lab.

Japan Executive Summary

There is no definition of data ownership in Japan's legal (hard law) regime. Since IPR law does not always govern mostly data, datasets, and metadata, most data does not automatically have licensing default rules. Therefore, it is crucial to clarify the terms of use and licensing conditions in an agreement. In case that the data, datasets, or metadata is also trade secrets, one should also set the terms and conditions which disclose such data to be protective enough to protect the confidentiality of the data.

As models mostly are works of programming language, the models should be licensable under the default rule in the Copyright Act. Similar to data, the model may be occasionally inventive enough to have a patent registered or can be protected as trade secrets. In that case, one should also pay attention to the licensing conditions in the Patent Act or the measures to keep the model confidential so that it could be a trade secret under the Unfair Competition Prevention Act. If the pre-trained language model produces training data to be seen as-is, it might results in copyright infringement if the data is creatively expressed (Article 30-4 of Copyright Act)

Text data mining, without violating the relevant laws, is not restricted in Japan. The use of the copyrighted work which aims not to have its expression perceived and does not conflict with the interest of the copyright owner is allowed under the Copyright Act. The Copyright Act altogether allows the use of copyrighted works for machine learning with no restriction to the academic or non-commercial use.

In Japan, with respect to model generating, and storing content and with respect to distribution of data and models, there is no restriction on the generation of data per se, unless it is a) generated with an illegal means (see C-1) or b) child pornography (Violation of Article Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children (児童買春、児童ポルノに係る行為等の規制及び処罰並びに児童の保護等に関する法律)). There is no restriction on storing data per se, unless the above-mentioned two situation (collected through illegal means, or child pornography). Distribution of the information could be a violation of multiple laws including the following: privacy , Information used for WMDs, missiles and conventional weapon development and production, Trade Secret or licensed information, or child pornography.

France Executive Summary

Intellectual Property

France has adopted two types of protections for databases in the French Code of Intellectual Property. A database has been defined as “a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means”.
First, databases can be protected through traditional copyrights law, provided that the author is at the origin of an original intellectual creation. Such originality should lie in the architectural structure of the database
Second, databases can be protected through a sui generis database right. The producer of a database benefits from a protection of the content of the database, provided that such producer made a substantial financial, material, or human investment in the constitution, verification, and presentation of the content of the database.

French law introduces exceptions and limitations to copyrights protections. If the rightsholder has divulged the work, then he may not prohibit:

- Analyses and short citations
- Temporary reproductions with a transitory or accessory character
- Reproductions and presentations of a work for conservation purposes
- ***Digital copies or reproductions text and data mining purposes***

French law also introduces exceptions and limitations to database right protections. If a rightsholder has made the database publicly available, then it may not prohibit:

- Extractions or reuse of a non substantial part of the content of the database
- Extractions and reuse of a substantial part for illustrative, research purposes
- Extractions or reuse of a database for conservation purposes
- ***Digital extractions, copies, or reproductions text and data mining purposes***

La France a adopté deux régimes de protection des bases de données dans le Code de la propriété intellectuelle. Une base de données a été définie comme "recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen".
Premièrement, une base de données peut être protégée par un droit d'auteur traditionnel, à condition que l'auteur soit à l'origine d'une création intellectuelle originale. Cette originalité réside généralement dans la structure architecturale de la base de données.
Deuxièmement, une base de données peut être protégée par un droit sui generis. Le producteur d'une base de données bénéficie d'une protection du contenu de la base, à condition que ce producteur ait fait un investissement financier, matériel ou humain substantiel dans la constitution, la vérification et la présentation du contenu de la base.

La loi introduit des exceptions et des limites à la protection des droits d'auteur. Si le titulaire du droit a divulgué l'œuvre, il ne peut pas interdire :

- *Les analyses et les citations courtes*
- *Les reproductions temporaires ayant un caractère transitoire ou accessoire*

- *Les reproductions et présentations de l'œuvre à des fins de conservation*
- *Les copies ou reproductions numériques à des fins de fouille de textes et de données.*

La loi introduit également des exceptions et limites à la protection des droits sui generis sur les bases de données. Si un titulaire de droits a mis la base de données à la disposition du public, il ne peut pas interdire :

- *Les extractions ou la réutilisation d'une partie non substantielle du contenu de la base de données*
- *Les extractions et la réutilisation d'une partie substantielle à des fins d'illustration ou de recherche*
- *Les extractions ou la réutilisation d'une base de données à des fins de conservation.*
- *Les extractions, copies ou reproductions numériques à des fins d'exploration de texte et de données.*

Licensing

Holders of a copyright over a software have the right to authorise another to effectuate (1) a permanent or temporary reproduction, (2) a translation, adaptation, arrangement, or modification of the software and the reproduction of such modified software, (3) the presentation to the market, gratuitously or at cost, of the exemplaries of the software. More broadly authors of copyrighted material are entitled to put their work at the gratuitous disposal of the public.

Holders of a database right have the right to subject the extraction and reuse of protected databases to license agreements.

However, unlike for patents and brands, France does not expressly recognise licenses for copyrights. We could therefore only suppose that a contract could be drafted based on generally applicable contract law principles.

Moreover, regarding public information, France has adopted special provisions present in the French Code of Relations between the Public and the Administration.

Les titulaires d'un droit d'auteur sur un logiciel ont le droit d'autoriser un tiers à effectuer (1) une reproduction permanente ou temporaire, (2) une traduction, une adaptation, un arrangement ou une modification du logiciel et la reproduction de ce logiciel modifié, (3) la présentation sur le marché, gratuitement ou à titre onéreux, des exemplaires du logiciel. Plus généralement, les auteurs d'œuvres protégées par des droits d'auteur ont le droit de mettre leur travail à la disposition du public à titre gratuit.

Les titulaires d'un droit sur une base de données ont le droit de soumettre l'extraction et la réutilisation des bases de données protégées à concessions de licence.
Cependant, contrairement à ce qui est le cas pour les brevets et les marques, la France ne reconnaît pas expressément les concessions de licence en matière de droits d'auteur. On ne peut donc que supposer qu'un contrat pourrait être rédigé sur la base des principes du droit commun des contrats.

Par ailleurs, en ce qui concerne les informations publiques, la France a adopté des dispositions spéciales dans le Code français des relations entre le public et l'administration.

Text Data Mining

France has adopted provisions that expressly regulated text and data mining. Provisions on work protected by copyright and on databases protected by a sui generis database right prohibit the rightsholder who has made his work or database publicly available to prevent its extraction, copy, and reproduction in a digital form for the purpose of text and data mining.

Text and data mining is defined as the “implementation of an automated analysis technique of texts and data in a digital form in order to extract information, including patterns, trends and correlations”. Text and data mining that does not require the consent of the rightsholder is limited to mining executed for scientific research purposes by research organisms and other public actors.

Additionally, if a database is protected by a copyright and is made publicly available by its rightsholder, the latter may not prohibit the extraction and re-use of the database made for research or private-study purposes by private individuals, in specific settings, and for no economic or commercial purpose.

A 2019 EU Directive has introduced mandatory provisions that need to be adopted by all Member states and that provide for text and data mining, either for scientific or more general purposes. However, most countries have failed to implement such provisions and are subject to legal proceedings by European institutions.

La France a adopté des dispositions qui réglementent expressément la fouille de textes et de données. Les dispositions relatives aux œuvres protégées par des droits d'auteur et aux bases de données protégées par un droit sui generis sur les bases de données interdisent à l'ayant droit qui a mis son œuvre ou sa base de données à la disposition du public d'en empêcher l'extraction, la copie et la reproduction sous forme numérique à des fins de fouille de textes et de données.

La fouille de textes et de données est définie comme la "mise en œuvre d'une technique d'analyse automatisée de textes et données sous forme numérique afin d'en dégager des informations, notamment des constantes, des tendances et des corrélations". La fouille de textes et de données qui ne nécessite pas le consentement de l'ayant droit est limitée aux fouilles effectuées à des fins de recherche scientifique par des organismes de recherche et d'autres acteurs publics.

En outre, si une base de données est protégée par un droit d'auteur et est mise à la disposition du public par son titulaire de droits, ce dernier ne peut pas interdire l'extraction et la réutilisation de la base de données effectuées à des fins de recherche ou d'étude privée par des particuliers, dans des cadres spécifiques, et sans but économique ou commercial.

Une directive européenne de 2019 a introduit des dispositions obligatoires qui doivent être adoptées par tous les États membres et qui prévoient la fouille de textes et de données, à des

fins scientifiques ou plus générales. Cependant, la plupart des pays n'ont pas mis en œuvre ces dispositions et font l'objet de procédures judiciaires de la part des institutions européennes.

Privacy

France adopted the EU's General Data Protection Regulation (GDPR) through an amendment to its its law n°78-17 (Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés). GDPR provisions became applicable in France from May 25, 2018.

The GDPR, which has the particularity of being extraterritorial (Article 3), created specific obligations surrounding the collecting and processing of personal data. Processing of personal data is only lawful under certain circumstances, including (among other conditions, defined by the GDPR Article 6-1) if consent of the data subject was obtained; if the processing is necessary for a contract; or if it was necessary for a task carried out in the public interest. The GDPR also specified that certain categories of personal data containing sensitive information (such as race, religious beliefs, or sexual orientation) are, by default, prohibited for processing. Only under certain exceptions laid out in GDPR's Article 9-2, including consent of the data subject, can personal data containing sensitive information be lawfully processed.

Data must be processed in a fair and transparent manner, for specific purposes, and for no longer than what is necessary (GDPR Article 5). The GDPR also provided that data gathered in the public interest or as part of research must abide by safeguards (Article 89), following the principle of data minimization and including pseudonymization when possible.

Consent of the data subject must be requested using clear and plain language, and must be easily withdrawable at any time (GDPR Article 7). When collecting data directly from individuals, the controller must provide the information listed in the GDPR's Article 13, including the existence of a right to withdraw consent and the right to be forgotten.

One specificity of the French adaptation of the GDPR to national law is the age of consent for personal data processing (without parental authorization); it is 16 for the GDPR (Article 8), and 15 for French law (*Loi n° 78-17 du 6 janvier 1978,* Article 45).

La France a adopté le Règlement général sur la protection des données (RGPD) de l'UE par le biais d'un amendement à sa loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Les dispositions du RGPD sont devenues applicables en France à compter du 25 mai 2018.

Le RGPD, qui a la particularité d'avoir une compétence extraterritoriale (article 3), a créé des obligations spécifiques autour de la collecte et du traitement des données personnelles. Le traitement des données à caractère personnel n'est licite que dans certaines circonstances, notamment (entre autres conditions, définies par l'article 6-1 du RGPD), si le consentement de la personne concernée a été obtenu ; si le traitement est nécessaire à un contrat ; ou s'il était nécessaire à une tâche effectuée dans l'intérêt public. Le RGPD a également précisé que certaines catégories de données personnelles contenant des informations sensibles (telles que

l'origine ethnique, les croyances religieuses ou l'orientation sexuelle) sont, par défaut, interdites de traitement. Ce n'est que dans le cadre de certaines exceptions énoncées à l'article 9-2 du RGPD, notamment le consentement de la personne concernée, que les données personnelles contenant des informations sensibles peuvent être traitées.

Les données doivent être traitées de manière loyale et transparente, pour des finalités déterminées et pour une durée n'excédant pas celle nécessaire (article 5 du RGPD). Le RGPD a également prévu que les données collectées dans l'intérêt public ou dans le cadre de la recherche doivent respecter des garanties spécifiques (article 89), en suivant le principe de minimisation des données et en incluant la pseudonymisation lorsque cela est possible.

Le consentement de la personne concernée doit être demandé en utilisant un langage clair et simple, et doit pouvoir être facilement retiré à tout moment (article 7 du RGPD). Lorsqu'il collecte des données directement auprès des personnes, le responsable du traitement doit fournir les informations énumérées à l'article 13 du RGPD, notamment l'existence d'un droit de retrait du consentement et le droit à l'oubli.

Une spécificité de l'adaptation française du RGPD au droit national est l'âge de consentement pour le traitement de données personnelles (sans autorisation parentale) ; il est de 16 ans pour le RGPD (article 8), et de 15 ans pour le droit français (Loi n° 78-17 du 6 janvier 1978, article 45).

Prohibited Content

The French law n°2004-575 (*loi n°2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique*) defines prohibited content in its Article 6-I-7. It includes content which negates crimes against humanity; content which encourages terrorism; content which encourages hate of individuals because of their race, sex, sexual orientation, gender identity or disability; child pornography; content encouraging violence (specifically including sexual and gender-based violence); and content which degrades human dignity. One aspect of the law is that service providers have to create reporting mechanisms to be made aware of prohibited content on their platforms (Article 6-I-7).

French law n°2004-575 was an adoption into national law of the European Union Directive on electronic commerce (*Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000*), which differentiates liability for content which is hosted, and content which is edited by a service provider. A service provider is not liable for content that it simply hosted on their platform, provided that they did not know that the content was illegal; or if they knew, that they deleted it immediately (*loi n°2004-575,* Article 6-I-2). On the other hand, if the content was edited by the service provider (if they had an active role in it), then they would be liable — following the French law on freedom of press, in which editors are liable for the content they publish (*loi du 29 juillet 1881 sur la liberté de la presse*, Article 42).

Content threatening national security is under specific scrutiny, as the European Parliament adopted its *Regulation 2021/784 on the dissemination of terrorist content online* on April 29, 2021. This regulation gives member states the power to issue removal orders for terrorist content, and

creates an obligation for service providers to remove the content within one hour of the order's issuance (Article 3). Service providers also have to submit a yearly transparency report concerning the mechanisms they put in place to prevent the spread of terrorist content (Article 7). However, content which was disseminated for research, educational, artistic or journalistic purposes is not considered as terrorist content by this Regulation (Article 1-3).

La loi n°2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique définit les types de contenus illicites dans son article 6-I-7. Il s'agit notamment des contenus qui nient les crimes contre l'humanité, des contenus qui provoquent la commission d'actes terroristes, des contenus qui incitent à la haine raciale, à la haine des individus en raison de leur sexe, de leur orientation sexuelle, de leur identité de genre ou de leur handicap, de la pornographie enfantine, des contenus qui incitent à la violence (notamment aux violences sexuelles et sexistes) et des contenus qui portent atteinte à la dignité humaine. L'une des dispositions de la loi est que les fournisseurs de services doivent créer des mécanismes de signalement pour être informés des contenus illicites qui pourraient se trouver sur leurs plateformes (article 6-I-7).

La loi française n°2004-575 a été une adoption en droit national de la directive de l'Union européenne sur le commerce électronique (directive 2000/31/CE du Parlement européen et du Conseil du 8 juin 2000), qui différencie la responsabilité pour le contenu qui est hébergé, du contenu qui est édité par un prestataire de services. Un prestataire de services n'est pas responsable d'un contenu qu'il a simplement hébergé sur sa plateforme, à condition qu'il n'ait pas su que ce contenu était illicite ; ou s'il l'a su, à condition de l'avoir supprimé immédiatement (loi n°2004-575, article 6-I-2). En revanche, si le contenu a été édité par le prestataire (s'il a joué un rôle actif), il sera responsable - conformément à la loi sur la liberté de la presse, selon laquelle les éditeurs sont responsables du contenu qu'ils publient (loi du 29 juillet 1881 sur la liberté de la presse, article 42).

Les contenus menaçant la sécurité nationale font l'objet d'une attention particulière, le Parlement européen ayant adopté le 29 avril 2021 son règlement 2021/784 contre la diffusion du terrorisme en ligne. Ce règlement donne aux États membres le pouvoir d'émettre des injonctions de retrait pour des contenus à caractère terroriste, et oblige les fournisseurs à retirer ces contenus dans l'heure qui suit (article 3). Les fournisseurs de services doivent également soumettre un rapport annuel de transparence concernant les mécanismes qu'ils ont mis en place pour empêcher la diffusion de contenus terroristes (article 7). Toutefois, les contenus diffusés à des fins éducatives, journalistiques, artistiques ou de recherche, ne sont pas considérés comme des contenus terroristes par ce règlement (article 1-3).

Belgium Executive Summary

In Belgium, an EU jurisdiction, most regulation that affects NLP research and researchers stems from the European Union. In anticipation of the newly proposed EU Artificial Intelligence Regulation, no specific legislation exists in this field in Belgium.

Intellectual property is an important challenge for NLP research. Datasets can be protected both by copyright (but only if the result of human intervention demonstrating the necessary degree of originality) and the *sui generis* database right. However, the existence of intellectual property rights does not rule out their use for NLP purposes per se: exceptions to such rights exist. Notably, the scientific research exception. If the scientific research exception to these intellectual property rights does not apply, the NLP researchers will need a license to use the protected data.

Licensing of data protected by intellectual property rights is permitted and governed by general contract law. Note however that licensing of copyrighted works requires the observance of more stringent legal conditions.

As soon as Belgium implements the Directive on copyright in the Single Digital Market, text data mining of copyrighted works will be explicitly permitted, for scientific research purposes only on the condition of lawful access, and in general upon lawful access and consent by the rightsholder.

Privacy of data subjects is governed by GDPR. GDPR imposes very strict standards on the use of personal data. It is essential that consent of the subject data is obtained if NLP processes such personal data. In the context of NLP research, article 89 provides for an exemption and more lenient standard, including with respect to the purpose limitation of the use of personal data.

Belgium - In French

En Belgique, une juridiction de l'UE, la plupart des réglementations qui affectent la recherche et les chercheurs en TALN proviennent de l'Union européenne. En prévision de la nouvelle proposition de règlement européen sur l'intelligence artificielle, aucune législation spécifique n'existe dans ce domaine en Belgique.

La propriété intellectuelle est un défi important pour la recherche en traitement automatique des langues. Les ensembles de données peuvent être protégés à la fois par le droit d'auteur (mais seulement si le résultat d'une intervention humaine démontre le degré d'originalité nécessaire) et par le droit sui generis des bases de données. Cependant, l'existence de droits de propriété intellectuelle n'exclut pas en soi leur utilisation à des fins de TALN : il existe des exceptions à ces droits. Notamment l'exception de recherche scientifique. Si l'exception de recherche scientifique à ces droits de propriété intellectuelle ne s'applique pas, les chercheurs TALN auront besoin d'une licence pour utiliser les données protégées.

La concession de licences sur des données protégées par des droits de propriété intellectuelle est autorisée et régie par le droit général des contrats. Notez toutefois que la concession de licences pour des œuvres protégées par le droit d'auteur nécessite le respect de conditions juridiques plus strictes.

Dès que la Belgique aura transposé la directive sur le droit d'auteur dans le marché numérique unique, l'exploration de données textuelles d'œuvres protégées par le droit d'auteur sera explicitement autorisée, à des fins de recherche scientifique uniquement à la condition d'un accès légal, et en général à la condition d'un accès légal et du consentement du titulaire des droits.

La vie privée des personnes concernées est régie par le GDPR. Le GDPR impose des normes très strictes sur l'utilisation des données personnelles. Il est essentiel que le consentement des données de la personne concernée soit obtenu si le PNL traite ces données personnelles. Dans le cadre de la recherche, l'article 89 prévoit une exemption et une norme plus clémente, notamment en ce qui concerne la limitation de la finalité de l'utilisation des données personnelles.

Switzerland Executive Summary

Despite its geographical location in the heart of Europe, Switzerland is not a member state of the European Union ("EU"). Hence, EU law is not directly applicable in Switzerland. Yet still, the indirect influences (e.g., via bilateral treaties) of EU law even on purely domestic issues are manifold. Further, many Swiss statutes are voluntarily adapted to, or at least inspired by, the respective regulations applicable in the EU.

Swiss intellectual property law may restrict NLP research. Datasets can for example be protected by copyright if they qualify as works under Art. 2 of the Swiss Copyrights Act. Unlike in the EU, however, there is no *sui generis* database right. The existence of intellectual property rights does not *per se* rule out their use for NLP purposes, as some exceptions exist. Notably, copyright protected data can be used for purely scientific research (Art. 24d CopA). If the scientific research exception to these intellectual property rights does not apply, the NLP researchers will need a license to use the protected data.

Licensing of data protected by intellectual property rights is permitted and governed by general contract law. There are no specific regulations on licensing agreements in the Swiss Code of Obligations. There are no specific form requirements for licensing agreements.

Text data mining of copyrighted works is explicitly permitted for scientific research purposes pursuant to the revised Art. 24d CopA, if the copyrighted material is accessed lawfully.

Swiss data privacy law is not governed by the EU GDPR, but by several other provisions on the federal and cantonal level. The strict standards on the use of personal data, however, are in practice very similar. According to Art. 13(2)(e) FADP, an overriding interest of the person processing the private data may justify the otherwise unlawful use if that person processes personal data for purposes not relating to a specific person, in particular for the purposes of research, planning and statistics and publishes the results in such a manner that the data subjects may not be identified.

South Africa Executive Summary

Intellectual Property

South Africa's Copyright Act 98 of 1978 and its amendments cover most of the IP issues. Under the Copyright Act, the "computer programs" are eligible for copyrights which benefit the data training sets and models for NLP researchers. Such copyrights need not to be registered and will

be effective immediately after programs are written; however, the contents and materials that the NLP researchers obtain to train the models might also be protected by its own copyright, which might cause the copyright conflicts. Thus, when a company publishes a pre-trained language model that is trained on the subset of a published dataset based on a crawl of the web in terms of IP, it is important to obtain the consent from the author of such web page content, especially for commercial use. In addition, copyright requires the “original" works from the model creators. Therefore, publishing datasets that contain plain text would be more vulnerable to copyright challenge than publishing datasets with additional information. On the other hand, when the dataset only refers to locations in another dataset, it is similar to an index, which is less likely to infringe other’s copyrights and could be a fair use if it is for research purposes. Be aware that when collecting personal information directly from persons through interviewing, it may trigger the protections under Protection of Personal Information Act (POPIA) as discussed in more detail below in the Privacy summary. Generally, POPIA prohibits the processing of special personal information such as religious beliefs; philosophical beliefs; race; ethnic origin; trade union membership; political persuasion; health; sex life; biometric information or other criminal behaviors and it requires e data training system to establish appropriate safeguards to protect such data. Moreover, if the NLP researchers train a language model on a borrowed book, it is safer to use the borrowed library books after the 50-year copyright duration expires or perhaps consider to use the materials in the open-source archive following the license agreement of the data pool.

Prohibited content

In addition to intellectual property and information privacy laws, legal constraints on data that can be (i) text mined, (ii) lawfully stored or distributed, and (iii) generated by AI can be found in the Constitution of the Republic of South Africa, the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, the Children’s Act, the Criminal Law (Sexual Offences and Related Matters) Amendment Act, 2007.

In light of the applicable legal framework, activities of data processing should be aware of prohibitions on (i) disclosure of national security information, (ii) child pornography, (iii) conveyance of hate speech, promotion of terrorism, unfair discrimination, war, and incitement of imminent violence.

Licensing

There is no fixed list on the types of IP that can be licensed in South Africa. Models, databases and their contents may qualify for copyright protection and be licensed through either an exclusive or a non-exclusive license agreement. Parties are free to negotiate the terms of the license for the whole or part of the bundle of the relevant IP rights, with jurisdictional restrictions and time limits. In addition, formal requirements apply for the execution of exclusive copyright licenses. The right to use copyrighted or licensed material can be transferred through agreements between the owner of the work and third parties.

Text Data Mining

Currently, in South Africa there are no specific exceptions governing the use of text data mining. In fact, in spite of being a member of WTO and Berne Convention, South Africa did not adopt any TDM exceptions and the three-step test contained in Article 13 TRIPS.

Moreover, in South Africa there are no laws or regulations governing web crawling. Therefore, to the extent that there are no specific regulations explicitly prohibiting web crawling, as long as such activity does not violate privacy, IP, or criminal laws, NLP researchers could legally crawl the web themselves.

Also, South Africa does not have a special regime applicable to social media platforms. Therefore, there is no definition of social media content or any provision that allows individuals to exactly determine what can be considered social media content, provided that individuals may be in any case protected by the constitutional right to privacy.

Privacy

The Protection of Personal Information Act ("POPIA") allows companies to process personal information when the owner of such data has consented and been notified of the purpose of collection and the responsible party. Personal information should be collected directly from data subjects, public records such as public webs or datasets, or another source agreed to by data subjects. The laws and regulations require companies to implement a set of policies and procedures to protect data. Additionally, South African laws provide stricter requirements on special personal information such as religious beliefs, race origin, biometric information, etc. Activities solely for research purposes may satisfy the public interest requirement and receive an exemption from the Information Regulator concerning further processing, notification for data subjects, prohibitions on processing special personal information, and other requirements under POPIA

[Detailed Playbooks Below]

BRAZIL

Rodrigo Canalli

Introduction

Brazil's Constitution provides for a federative republic with a democratic system of government limited by formal and material constitutional constraints, such as fundamental rights, separation of powers, rule of law and judicial review. Notwithstanding the federative organization, most of the relevant laws related to IP, data protection, privacy, licensing and content regulation are enacted by the federal government (the Union). The Brazilian legal system is largely civil law. It means that, as a general rule, statutes have precedence over case law. However, decisions of the Supreme Federal Court on constitutional matters delivered though specific procedures are binding to lower courts.

In April 2021, the federal government introduced the Brazilian Artificial Intelligence Strategy - BAIS (Ordinance n. 4617/2021 of the Ministry of Science, Technology and Innovation), which aims to guide state actions toward the promotion of research and development of artificial intelligence solutions and establish ethical standards for AI development and governance. Among its main goals is to avoid regulatory actions that needlessly limit innovation, adoption and development of AI. The document recognizes the transversality of AI impact on society, extending through sectors as different as scientific research, education, public governance, labor relations, entrepreneurship, and security. Especial concern with adherence to ethical principles in AI development and use are highlighted, notably regarding the need to avoid algorithmic biases that may create or reinforce prejudices. The Brazilian Strategy was designed in alignment with the OCDE principles for responsible AI development: (i) inclusive growth, sustainable development and well-being, (ii) human-centered value and fairness; (iii) transparency and explainability, (iv) robustness, security and safety, and (v) accountability. Among the proposed guidelines, the following are particularly relevant to researchers working with natural language processing technologies:

(a) Encouragement of ethical AI through financing of research projects dealing with equity/non-discrimination, accountability and transparency.

(b) Encouragement of partnerships between the government and the private sector.

(c) Identification and elimination of algorithmic bias.

(d) Elaboration of a data quality control policy for AI training systems.

(e) Mapping of legal and regulatory barriers to AI development as to identify legislation in need of update.

(f) Promotion of sandboxes and regulatory hubs.

(g) Creation of an artificial intelligence observatory in Brazil.

(h) Support to the use of representative datasets to train and test models.

(i) Facilitation of access to open government data.

(j) Improvement in quality of the available data, facilitating the detection and the correction of algorithmic biases.

(k) Encouragement of the dissemination of open-source codes capable of verifying discriminatory trends in data sets and machine learning models.

List of relevant legislation

Law n. 9279/1996 - Authors' Rights Law

Law n. 9609/1998 - Computer Software Intellectual Property Protection Law

Law n. 9610/1998 - Industrial Property Law

Law n. 11484/2007 - Integrated Circuit Topography Law

Law n. 12414/2011 – Positive Credit Reporting Law

Decree n. 8771/2016 – Federal Government Open Data Policy

Ordinance n. 46/2016 of the Information Technology Secretariat – Brazilian Public Software Policy

Law n. 13709/2018 – General Personal Data Protection Law

Ordinance n. 4617/2021 of the Ministry of Science, Technology and Innovation – Brazilian Artificial Intelligence Strategy - BAIS

A. IP Questions

1. Are the training data sets and models protected by IP rights and if so which IP rights?

Issue: whether IP laws protection extends to data training sets and models and what IP rights would apply.

Rules

As a member of WTO, Brazil has amended its IP legislation to conform with the TRIPS Agreement. Main IP related statutes are: (i) the Authors' Rights Law (Law n. 9610/1998), which covers copyrights and moral rights on creative works of authorship, (ii) the Computer Software Intellectual Property Protection Law (Law n. 9609/1998), which specifies the application of copyright protection to computer software, (iii) the Industrial Property Law (Law n. 9279/1996), which provides for patent law, industrial design law, trademark law, geographical indications, and anti-competitive practices, and (iv) the Integrated Circuit Topography Law (Law n. 11484/2007, chapter 3), which regulates the intellectual property of layout designs (topography) of integrated circuits.

Analysis

Machine-learning algorithms are trained, validated and tested with input from datasets. Training datasets are the sets of data from which the algorithm "learns". In Brazil, particular software expressions are protected by copyright (Law n. 9.610/1998, article 7, XII) either in the form of source code or executable code (Lei n. 9609/1998, article). However, because the Authors' Rights Act (Law n. 9610/1998) provides that "ideas, systems, methods, projects or mathematical concepts" are not protected by copyright (article 8, I), the protection does not extend to the algorithms and models they implement.

Regarding datasets, under article 7, XIII and § 2, of the Authors' Rights Act (Law n. 9610/1998), copyright protection of "collections or compilations, anthologies, encyclopedias, dictionaries, databases and works alike" is limited to expression that is original and creative in the "selection, organization or arrangement of their content" and "does not cover the data or materials themselves." Therefore, the owner of the copyright on a database has the exclusive right regarding only "the form of expression of the structure of that database."

Furthermore, the inclusion of content protected by copyright in datasets and its storage require previous and express authorization from the copyright holder (Law n. 9610/1998, article 29, IX).

Patent protection does not extend to "discoveries, scientific theories and mathematical methods" (Law n. 9279/1996, article 10, I), "purely abstract conceptions" (Law n. 9279/1996, article 10, II), nor "computer software in itself" (Law n. 9279/1996, article 10, V). It is usually understood as to deny patent protection to software, while allowing for patentability of specific implementations of hardware and software. To the extent that training models may be understood as educational materials, it is noteworthy that Brazilian legislation also prohibits patenting of "educational … schemes, plans, principles or methods" (Law n. 9279/1996, article 10, III).

Conclusion

Under Brazilian law, training data sets have limited intellectual property protection in the form of copyright, covering only the expression of the structure or design of that dataset, and provided that the content selection, organization and arrangement meets the requirements of an original and creative work in itself.

Unauthorized inclusion of copyrighted content in datasets is illegal.

Machine learning training models have no copyright or patent protection. However, software implementing training models, either in the form of source code or executable code, has copyright protection.

2. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?

As noted above, under Brazilian law the inclusion of copyrighted content in datasets and its storage require previous and express authorization from the copyright holder (Law n. 9610/1998, article 29, IX). This makes web crawled datasets vulnerable to copyright infringement claims since they copy protected content from various websites.

More generally, under Law n. 9610/1998, article 29, any total or partial reproduction, editing, adaptation, transformation, translation into any language and distribution, among other uses, require previous and express authorization from the author. The right of the copyright owner is infringed whenever a protected work is the object of unauthorized or fraudulent reproduction, disclosure, use, editing, sale, acquisition, storage, broadcasting etc.

Because it is the dataset, and not the model, that reproduces the content, it is unclear whether publishing a language model that was trained on such a dataset would suffer from the same vulnerability. Since statutory law does not provide for a specific guidance, legal status of language models trained on web crawled data sets remains uncertain. Non-commercial and scientific purposes may be considered as a positive factor in eventual controversy, but there is no express exception in statutory law. Furthermore, if the ubiquitous use of web crawling by search engines serves as a parameter, language models would have no trouble using such datasets for training.

Conclusion

Legal uncertainty prevents reaching a conclusive answer about such legal concerns.

B. Licensing Questions

1. Are data sets licensable? Are models licensable?

Under the article 7, XIII and § 2, of the Authors' Rights Act (Law n. 9610/1998), datasets are copyrightable. The copyright protection over datasets, however, is limited to expression that is original and creative in the "selection, organization or arrangement of their content" and "does not cover the data or materials themselves." The owner of the copyright on a database holds exclusive rights of use regarding "the form of expression of the structure of that database", including the right to license the use of the copyrightable elements of a dataset to third-parties (Law n. 9610/1998, articles 49 and 87).

The copyright on the expression of a database structure does not preclude existing copyright on the data contained therein, which also can be licensed by the respective owner (Law n. 9610/1998, article 49), however.

Software is protected by copyright (Law n. 9.610/1998, article 7, XII) and licensable (Lei n. 9609/1998, article 9), either in the form of source code or executable code. However, because the Authors' Rights Act (Law n. 9610/1998) provides that "ideas, systems, methods, projects or mathematical concepts" are not protected by copyright (article 8, I), the protection does not extend to the algorithms and models they implement.

Copyleft and free software licenses are generally admitted and encouraged by policy makers and technology community in Brazil.

As discussed above, patent protection does not extend to "discoveries, scientific theories and mathematical methods" (Law n. 9279/1996, article 10, I), "purely abstract conceptions" (Law n. 9279/1996, article 10, II), nor "computer software in itself" (Law n. 9279/1996, article 10, V). This

means that patent protection is not available to software in itself, but patentability of specific implementations of hardware and software is not excluded. To the extent that training models may be understood as educational materials, it is noteworthy that Brazilian legislation also prohibits patenting of "educational … schemes, plans, principles or methods" (Law n. 9279/1996, article 10, III).

Conclusion

Under Brazilian law, datasets have limited intellectual property protection in the form of copyright, covering only the expression of the structure or design of that dataset, and provided that the content selection, organization and arrangement meets the requirements of an original and creative work in itself. To the extent that they are copyrightable, they are licensable.
The content of datasets can be the object of parallel licensing by the respective copyright owner. Although software is subject to copyright, algorithmic models are not. Patent protection also does not extend to either software or algorithms.

E. Prohibited content

1. What types of data may be prohibited from being (i) text data mined, (ii) stored or distributed, or (iii) generated?

This section analyzes the presence of legal constraints other than those related to intellectual property and information privacy laws in the researched jurisdiction. The issue is whether the jurisdiction provides for content-based prohibitions restrictions on (i) data that can be text mined for the purpose of training machine learning algorithms, (ii) data that can be lawfully stored or distributed, and (iii) the data generated by AI.

Rules

The Brazilian constitution offers sound protection to freedom of expression, but not without some caveat. It provides for freedom of expression of thought, but forbids anonymity (article 5, IV) and assures the right of reply to offenses, as well as compensation for pecuniary or moral damages to reputation (article 5, V). The expression of intellectual, artistic, scientific, and communication activities is free, and not subjected to censorship or license (article 5, X). Notably, the constitution establishes access to information as a constitutional right (article 5, XIV).

The law punishes with imprisonment the delivery to a foreign government, its agents, or a foreign criminal organization, of documents or information classified as secret or top secret when such disclosure could endanger the preservation of the constitutional order or national sovereignty. Anyone who helps a spy incurs the same penalty. However, it is not a crime if the act has the purpose of exposing crimes or human rights violations (Law n. 14197/2021, article 2).

Undisclosed information submitted to government authorities for approval of a regulated product's marketing (e.g., test results) is protected. The unauthorized use of such information can be regarded as anticompetitive practice (Law n. 10603/2002).

Producing, reproducing, directing, photographing, filming or recording, by any means, explicit sex or pornographic scene involving children or adolescents (meaning any person under 18 years old) are crimes punishable with imprisonment and fine. The acts of possessing, storing, making available, transmitting, distributing, publishing or disseminating such materials are also crimes punishable with time in prison and fine. The penalties are extensible to those who provide the means or services for the storage of child pornography. The prohibition includes simulated child pornography (Law n. 11829/2008).

Obscenity is not prohibited in itself. However, even if no underaged is involved, the law punishes with imprisonment and fine those who make available, transmit, distribute, publish or disseminate, by any means, photography or video depicting a scene of rape, or a scene of sex, nudity or pornography, without the depicted persons' consent (Law n. 13718/2018).

Hate speech, understood as speech inducing or inciting discrimination based on race, color, ethnicity, religion or national origin, is criminalized. Commerce, distribution or conveyance of the swastika symbol, with the purpose of spreading Nazism, is expressly forbidden (Law n. 7716/1989). In 2019, the Supreme Federal Court decided that the hate speech legislation can be applied to cases involving homophobic and transphobic speech (STF, ADO 23 and MI 4733).

Brazil is among the jurisdictions where there are criminal defamation laws. The Penal Code provides for three degrees of defamation: injury (lowest penalty), defamation and calumny (highest penalty). Promotion of terrorism or of a terrorist organization is a crime punished with imprisonment and fine (Law n. 13260/2016). Incitement or induction to suicide is also a crime.

Analysis

In light of the laws in place, activities of data processing should be aware of prohibitions on (i) disclosure of secret or top-secret information (state secret) or information protected from anticompetitive practices, (ii) child pornography, dissemination of scenes of rape and unconsented dissemination of scenes of sex, nudity or pornography, (iii) conveyance of hate speech, defamation, promotion of terrorism and incitement or induction to suicide.

(i) Text mining of secret or top secret classified information, even for the purpose of training machine learning algorithms, may be held a crime against national sovereignty. The same can be applied to those who help someone to have access to such classified information by supplying the technological tools. Although the law does not prohibit the storage of this kind of data, its distribution is a crime, unless it is done with the purpose of exposing crimes or human rights violations.

Unauthorized text data mining and distribution of undisclosed information submitted to government authorities for approval of a regulated product's marketing (e.g., test results) can be prohibited as anticompetitive practice. Because they are related to government held data, these prohibitions have no application to data generated by AI.

(ii) Brazilian definition of child pornography as explicit sex or pornographic scene involving children or adolescents, includes (i) any situation that involves a child or adolescent in explicit

sexual activities, real or simulated, or (ii) the display of a child or adolescent's genitals for primarily sexual purposes.

Child pornography legislation is aimed at the prohibition of images, photographs, films or recordings depicting an underaged person in sexual activities. Likewise, the legislation prohibiting dissemination of scenes of rape and unconsented dissemination of scenes of sex, nudity or pornography does not concern textual content. Therefore, they do not extend to texts, even if their content involves textual description of child pornography or rape. Violation of these laws by text data mining activity is unlikely. In fact, regarding plain text data mining, storage, distribution and generation is out of the scope of criminal legislation on child pornography, dissemination of scenes of rape and unconsented dissemination of scenes of sex, nudity or pornography. However, the retrieval of textual metadata identifying the actual location of such content (e.g., URLs) or inclusion of such meta data in the AI generated content may incur in distribution, publishing or dissemination of child pornography, and the design of mechanisms to avoid this outcome are advised.

(iii) Hate speech, defamation, promotion of terrorism and incitement or induction to suicide are often conveyed in the form of texts, therefore requiring special attention when it comes to text data mining, distribution and generation. Although the text mining of data containing any of these contents, or their storage, would not be unlawful in themselves, the use of such data in algorithm training might lead to further generation and distribution of data comprising hate speech, someone's defamation, promotion of terrorism and incitement to suicide. If the content generated by AI amounts to any of these, it would likely be unlawful.

A model trained on such data could eventually be deemed useless or unlawful, depending on the degree of the content generated by it is irremediably tainted by prohibited content.

Conclusion

Brazilian laws include provisions that would prevent certain types of data from being text data mined, stored, distributed or generated, based or their content:

(i) Secret or top-secret information (state secret) and information protected from disclosure by legislation against anticompetitive practices may be prohibited from being text mined or distributed.

(ii) Child pornography, scene of rape and scene of sex, nudity or pornography that were made public without the consent of depicted person raise little concern regarding data mining, storage, distribution and generation of plain text data. The retrieval of textual metadata identifying the actual location of this type of content (e.g., URLs), or the inclusion of such metadata in the AI generated, may amount to a crime.

(iii) Conveyance of hate speech, defamation, promotion of terrorism and incitement or induction to suicide require special attention. Text data mining and storage of such content is not unlawful in itself, but distribution and generation can be.

3. Is there a legal restriction on the distribution of data for national security reasons?

As examined above, there are restrictions on the distribution of data for national security reasons. Text mining of secret or top secret classified information, even for the purpose of training machine learning algorithms, may be held a crime against national sovereignty. The same can be applied to those who help someone to have access to such classified information by supplying the technological tools. Although the law does not prohibit the storage of this kind of data, its distribution is a crime, unless it is done with the purpose of exposing crimes or human rights violations. Because it is related to government held data, the prohibition has no application to the data generated by AI.

Therefore, national security reasons may impose restrictions on the distribution of data produced or held by the government and its agencies.

COLOMBIA

Daniel Molano and Maria Castillo

Introduction

Colombia has civil law legal system that is governed, hierarchically, by the following norms: (1) the Constitution of Colombia; (2) the laws enacted by the Congress; (3) the regulations issued by the government to implement and elaborate the content of the laws enacted by the Congress; and (4) judicial decisions.

In Colombia, artificial intelligence ("AI") and data are regulated by various areas of law and rules of different hierarchies. The rules governing AI involve the application of privacy law, Intellectual Property ("IP"), and other principles and policies that specifically govern the AI market in Colombia. All of these rules are contained in different sources of law, but primarily in the Colombian Constitution, the specific statutes governing such areas of law, and the respective regulatory provisions issued by the government to apply the statutes. This document identifies the applicable law to AI in Colombia, and answers specific questions related to the application of such laws in particular cases.

First, the regulation of AI involves the application of Colombian privacy laws. In Colombia, habeas data and privacy are governed by articles 15 and 20 of the Constitution, the Law 1266 of 2008, the Law 1581 of 2012, the Decree 1377 of 2013, and the Decree 1074 of 2015. Those rules establish the general framework applicable to data collection and treatment within the Colombian jurisdiction. Also, the decision C-748/11 issued by the Constitutional Court of Colombia constitutes the landmark case regarding privacy law in the country and must be considered in application of privacy law.

Additionally, the implementation of AI systems in Colombia demands compliance with IP laws, particularly in connection with the protection of software by copyright laws, as well as to legally execute contracts that have the AI product as an object. The Law 23 of 1982, the Andean Community Decision No. 351 of 1993, the Decree 1360 of 1989, and the Andean Community Decision No. 486 of 2000 govern the development of AI in Colombia within the field of IP and provide the legal requirements to protect AI under that area of law.

Furthermore, the Colombian government has established a set of policies applicable to AI and has also subscribed international agreements committing to apply certain principles regarding AI. Since april 2018 a big data national policy was approved as part of an industrial revolution, to promote economic growth and to accomplish digital objectives from the World Economic Forum, 2016 . This policy (CONPES 3920) was developed based on previous laws and policies regarding personal data, open data and data sets as well as a part of an effort of 'transparency' by the government to give open access to all the data compiled by its entities as a general rule. In November 2019, the National Council of Economic and Social Policy of the Republic of Colombia issued a document (CONPES 3975) setting the governmental policy applicable to digital innovation and AI in the country. Further, in 2019, Colombia adopted the OECD Principles on Artificial Intelligence, that set forth a series of principles for responsible stewardship of trustworthy AI and national policies and international co-operation for trustworthy AI. Although these documents only include general policies and principles, they are useful tools to grasp certain definitions and concepts regarding AI in Colombia.

Finally, Colombian agencies have put in place a set of spaces for companies to develop AI projects under the supervision of the authorities. That is the case of 'Innova SFC,' a space created by the Superintendency of Finance that allows Fintech projects to be developed with the assistance and supervision of the regulator. Similarly, the Superintendency of Industry and Commerce recently created the Sandbox on privacy by design and by default in Artificial Intelligence projects to promote the development of innovative AI projects under the supervision and assistance of the supervisory agency.

For the purposes of this document, the following are the most relevant applicable rules:

- The Constitution of Colombia
- Law 23 of 1982 – Copyright Law
- Law 1266 of 2008 – Financial Data Protection Law
- Law 1581 of 2012 – Data Protection Law
- Law 1621 of 2013 - Intelligence and Counterintelligence Law
- Law 1712 of 2014 - Access to Public Information Law
- Resolution 1519 of 2020 - Standards for the Use of Public Information Minister of Technology.
- Andean Community Decision No. 351 of 1993 – Copyright Law of the Andean Community
- Andean Community Decision No. 486 of 2000 – Industrial Property Law of the Andean Community
- Decree 1360 of 1989 - Regulation of Software under Copyright Law
- Decree 1377 of 2013 - Regulation of Data Protection Law
- Decree 1074 of 2015 - Compilatory Decree of the Commercial
- CONPES 3920 - National Policy of Data Usage (BigData)
- CONPES 3975 - National Policy of Digital Transformation and Artificial Intelligence

A. IP Questions

1. Are the data training sets and models protected by IP rights and if so which IP rights?

If the model has enough creative expression to qualify as a copyrightable work, data training sets are likely to be considered software under Colombian copyright laws and thus be protected by Intellectual Property ("IP") rights.

In Colombia, copyright is primarily governed by Law 23 of 1982 and the Andean Community Decision No. 351 of 1993. According to article 2 of Law 23 of 1982, any literary work created by a natural person that is original and could be fixed in paper or be reproduced through any other known or to be known means of distribution is protected by copyright. Further, the Decree 1360 of 1989 regulates the application of Law 23 of 1982 to software. Article 1 of such Decree establishes that software is considered a literary work under Law 23 of 1982, and thus is protected by copyright laws when it falls within the definition set forth in articles 2 and 3 of the Decree. Similarly, in accordance with article 4(l) of the Andean Community Decision No. 351 of 1993,

computer programs are protected by copyright if they fall within the definition contained in article 3(20) of such Decision.

Data sets are likely to be considered software or part of software according to the definitions set forth in articles 2 and 3 of the Decree 1360 of 1989 and article 3(20) of the Andean Community Decision No. 351 of 1993. Article 2 of the Decree establishes that software is integrated by any or all of the following three components: computer program, description of the program, and supplementary or auxiliary material. Data training sets are likely to be considered computer programs under the definition of article 3(a) of the Decree that provides that computer programs are "the expression of an organized set of instructions, in natural or codified language, regardless of the means in which they are stored, whose objective is to make a machine capable of processing information, indicating, performing, or obtaining a function, a task or a specific result." Thus, to the extent that datasets are part of a system that permits to perform any of these tasks, they are likely to be considered software.

Similarly, data training sets could also be deemed software under article 3(20) of the Andean Community Decision No. 351 of 1993 that defines computer programs or software as "[…] the expression in words, codes, plans or any other form of a set of instructions which, on being incorporated in automated reading apparatus, is capable of causing a computer an electronic or similar device capable of processing information to execute a particular task or produce a particular result."

Although protection of copyright law begins when the software is created, the developer shall register such work at the National Directorate of Copyright -DNDA- in order to give publicity to the software and validly celebrate agreements whose objective is the transfer or distribution of the software rights.

Additionally, data training sets and models may be protected through trade secrets or business secrets. Article 260 of the Andean Community Decision No. 486 of 2000 defines business secret as "[a]ny undisclosed information that a natural person or legal entity legitimately holds, that may be used in any productive, industrial or commercial activity and that is capable of being passed on to a third party […]." According to that article, such information will be considered business secret if: (a) is secret; (b) has a commercial value due to its secrecy; and (c) was kept in secrecy by its legitimate holder through the usage of the reasonable measures to maintain confidentiality.

In conclusion, to the extent that data training sets or models have the necessary creative expression to qualify as a copyrightable work, they can be protected as software under Colombian copyright laws and can be negotiated through agreements when they have been previously registered at the DNDA.

2. **Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?**

Colombian regulation provides that copyright protection is automatic from the creation, which applies to scientific creations, and if registered in the DNDA[1] it is to declare but no to create the protection. Therefore, the owner of the information compiled in the data set is who compiled the information (except for personal data) and depending on the contract with whom might be the final recipient of the information contended in the data set. Not all the information content in data sets is declared as protected.

In general terms, data sets understood as software under the definition done by the Andean Community of Nations, if they are sufficiently original organized, are protected and copyrightable. When protected (since its creation) to modify it a license with transference rights is necessary.

However if what is being used is a data set from a public institution, those data sets have open licensing and can be used without any problem.

3. **Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g. HTML tags or document structure)?**

Assuming that the datasets are copyrightable because they have enough creative expression to be protected under copyright laws, there are no legal differences between publishing datasets that contain plain text only and publishing those with additional information. As previously mentioned, when data sets are copyrightable because they can be considered software under the Colombian regulations, the protection covers both source code and object code, regardless of the additional information that could be contained in the datasets. Consequently, when they are copyrightable, datasets are protected even if they contain additional information such as that included in HTML tags or document structures.

4. **What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?**

There are no specific rules governing the publication of metadata in Colombia. However, this kind of publication may raise concerns regarding privacy and IP laws. Assuming that the metadata could be linked or associated to any determined or to be determined natural person, it will be considered personal data and thus be governed by Law 1581 of 2012 and the Decree 1377 of 2013. Additionally, if the metadata is part of a copyrightable work, the person publishing the information must comply with the Colombian IP laws, that require authorization from the creator of the work to make any publication of such work.

In addition, the Colombian government has briefly included metadata as part of its policy regarding usage of big data. The Colombian *National Policy Of Data Usage (BigData)*, set forth in the document CONPES 3920 issued by the National Council of Economic and Social Policy of the Republic of Colombia, includes metadata as part of the governance policies that must be put in place to make use of big data. However, this document does not provide any information

[1] Copyright National Dictetorate

regarding the publication of metadata subtracted from a dataset and only includes the description of data usage as part of the governance policy necessary to use big data.

In conclusion, the main legal concerns regarding the publication of metadata subtracted from datasets are those related to privacy and IP laws.

5. **Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR/Pile) (for example, such a dataset may contain information like: in the nth entry of C4, there is a "<b>" html tag after the mth character)?**

What is being protected by Colombian regulation is the way information is organized under the software definition. Therefore, publishing datasets that only refers to other datasets locations but not the information itself or its organization is permitted. Always that the location information is not substantially relevant to the data set it can be used. Appropriation of resources might be an infringement of copyright. When one element is being taken out and transforming it into another context. Article 5 of Law 23 of 1982 requires authorization to transform. Therefore, who transforms the element is the owner of the new product.

6. **What are the IP risks related to data collection directly from persons? For example when you interview people or they donate data etc.**

In terms of direct data collection of information, it is necessary to differentiate from personal data, regulated by the Statutory Law of Personal Data (1581/2012). In these cases, the owner of this information is the person itself and when compelling this kind of information its use must be advised to the person sharing it (the information) and express authorization must be obtained, this authorization can be revoked anytime and must be removed from any source of information immediately. Also must be advised to everybody using that information that it cannot be sued and needs to be removed from the database.

When collecting other kinds of information, it depends on the contract about the use of the information (usually called terms and conditions). Therefore if the information that is being directly collected is for exclusive use, the owner is who compiles the information and if sufficiently originally organized, it is copyrightable.

7. **Can NLP researchers train a language model on borrowed library books – e.g., from the Internet Archive or other online book lender?**

Most likely, no. Books are literary works protected by copyright and their content cannot be reproduced or shared without authorization of the author. Article 2 of Law 23 of 1982 and article 4(a) of the Andean Community Decision No. 351 of 1993 expressly protect artistic and literary works expressed in writing as a copyrightable work. In consequence, unless the work is in the public domain, NLP researchers cannot train language models on borrowed library books.

Books and literary works that belong to the public domain could be used to train a language model under the Colombian IP laws. In Colombia, article 21 of the Law 23 of 1982 protects the economic rights of the author until their death and 80 years after such an event. After that period, the work enters the public domain, and although the moral rights of the work remain to the author, the economic rights can be freely exploited without their authorization. Therefore, a language model could be trained with books consulted from the Internet Archive or other book lender as long as (1) the book belongs to public domain, in the terms set forth in article 187 of the Law 23 of 1982; and (2) the way in which the library provides the information its server is not itself protected by copyright law. In any other case, the authorization of the author would be necessary to train the model based on the books.

Moreover, NLP researchers cannot train language models on borrowed library books to the extent it would appear that none of the fair use exceptions in Colombia apply to such cases. In Colombia, copyrighted works can be used without the authorization of the author in a set of specific cases. Among them, there are four exceptions that allow the usage of literary works for academic or teaching purposes. Those include (1) the reproduction and communication of literary or artistic works; (2) the reproduction of articles or extracts of literary works through reprographic means; (3) the communication or interpretation of works; and (4) the reproduction or setting of conferences and lessons. Such exceptions are limited, and only apply to non-for-profit activities and for academic or teaching purposes. Therefore, it would appear that such exceptions do not cover the training of a language model with information on borrowed library books, and thus such training would not be permissible in those cases.

Additionally, a language model trained on borrowed library books could be considered a derivative work if it has enough creative expression to be copyrightable, as provided in article 8(j) of Law 23 of 1982. According to this latter rule, a derivative work is the one that derives from the adaptation, translation, or any transformation of the original work, as long as it constitutes a new independent creation. In such a case, article 5(a) of the cited law requires the adaptation, translation or transformation of the original work to be authorized by the author, if it does not belong to the public domain. Therefore, even when the derivative work is copyrightable, the NLP researchers need authorization from the author.

In conclusion, unless the information used to train the language model is contained in books that belong to the public domain and are no longer protected by the copyright laws regarding the economic rights of the work, NLP researchers cannot train a language model on borrowed books from a library without the explicit authorization of the author.

8. How do rights on the source data (e.g. copyright) transfer to the trained model? Will this depend on where training occurs? Will this depend on where data is gathered?

Assuming that the source data is or has copyrightable material, the most typical way to legally transfer the rights would be through (1) contracts of assignment of copyrighted material; or (2) license agreements. These agreements (A) allow third parties to use the copyrighted material under certain conditions, in the case of licenses; or (B) transfer the ownership of the economical rights of the copyrighted work. Thus, the relevant aspect to consider when transferring data to the trained model is not where such transfer occurred, but rather if there was a valid

agreement to make such transfer. Similarly, the need to have a valid agreement to use the copyrighted material is independent of where data is gathered.

B. Licensing Questions

1. **Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?**

Whether data sets and models are licelceable and what are the limitations to face depends on who is the owner of the information in line with copyright regulation and if this is public regulated information which people have the right to access to, what is protected is not the date itself (underlined data) but how data sets are organized. If they are **sufficiently original in their organization,** they can be subject to copyright protection.

In general terms, data sets from governmental entities are public with open licensing. Therefore, there is no limitation on its use or access unless a special regulation dictates otherwise which is the case on personal data where certain types of information are not accessible by everybody. This protected information includes names, identification numbers, race and sexual orientation among others.

Date sets from private persons, can be licenced and its licenses are available during the same time period copyrights are protected over them. However, licensing regulation depends mostly on privacy law depending on contract terms which only requisite is that licenses must be written according to article 20 of law 23 of 1982.

Licenses can be exclusive, not exclusive, for a limited time or with transformation possibilities (among others) which requires a perpetual permission for future developments always depending on contractual terms.

Under this interpretation, data sets and models are licenceable as scientific creations but protection will only be valid according to specific private contracts.

2. **What are the relationships between laws, licenses, and terms of use? Which is more binding?**

In Colombia, laws regulate and set forth the general requirements for licenses to be binding and valid under Colombian law, licenses constitute private binding agreements between the parties that participate in the execution of them, and terms of use are specific agreements that persons collecting, processing, and sharing personal data must celebrate with those providing the information.

Laws are more binding than licenses or terms of use, but they usually contain general rules on how to legally engage in certain activities. Licenses and terms of use are agreements that are only binding to those who have executed them, and exceptionally to third parties. While licenses tend to be broad agreements typically designed to permit and/or transfer the usage of

intangible assets from one person to another, terms of use are specific agreements in which a person accessing to a digital portal authorizes the way in which the portal provider manages the portal and the information collected from the public through such portal.

In Colombia, terms of use are often the means to obtain authorization to collect, process, and share personal data in compliance with article 5 of the Decree 1377 of 2013 that requires such authorization by the owner of the data. Therefore, such terms tend to be set forth by the owner of the website and are not negotiated between the parties.

Notwithstanding the above, since there is no legal definition of terms of use, such terms could include agreements similar to those of a license to use a copyrighted material.

3. **What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?**

Under Colombian law, a license is valid as long as it complies with the general requirements of any contract under the Colombian contractual regulation. Article 1502 of the Colombian Civil Code requires (1) legal capacity; (2) valid consent; (3) licit consideration and object; and (4) compliance with formalities for a contract to be valid. In the case of licenses whose object is copyrightable dataset, the parties shall celebrate the contract over the work's patrimonial right and not over the moral rights of the work, that cannot be transferred in any way.

Assuming that NLP researchers have access to the license, they can determine its applicability to a re-published or derived dataset by accessing the terms of the license and reviewing the terms of the contract. If the dataset is copyrightable, it cannot be used without the express authorization of the creator of the work. Such authorization is typically granted through licenses of different kinds. Typically, copyrightable work that is available to any NLP researcher to use with no limitation is accessible through a creative commons license. Works that have these kinds of licenses are identified with a symbol that permits access to the terms and scope of the contract. Meanwhile, if there is not such license, the NLP researcher must try to identify the kind of license or agreement that governs the distribution of the work, because such a contract will be the one determining whether any license applies to any re-published or derivative dataset.

In addition, there may be datasets that are publicly available to NLP researchers and that do not require a license to be used. Public data in Colombia is regulated by the Law 1712 of 2014, that obligates governmental agencies to make publicly available certain information collected and processed by such agencies. Such data can be used by NLP researchers under certain limitations provided by the cited law and by the open licenses or terms of use established by each institution.

In conclusion, when the dataset is a copyrightable work, NLP researchers can know the scope of the license by accessing it, keeping in mind that any derivative work would require authorization by the creator of the work, unless a legal exception applies. Conversely, when the information accessed by the researcher is open data, the way to use such data will come from the law and the licenses or terms of use that each governmental agency establishes to grant access to the information.

4. What about if the users download or copy their own data and then provide it to NLP researchers directly?

As explained before in the matter of personal data, statutory law 1581 of 2012 requires that in matters of handling personal data there is an express authorization on its use, in these cases the owner of the information and owner of the data is the person, who can know, update and verify their personal data as well as revoke the authorization of its use at any time.

When it comes to personal data, as it is owned by the person, he may use it in the way he prefers (it cannot be sold) without the need to obtain authorization for the handling of his own data.

In the case of data that is not defined as personal data, the owner of the information is the one who collected it and therefore, it is necessary to refer to the agreement entered into with the person who provided the information. If the destination of the information was exclusive, then the owner will not be able to download the data and deliver it to anyone since there would be a copyright violation, if the information was not collected with exclusive use or through confidentiality agreements, they could be downloaded and shared but only those in which that person is the resource

5. Does the license that the dataset is shared under override the terms and conditions?

Under Colombian privacy and IP law, licenses do not necessarily override the terms and conditions applicable to a data set. As previously mentioned, licenses are different from terms and conditions. While licenses are binding agreements to transfer the use of a set of intangible assets over certain work, terms of use and conditions are agreements in which the owner of a digital portal establishes the way in which they will manage the portal and the information collected from the public through it. Moreover, establishing terms of use or a policy describing the use of the data that is collected through a website or a mobile application is a legal obligation set forth in article 13 of the Decree 1377 of 2013, which is applicable regardless of any licensing that may govern the use of the data set. However, to the extent that licenses can be freely negotiated between the parties, it would be possible to include covenants regarding the obligation to apply certain terms of use or to adopt a new set of terms to any person accessing the portal. Nevertheless, regardless of the content of the license agreement, any person collecting, processing, using or sharing the personal data must put in place a policy indicating the terms of use of such information.

Additionally, licenses normally regulate the relationship between the owner of the intangible asset whose usage is being transferred and the person receiving such asset, while the terms of use govern the relationship between those accessing the digital platform and the one providing such platform, regardless if the latter is the owner or the licensee. In consequence, the license under which the dataset is shared will coexist with the terms of use applicable to those entering the portal or digital platform.

In conclusion, terms of use are agreements that must be adopted and structured to legally collect, process, use, and share personal data, and they cannot be completely overridden by licenses, although such licenses may include covenants obligating the parties to modify or amend certain provisions of such terms.

6. Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?

Most likely yes, but not necessarily. Access to licensed or copyrighted material is often limited. The right to use copyrighted or licensed material can be transferred through agreements between the owner of the work and third parties. In Colombia, such kinds of agreements are normally (1) contracts of assignment of copyrighted material; or (2) license agreements. Those contracts will be the ones determining whether it is possible for NLP researchers to use the licensed or copyrighted material to train a language model.

Contracts of assignment of copyrighted material are regulated by article 183 of Law 23 of 1982, as it was amended by article 30 of the Law 1450 of 2011. Through this contract, the owner of a copyrighted work transfers the economic rights of its creation to a third party. According to the cited article, to execute this kind of agreement the parties are required to put the contract in writing, to indicate the rights that are being transferred, and to register the contract at the DNDA. To the extent that the norm does not impose any further restrictions on the contract, the parties are free to set the terms of the usage. Moreover, if no limitations are included in the agreement, the person that receives the rights over the work can use it without any limitations. Therefore, if there are no explicit limitations included in the contract, the NLP researchers can legally train a language model with the data. Furthermore, in this agreement the owner of the copyrighted work transfers the rights over the work and is no longer entitled to exercise such rights unless a limitation is set forth in the contract.

License agreements, also called ‘authorizations to use’, allow third parties to access copyrighted material with the authorization of the owner. Contrary to the contract of assignment, where the rights are ceded to the acquiring party, the license agreement only allows third parties to use the copyrighted work for free or in exchange for a fee. In fact, the DNDA issued the Concept 1-2016-66819 of 2016, in which it clarified that while the contract of transfer or assignment transferred some or all the rights of the work to the acquiror, the license agreement did not. Nevertheless, if this latter agreement does not include any particular limitations to the usage of the work, the NLP researcher could give any reasonable use to such material, such as using the content of it to train new models.

In those cases where the terms of the agreement do not restrict the ways in which the copyrighted work can be used, the language model trained by NLP researchers could be considered a derivative work, as defined by article 8(j) of Law 23 of 1982. Assuming that such work has enough creative expression to be copyrightable, the language model trained by the NLP researcher would be protected by copyright law as a separate work, according to article 5 of the Law 23 of 1982. In consequence, the derivative work created by NLP researchers can be

published and distributed without limitation. They will be the owners of all the legal rights attached to such derivative work.

7. Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?

There can be problems about text data mining when the information used after the mining is protected by copyrights such as books or articles. Since these pieces are protected, cannot be reproduced or used without authorization therefore data is not licensable because it is being transformed and authorization is needed to be the owner of the product or new [software]piece and being able to license it.

If the model or data sets are original enough, they can be licensed and therefore it could be licensed under a royalty free license if no work from another person is being used or if it was authorized to be transferred. The requirements for originality are low, and will likely be met by the model or data.

C. Text Data Mining and Fair Use Questions

1. What are types of legally permitted text data mining?

In Colombia there are no specific regulations governing the use of text data mining. This lack of regulation renders applicable the general rule of law set forth in the opinion C-893/03 of the Constitutional Court of Colombia, according to which private parties can act freely and without any limitation as long as their conduct is not prohibited by the Constitution or any other laws.

In the case of text data mining, the limitations may come from different sources of law. First, the usage of large sets of data requires compliance with privacy laws. These laws limit the collection, storage, access, and sharing of the information to persons that have expressly been authorized by the rightsholder of the data or by the law to handle such information. Also, even when authorization has been granted, the person using the databases must give special treatment to sensitive information as provided in Law 1581 of 2012 and the Decree 1377 of 2013. It shall also avoid including information that cannot be collected, stored, accessed, or shared according to that law, that is, private information about minor children. Furthermore, according to articles 17(k) and 18(f) of Law 1581 of 2012 and article 13 of the Decree 1377 of 2013, the person who is processing personal data must have a manual of data processing and terms of use in place, informing the rightsholder about the procedures to correct the data and withdraw the authorization. Such manuals must clearly disclose the ways in which the data will be processed and shared. Failure to comply with these rules may result in sanctions by the supervisor, the Superintendency of Industry and Commerce -SIC-.

Additionally, any person handling data set models that are copyrightable must comply with Colombian copyright laws. When such laws apply, those who are processing work that is copyrightable must be authorized by the rightsholder to use it. As previously mentioned, although

information itself is not copyrightable under Colombian laws, it can be protected by copyright laws when it has enough creative expression to qualify as a copyrightable work.

Furthermore, projects that involve text data mining may participate in the *Sandbox on privacy by design and by default in Artificial Intelligence projects* put in place by the Superintendency of Industry and Commerce in 2021. This program allows projects that are in the design stage and have not been completed to continue the development of the project with the specialized advice from the supervisor. However, to participate in this program the applicant must comply with the following requirements: "(1) they must be AI projects; (2) the project focus on e-commerce and/or advertising and marketing; (3) the project involves the processing of personal data; (4) the project is in a design stage; (5) the project has not started the processing of personal data; and (6) the project has not been completed or has been put into operation." In addition, if the AI project involves FinTech, the developer could also apply to the sandbox put in place by the Superintendency of Finance of Colombia, that allows companies to develop the project under the supervision and with the assistance of the financial supervisor.

In conclusion, there are no specific prohibitions to any specific types of data mining and all of them are permitted as long as they comply with Colombian regulations. Moreover, Colombia provides alternatives to develop AI projects with the help of the supervisor in the so-called sandboxes.

2. **Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question?**

In Colombia there are no laws or regulations governing web crawling. In consequence, there are no explicit prohibitions specifically applicable to such activity. However, since data crawling from the internet typically involves downloading HTML content from the website to iterate on the addresses that can be found in such text, if such a process requires collecting, processing, using or sharing personal data, the researcher must comply with the privacy laws of Colombia.

To the extent that there are no specific regulations prohibiting web crawling in Colombia, as long as such activity does not violate privacy, IP, or criminal laws, NLP researchers could legally crawl the web themselves.

Assuming that the applicable laws and regulations to data crawling are privacy and IP laws, there are no regional restrictions applicable to specific jurisdictions in Colombia. Laws enacted by the Congress of the Republic and the regulations issued by the government regulating such laws are applicable to all the regions of the country. Privacy and IP laws belong to this category of laws, and thus apply to all regions and jurisdictions with no distinction.

3. **If NLP researchers can crawl data ourselves, what are the limits regarding using this data for training a language model? Can NLP researchers redistribute the data afterwards? If so, under which license and in which geographical regions?**

As mentioned in *section C(2)*, web crawling is not expressly prohibited in Colombia, nor is web crawling to train language models. Regardless of the lack of regulation, in some cases, crawling the web to train language models may require compliance with privacy, IP, and criminal laws. Such laws may set limits regarding the use of the information subtracted from websites by downloading HTML that will be used to iterate the addresses contained within, and train the language model.

Privacy laws restrict the use of personal information. In particular, they establish a series of obligations aimed at protecting *habeas data* right of those providing their information, requiring from those collecting, processing, using and sharing such data to (1) put in place a series of measures to protect the information and the rights of the owner of the data; (2) adopt special protections and procedures to process certain types of data; and (3) avoid the usage of certain information whose usage is expressly prohibited by law.

Additionally, assuming that through crawling the web NLP researchers gain access to copyrighted material, they must comply with Colombian IP regulations. According to those rules, any use of copyrighted material must be authorized by the owner of the work or the owner of the economic rights attached to such work to be legally used. Thus, to legally train a language model with information crawled from the web, NLP researchers must be authorized to use the copyrighted material.

Finally, criminal law sets forth a series of rules prohibiting illegal or fraudulent access and use of private information. As mentioned in *section E(2) below*, there are a number of conducts regarding data and use of information that can be considered a crime. NLP researchers are prohibited to engage in any of the conducts that are mentioned in such norms.

In conclusion, although there are no specific norms regulating web crawling to train language models in the country, when NLP researchers crawl the web to train legal models, they must consider the limitations described in this document in regard to the application of certain laws and regulations.

4. What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify? The edition conversations of Wikipedia?

Colombia does not have a special regime applicable to social media platforms. Therefore, there is no definition of social media content or any provision that allows individuals to determine whether the news section of a website or the conversations in Wikipedia can be considered social media content.

Notwithstanding the above, the Congress of Colombia is currently seeking to enact a special regime applicable to social media content. In 2019, Bill No. 176 of 2019 was submitted before the Chamber of Representatives, seeking to create a special regime applicable to the terms of use and the appropriation of social media content. According to the article 3(a);(c) of the draft that was submitted, social media consist of “ways of social interaction defined as the dynamic exchange between people, groups and organization in complexity contexts,” and internet social

media is “the result of the usage of social media services to reproduce and create new social relationships.” Under those definitions, both the comment section of a news website and the conversions in Wikipedia would be considered internet social media content. However, this Bill has not yet been approved and is far from becoming a law of the Republic of Colombia.

The major risk associated with data scraping of social media content is the violation of Colombian criminal, privacy and IP laws. Notably, the most relevant risk associated with social media scraping is the commission of a crime for accessing unauthorized data (*see section E(2) below*). Indeed, although scraping social media is not per se illegal in the country, it can ultimately result in the commission of a crime when, for instance, a person accesses private information without authorization of the owner of such data.

5. Does the consent override the Terms of Use?

Consent does not override the Terms of Use. In Colombia, Terms of Use typically are agreements including disclosures about the rules applicable to a website or a mobile application. Such terms can be established separately from the privacy policy that every person collecting, processing, using or sharing personal data must have in place according to article 13 of the Decree 1377 of 2013, or they can be part of such privacy policy, either because this latter is included within the Terms of Use or because such terms include references to the privacy policy. Consequently, contrary to override the Terms of Use, consent by those providing their personal data and accessing to websites or mobile applications is a necessary requirement within Colombian privacy laws, according to article 9 of the Law 1581 of 2012 and article 5 of the Decree 1377 of 2013.

However, in many cases, Terms of Use include provisions that go beyond the scope of privacy law. In those situations, although consent may not be expreselly mandated by privacy laws, it is required under civil law regulations, according to which consent is a necessary condition to validly execute binding agreements. Therefore, even in those cases where the privacy law does not mandate consent, the civil law imposes the obligation to consent to the terms of the agreement.

In conclusion, consent does not override the Terms of Use put in place by any particular website or mobile application. Rather, it constitutes a necessary requirement for developers to legally use personal data provided by those accessing the websites or the mobile applications, as well as a condition to validly celebrate binding agreements such as those the Terms of Use.

6. Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can’t be scrapped?

There is no regulation in Colombia regarding strapping or crawling in general terms and there is no difference about doing any of them on social media. However, personal data laws need to be observed and this data can only be used under the terms and conditions accepted by owners. Also children's information cannot be used based on the category of subject of special constitutional protection and under the superior interest of children under 18 years.

D. Privacy Questions

1. **Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?**

Personal Identifiable Information is considered personal data under Colombian regulation. Therefore its use is regulated and depends upon the given authorization by the owner. Datasets, databases, sites etc containing personal information cannot be used without owner's authorization and when authorization exists, its distribution across bordes is limited and article 27 of law 1581 of 2012 is applied.

Transferring personal data to third party countries is prohibited when those countries does not have, at least, the same level of personal data protection as Colombia has some exception apply when there is explicit and unequivocal authorization for its transfer, when medical treatment or public health require, financial reasons (bank transfer), because of international treaties based on the principle of reciprocity, transferences needed for the execution of a contract, public interest and judicial processes.

2. **Are there legal concerns around publishing a pre-trained language model that is trained on a subset of common crawl, a published dataset based on a craw of the web in terms of PII?**

In Colombia, there are no rules regulating language models or crawling processes. Therefore, there are no specific laws or regulations governing the use of a subset of common crawl to train a language model that is meant to be published. However, when such models use Personal Identifiable Information ("PII"), they are subject to Colombian privacy laws.

PII is considered personal data under the Colombian privacy laws, and any system involving collection, processing, usage or sharing of personal information will be subject to Colombian privacy laws. Article 2 of the Law 1581 of 2012 establishes the scope of application of the general privacy law of Colombia, providing that the principles and provisions set forth in it are applicable to any personal data registered or being part of a database that is not explicitly exempted of such regulation (*see Section E(2) below*). Article 3(c) of such law defines personal data as "any information linked or that can be associated with one or more determined or to be determined natural persons." In consequence, any pre-trained language model that uses PII has to comply with Colombian privacy laws.

Notwithstanding the above, PII information can be used under the Colombian privacy laws as long as those who use such information comply with all the requirements and provisions set forth in the laws and regulations. Thus, the major concern regarding PII and publishing pre-trained language models that have been trained on a subset of common crawl is in connection with specific types of data. Such kinds of data are primarily (1) sensitive data; and (2) data of minor children. In the first case, the major concern is about the way such information is used. In effect, collecting, processing, using and sharing personal sensitive data is per se prohibited, unless certain particular conditions are met. Article 5 of the Law 1581 of 2012 seeks to protect all kinds

of sensitive data in such a way that only in special circumstances and applying certain measures it could be used. Therefore, the first concern regarding PII would be in connection with sensitive data.

Additionally, any language model that has crawled the web to pre-train such a model cannot use private information regarding minor children. In Colombia, articles 7 of the Law 1581 of 2012 and 12 of the Decree 1377 of 2013 protect the information of minor children by prohibiting the usage of their private information, and restricting the use of their public information. In consequence, information of minor children can only be used if it is public and if those managing the information (1) respect the superior interest of minor children, as provided in the Colombian Constitution; and (2) respect the fundamental rights of minor children.

In conclusion, although Colombian regulations do not prohibit crawling the web to pre-train language models with the information contained therein, such models must observe Colombian privacy regulations, specially those concerning the treatment of sensitive information and information of minor children.

2. **Are there concerns around distributing models that have PII stored within it? That can expose such PII?**

The main concern around the distribution of models that have PII stored is compliance with privacy laws. As it has been mentioned in other sections, any person that collects, processes, uses and shares personal data must comply with Colombian privacy laws. Such laws protect the integrity of the information provided, and more importantly, the privacy of the persons providing such information. In that sense, the major concern regarding distribution of models with PII is the possible violation of the fundamental right of *habeas data* set forth in article 15 of the Colombian Constitution. In fact, all the protections established in the various norms that regulate privacy in Colombia are intended to protect such a right, which grants any Colombian citizen the right to privacy, and to know, update, and correct any information that is stored in public or private databases. In consequence, the most important factor to consider when distributing models with PII is to comply with privacy laws.

Nevertheless, when those collecting, processing, using and sharing personal data comply with all applicable regulations and have put in place all the required measures to guarantee the *habeas data* right of the individuals, it is possible to make such distribution within the country. Indeed, the most important aspect to consider when those models are created and distributed is that the people whose data is being collected have access to the procedures to know, rectify, modify, delete, and make any changes in connection with its personal data. Moreover, the person distributing such models must design and implement such measures since the inception of the project, according to articles 2.2.17.1.6(5) and 2.2.17.5.5 and of the Decree 1078 of 2015 that establishes privacy by design and by defect as a principle to provide public digital services. Although privacy by design and by default as set forth in such articles does not apply to persons that do not provide such services, privacy laws require compliance with data protection regulation to be implemented since the beginning of the project. Therefore, such measures shall be put in place even when there are not public digital services involved.

In conclusion, any person distributing models that have PII within them has to comply with privacy regulations and make sure that they are implementing all the necessary measures to protect the privacy of the individuals whose information is collected and processed, as well as to include processes and mechanisms to guarantee the exercise of the fundamental right of *habeas data*.

3. **What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exception for research purposes?**

NLP researchers must comply with the Colombian privacy laws primarily contained in the Law 1581 of 2012, the Decree 1377 of 2013, the Decree 1078 of 2015, and the Decree 1413 of 2017. According to articles 17(d) and 18(b) of Law 1581 of 2012, persons collecting, processing, and sharing personal data must put in place the necessary measures to impede such data to be modified, lost, consulted, used, or accessed in a fraudulent or unauthorized manner. Although article 19 of the Decree 1377 of 2013 provides that the supervisor has the competence to provide instruction on how the security measures must be put in place, such authority has not established specific means to comply with the aforementioned duty. In case a data breach occurs, articles 17(n) and 18(k) of Law 1581 of 2012 establish that those collecting, processing, and sharing personal data must inform the supervisor, the Superintendency of Industry and Commerce, about the breach. Moreover, the NLP researcher, or any person whose databases were breached, must take all the reasonable measures to repair the breach and recover the information. Additionally, according to article 9 of the Decree 1377 of 2013, the rightsholder of the personal data can withdraw their authorization at any moment and request the suppression of the personal data contained in the databases and following the privacy policies adopted by the company in compliance with article 13 of the Decree 1377 of 2013. Currently, Colombian privacy laws do not establish any exception for research purposes.

In conclusion, in Colombia, NLP researchers who collect, process, use or share personal data are obligated to takedown any personal data in case of a data breach or withdrawal of the authorization by the rightsholder. The measures to takedown such information must follow the privacy policy of the company, the privacy laws, and the regulations or instructions provided by the supervisor.

4. **What are the privacy risks related to data collection directly from persons? For example when you interview people or they donate data etc.**

The privacy risks related to the direct collection of data have to do with the explicit authorization of the owner of the data when it is personal (the individual) or carry it out without licenses when it is protected information.

When personal information is obtained directly from people without expressly informing the purposes and uses and without obtaining express authorization, whoever collects that information or distributes it may be subject to administrative sanctions or civil lawsuits.

Likewise, if data is used or collected when it is protected by copyright, there is a risk of assuming costs due to lack of licenses in possible civil lawsuits.

E. Prohibited content

1. What types of data may be prohibited from being text data mined?

Colombian laws and regulations do not directly prohibit any types of data to mined. However, there are several norms limiting the way in which data is collected, processed, used and shared. To that extent, the limitations and prohibitions over data mining may come from other norms, particularly from privacy law and laws regarding the use of data.

First, private data and 'semi-private' data cannot be mined if the owner of the information does not authorize the treatment of their personal data in the terms set forth by the person collecting, processing, using and sharing the data. According to subsections (g)-(h) of article 3 the Law 1266 of 2008, private data is the one that, by its intimate or reserved nature, is only relevant for the owner, and 'semi-private' data the one that has not an intimate, reserve or public nature and whose knowledge or distribution may not only be relevant to its owner but for an specific sector, group of people or the society in general. To legally process this data, the person collecting, processing, using or sharing the data must comply with all privacy laws and regulations. These rules require (1) that the person treating the personal data discloses the ways in which the provided data will be used through a privacy notice; (2) that the owner of the data authorizes such usage; (3) that those treating personal data adopt a privacy policy and manual establishing (A) the identification data of the person treating the data; (B) the use that will be given to such data, when such information has not been provided in the privacy notice; (C) the rights that the owner of the data has; (D) the name or the person or area in charge of the data and claims in connection with the data and the exercise of the rights that the owner has; (E) the procedures available to the owner of the data to exercise their rights; and (F) the date in which the privacy policy is in effect. If those treating personal data do not comply with these and any other requirements that may be necessary to collect and use the data, the treatment will be prohibited. In consequence, absent compliance with such laws and regulations the data cannot be used to be text data mined.

Additionally, there are certain types of data whose use is prohibited or highly limited, and in which data mining would be prohibited as well. That is the case of private data of minor children, that is strictly prohibited, and sensitive data, which is prohibited but can be used under certain circumstances and applying specific protective measures. Moreover, there are certain types of data that are not governed by privacy laws and are either prohibited to be used or subject to a special regime (*see section E(2) below*).

Finally, any data that has been obtained or accessed through illegal means cannot be used to be text data mined. In effect, if the information that will be mined has been obtained by the commission of any of the crimes listed in *section E(2)* below, the data cannot be mined. Moreover, if the person has knowledge of the commission of a crime in the access of such information, it is obligated to report it before the respective authorities.

2. What types of data may be prohibited from being stored or distributed?

In Colombia, there are different situations and types of data that cannot be stored or distributed. First, no person can store or distribute private personal data without the authorization of the rights holders. Also, private information of minor children (persons under the age of eighteen years) cannot be stored or distributed in accordance with article 7 of the Law 1581 of 2012 and article 12 of the Decree 1377 of 2013. Furthermore, storage and distribution of sensitive data is prohibited by article 6 of the Decree 1377 of 2013, unless there is an applicable exception of those provided in article 6 of the Law 1581 of 2012. According to article 5 of such law, sensitive data includes, but is not limited to, any data that reveals racial or ethnic origin, political orientation, religious or philosophical convictions, memberships to unions, social or human rights organizations, organizations that promote the interest of any political party. that guarantee the rights of opposition political parties, as well as data related to health, sexual life and biometrical information.

Additionally, there are some types of data that are not covered by the general privacy laws. Such data includes (a) databases or files that are exclusively held in a personal or domestic environment; (b) databases containing national security data, as well as information related to the prevention, detection, monitoring and control of money laundering and financing of terrorism, whose access is prohibited and can be denied in accordance to article 19 of the Law 1712 of 2014; (c) databases whose primary purpose is to provide services of intelligence or counterintelligence, which are reserved for a maximum of 45 years according to article 33 of Law 1621 of 2013; (d) databases held by journals and other editorial contents; (e) databases containing financial information, which are regulated by the Law 1266 of 2008 that provides a similar regulation to that established in the Law 1581 of 2012 and the Decree 1377 of 2013; and (f) databases and files that are governed by the Law 79 of 1993, regarding the national census.

Furthermore, certain actions related to IT may constitute a crime under Colombian laws. The Law 1273 of 2009, that partially amended the Colombian Criminal Code, establishes a set of conducts related to data and databases that constitute a crime. Among these, article 2069A-J of such law considers as a crime (a) the fraudulent or abusive access to computer systems; (b) illegitimate obstruction of a computer system or a telecommunication network; (c) intentional harm to computer systems; (d) usage of malware; (e) usage of personal data without authorization; (f) impersonation of websites to get access to personal data; (g) theft or fraud using computer systems; and (h) unauthorized transfer of assets through computer systems.

In addition, according to article 18 of the Law 1712 of 2014, public officials that have access to confidential information by virtue of their work in the government cannot distribute or reveal such information to third parties. That is the case, for example, of governmental officials who get access to trade secrets.

In conclusion, any information that has been obtained in violation of the aforementioned rules or any other regulations that may be applicable to special circumstances or areas of practice is prohibited and can result in sanctions that go from monetary sanctions to imprisonment, depending on the nature and severity of the violation.

3. **What types of data may be prohibited from being generated?**

Colombia does not prohibit any particular type of information to be generated. However, the same restrictions applicable to the information that cannot be stored or distributed described in *section E(2)* are applicable to the generation of new data. Therefore, data that is generated in violation of the prohibitions mentioned in the previous section will likely be sanctioned by the Colombian authorities.

4. **What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?**

The essential legal norms that are implicated by prohibited content are (1) constitutional norms protecting *habeas data*, and confidentiality and information of minor children; (2) civil and commercial laws and regulations governing privacy and personal data; (3) criminal laws protecting fraudulent or unauthorized access to data; and (4) laws regulating national security. Each of these norms is intended to protect different interests, but they all seek to protect the constitutional right of *habeas data* and to prevent the usage of private information for illegitimate or criminal purposes.

1. The law 1581 of 2012 defines sensible data as the data belonging to an individual which wrong use might cause his or her discrimination. Since what this norm is looking for is to protect privacy, as a general rule the treatment of sensible data is prohibited. There are some exceptions to this prohibition as when explicit authorization is given, when the owner interests need to be protected, when data is needed to execute activities of an NGO referring only to its members, when it is necessary for a judicial process or for historical, statistical or scientific purposes, in this cases information of the owner needs to be suppressed.
2. Only because of the superior interest of children under 18 years of age can data be collected. It is prohibited the treatment of children's personal data except when it is public information.
3. Under article 4 of Law 1581 of 2012, the treatment of incomplete data that can lead to a mistake is prohibited.
4. Transfer of personal data to third party countries is prohibited with exceptions under article 26 of Law 1581 of 2012.
5. Chapter 1 of Law 1273 of 2009 (modifying the criminal law)regulates confidentiality, integrity and availability of sta and informatic systems. It prohibits the abusive access to an informatic system without authorization (whether or not it has security measures). Also, prohibits the violation of personal data. None without authorization can compile, subtract, offer, sell, exchange, send, buy, intercept, divulgue, modify or use codes, personal data, databases or similar.
6. Child pornography is explicit prohibited by the law 360 of 1997 and the criminal code tittle IV.

5. **What types of licensing or other control mechanisms would be preferred under the applicable jurisdictions?**

Colombia does not have any preferred control mechanisms to supervise the development and implementation of AI or Machine Learning projects. However, in recent years, the authorities that oversee those who develop such kinds of systems have encouraged innovation through the so-called *Sandboxes* and supervised spaces.

In particular, the Superintendency of Finance ("SFC") and the Superintendency of Industry and Commerce ("SIC") have created spaces for companies and individuals to develop AI projects and any other project involving technology in a supervised and controlled manner.

The SFC has led the development of these spaces. Since 2017, this agency has developed four spaces for companies under its oversight to innovate and develop projects with the help of the supervisor under a project called 'Innova SFC.' These four spaces are (1) 'elHub,' that allows companies to receive support, guidance and feedback from the SFC in any project involving financial and technological innovation; (2) 'laArenera,' which is a sandbox that permits companies to develop their projects under the supervision and guidance of the SFC; (3) 'regTech,' that consists of a regulatory technology project that requires the usage of technology to booster innovation within the supervisor by optimizing the internal and supervision processes; and (4) 'supervised testing space' (*espacio de prueba controlada*), that allow companies supervised by the SFC to test innovative technologies with the support and supervision of the SFC. Although this initiative is limited to innovative developments in Fintech, other authorities have followed the example of the SFC.

The SIC, which is the government agency that supervises compliance with consumer, privacy, antitrust, and industrial property laws in the country, recently implemented a sandbox aimed at promoting the development of AI projects. In 2021, that agency created the *Sandbox on privacy by design and by default in Artificial Intelligence projects*, allowing projects that are in the design stage and have not been completed to continue its development with the specialized advice from the supervisor. Similar to the project created by the SFC, the SIC's sandbox seeks that companies innovate while duly complying with all applicable laws, especially privacy laws.

In conclusion, Colombia prefers to use mechanisms of control that allow companies to develop their AI projects with the help and guidance of the supervisor.

6. Is there a legal restriction on the distribution of data for national security reasons?

Yes, there is a legal restriction regarding the distribution of data for national security reasons. According to article 33 of Law 1621 of 2013 the information collected, processed and used by the intelligence and counterintelligence entities is reserved and cannot be published, shared or distributed. Article 2 of such law defines intelligence and counterintelligence as the recoleccion, processing, analysis and distribution of information aiming to protect human rights, and to protect and combat internal or external threats against democracy, the constitutional and legal regime, the national security and defense, and any other ends included in such law. Therefore, NLP researchers cannot and must not access information that is subject to reserve for national security.

Additionally, article 19(a)-(b) of the Law 1712 of 2014 provides that information regarding national security and defense, as well as public safety, must be kept in reserve, and any request of disclosure must be denied. Therefore, information regarding national security cannot be distributed or shared under Colombian law.

7. Would a model trained on the data also fall under such restrictions?

Yes. Models trained on data that is subject to reserve are illegal and can be considered a criminal infraction under Colombian laws. Therefore, the prohibitions mentioned in *section E(6)* also apply to data models trained with such information while the reserve is in place.

CANADA

Sally Kang

Introduction

In Canada, the law comprises two components: 1) case law; and 2) legislation (e.g., statutes). Both are primary sources of Canadian law.

Case Law

Case law consists of the written decisions of the various levels of courts across the country. As a common law jurisdiction, Canadian courts adhere to the principle of *stare decisis*, which requires that they follow the previous rulings (i.e., precedents) of higher courts in their province or territory and/or the Supreme Court of Canada on the same issue. Decisions from the same level of court or other provinces or jurisdictions are persuasive in that they may assist a court in reaching a decision.

Legislation

Canada is a federation – a union of several provinces and territories with a central government. *Constitution Acts, 1867 to 1982* set out the division of powers between the provincial/territorial legislatures and Parliament (i.e., the federal government). Parliament makes and modifies laws for all of Canada, while the legislatures in each of the ten provinces and three territories handle the laws applicable in their respective province/territory. Laws enacted at either the federal or provincial/territorial level are called statutes, legislation, or acts. When Parliament or a provincial/territorial legislature enacts a statute, the statute supersedes the common law case law and precedents dealing with the same subject or issue. In addition to developing common law by referring to and setting precedents, courts are tasked with interpreting and applying statutes.

Because patent and copyright fall within the scope of the federal government's powers, legislation concerning them is enacted by Parliament and applies nationwide. The division of powers in regards to privacy and data are more ambiguous; accordingly, both Parliament and the provincial/territorial legislatures have enacted legislation in relation to these issues. Namely, there is federal legislation that governs how the federal government handles personal information (see *Privacy Act*); each province/territory has its own laws regarding provincial government agencies' handling of personal information. Parliament has also enacted legislation that addresses how businesses handle personal information and is to be applied nationwide; that said, some provinces (i.e., Alberta, British Columbia and Quebec) have their own private-sector privacy laws that may override the federal legislation in some cases.

Relevant Legislation

Federal:

- *Patent Act*, RSC, 1985, c P-4
- *Copyright Act*, RSC, 1985, c C-42
- *Privacy Act*, RSC, 1985, c P-21
- *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5 ("PIPEDA")
- Bill C-11: *Digital Charter Implementation Act, 2020*

Provincial:

- Alberta: *Personal Information Protection Act*, SA 2003, c P-6.5
- British Columbia: *Personal Information Protection Act*, SBC 2003, c 63
- Quebec: *Act respecting the Protection of Personal Information in the Private Sector*, CQLR, c P-39.1

Questions

A. IP Questions

1. Are training datasets and models protected by IP rights. If so, which IP rights?

Applicable legislation:

- *Copyright Act*, RSC, 1985, c C-42

Analysis:

Machine-learning algorithms are trained, validated and tested with input from datasets. Training datasets are sets of data from which an algorithm "learns."

The *Copyright Act* ("the Act") provides that computer programs/software are copyrightable subject matter. More specifically, source code (human-readable) and object code (machine-readable) are both protected as "literary works" under the Act (see sections 2 and 5(1) of the Act). It should be noted that even if datasets do not qualify as computer programs/software, they - and those used to train language models - would nevertheless be copyrightable on the basis of being compilations of literary works (see subsection 2.1(1) of the Act).

In addition to being a copyrightable subject matter, to receive copyright protection, the work in question must have sufficient expression, fixation, originality, and authorship.

Training datasets would appear to have sufficient expression because they are not mere ideas, an important dichotomy and distinction in copyright law. It is a basic principle of copyright law that copyright protection extends to *expression* in a work and not to ideas, procedures, methods of operation or mathematic concepts. Copyright does not protect basic ideas or information, but rather the form in which they are expressed. Datasets can comprise basic ideas and/or information; that said, the *compilation* of such ideas and/or information can attract copyright protection because their *selection and arrangement* may constitute sufficient expression.

Assuming that training datasets can and will exist in some concrete, tangible form (e.g., written down), they will satisfy the fixation requirement.

Lastly, so long as the training datasets are the result of skill (use of one's knowledge, developed aptitude and/or practiced ability in producing the work) and judgment (use of one's capacity for

discernment, ability to form an opinion and/or evaluation by comparing different possible options in producing the work) that are not trivial or mechanical, they will satisfy the originality requirement. As alluded to previously, although datasets may be a compilation of data, so long as skill and judgment are employed in both the selection and arrangement of the data, the datasets will be considered original.

As for authorship, in order for copyright protection to extend to the training datasets, the datasets must be authored by a person – i.e., a human must be responsible for the creation of the original fixed expression. Assuming that an NLP researcher or AI engineer has compiled the dataset, the authorship requirement will be met. On the other hand, if the datasets are entirely machine-generated, the authorship requirement will likely not be satisfied and the datasets not protected by copyright.

As computer programs/software and/or compilations of copyrighted/non-copyrighted (i.e., public domain) works created by an NLP researcher or AI engineer, copyright will likely subsist in the training datasets. That said, copyright in the compilations/datasets does not give the compiler rights in the underlying content (see subsection 2.1(2) of the Act). Accordingly, if copyrighted works are included in the datasets and authorization to use the works from their respective copyright owners (via license or otherwise) has not been granted, then the author of the datasets is likely liable for infringement with respect to the copyrighted works used in the datasets by reproducing them in the datasets.

Despite the likelihood of infringement, the author(s) of the datasets may be able to avail themselves to the fair dealing defense by successfully demonstrating that their use of the copyrighted works was fair. Fairness is assessed based on the following factors: 1) purpose (private vs. commercial); 2) character of the dealing; 3) amount of the dealing; 4) alternatives to the work that could have been used instead; 5) nature of the work (Published? Confidential?); and 6) market effect. An NLP researcher's fair dealing defense will likely succeed if it can demonstrate that: 1) the purpose was non-commercial; 2) distribution of the copyrighted works would be circumscribed; 3) a small, limited portion of the works was appropriated; 4) the works were necessary in achieving the authors' purpose such that alternatives were insufficient; 5) the works are not confidential; and/or 6) the copying of the works likely will not deleteriously impact the copyright owners' respective markets for their works (e.g., declining sales, etc.). Note that not all of the foregoing must be demonstrated and there is no one determinative factor – courts will balance the factors, and assess whether the evidence presented on the whole and considered together weighs in favor of finding fair dealing in the use of the copyrighted works.

Conclusion:

Copyright can subsist in training datasets because they are either/both computer programs/software and compilations; however, the author(s) of the datasets do not acquire copyright in the works that comprise the datasets. Moreover, should any portion of the datasets comprise copyrighted works, the copyright owners' right to reproduce their respective works may be infringed and the dataset author(s) liable for that infringement. That being said, the defense of fair dealing may be raised to escape liability.

2. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?

Applicable legislation:

- *Copyright Act*, RSC, 1985, c. C-42

Analysis:

Under the analysis in A.1, it was established that copyright can subsist in training datasets. Recall that even if a published dataset comprises solely of public domain (i.e., non-copyrighted) works, the dataset as a compilation of these works may attract copyright protection, if the selection and arrangement of the works within the dataset is sufficiently original and all other requirements for copyright protection (i.e., fixation, expression and authorship) are satisfied. Accordingly, a "subset of a published dataset based on a crawl of the web" may attract copyright protection, and breaching the dataset creator's (i.e., copyright owner) exclusive right to make use of that work may constitute infringement, unless the right has been licensed or otherwise granted to the database creator by the copyright owner.

However, whether the dataset's use in relation to training the language model engenders infringement is murky. Subsection 3(1) of the *Copyright Act* enumerates the copyright owner's exclusive rights, stating that such an owner has "the sole right to produce or reproduce the work or any substantial part thereof in any material form whatever, to perform the work or any substantial part thereof in public or, if the work is unpublished, to publish the work or any substantial part thereof, and includes the sole right (a) to produce, reproduce, perform or publish any translation of the work . . . ." (emphasis added).

Under this provision, an argument could be made that the specific exclusive rights that may be infringed by training the language model on a subset of a published dataset are the dataset creator's right to perform or publish the translation of the work. In learning and arguably interpreting the dataset and transforming it into a language model, the dataset has been "translated" into the language model. The subsequent publication of the language model may constitute "publishing a translation of the work," while its actual use may be considered a "performance" of the subset of the dataset used to train the language model. If one or both of these analyses are accepted, then publishing the language model, without the proper authorization/license from the copyright owner, would infringe the copyright in the dataset subset.

Given its novelty, there is no jurisprudence to date that can speak, either directly or tangentially, to whether the foregoing analysis has been applied and/or accepted by the courts to find infringement in the use of a previously published dataset for training a language model. In any event, if the dataset creator were to pursue an infringement claim, the fair dealing defense may be invoked by the creator of the language model to avoid liability.

The analyses under both A.1 and A.2 appear to be supported by the fact that the House of Commons' Standing Committee on Industry, Science and Technology issued, in June 2019, a report setting out a number of recommendations related to AI, one of which was that the *Copyright Act* be amended to facilitate the use of a work or other subject matter for the purpose of

informational analysis, and make the list of permitted purposes under the fair dealing exception illustrative rather than exhaustive.

Conclusion:

Whether publishing a language model trained on a subset of a published dataset will engender copyright infringement liability is a novel issue that does not have much in the way of supporting jurisprudence. That said, a reading of the *Copyright Act* suggests that the provision regarding copyright owner's exclusive rights can be construed as indicating that the publishing of a language model constitutes publishing a translation of the subset of the dataset on which the model was trained, and that the actual use of the model constitutes performance of the dataset. To the extent that these propositions are persuasive, the language model's publication may constitute copyright infringement in the dataset. However, fair dealing may be raised as a defense to escape liability.

3. What are the IP risks related to data collection directly from persons? E.g., when you interview people or they donate data, etc.

Applicable legislation:

- *Copyright Act*, RSC, 1985, c C-42

Analysis:

Assuming that 1) the data collected directly from persons will be incorporated into a training dataset, and 2) the data collector also creates/compiles the dataset, the primary issue is whether the data collector/dataset compiler is the author of the work.

Recall that an author is the person responsible for the creation of original fixed expression; recall also that a dataset may be deemed a sufficiently original fixed expression to which copyright protection adheres.

In instances of interviews/direct data collection, the interviewer/data collector who reduced the interviewee's oral statements to a fixed form is the author of the work. However, it is imperative that the interviewer/data collector do more than merely fixing the work on behalf of the interviewee or acting as a "scribe." Rather, they must play a significant role in the interview process, prompting answers with questions, guiding the conversation, and subsequently arranging the material into a fixed expression. Where the foregoing can be demonstrated, the interviewer/data collector would likely be found to be the author, thereby owning the copyright in the interviewee's statements and/or provision of information as fixed in a tangible form.

Conclusion:

Direct collection of data - by way of an interview, for instance - does not appear to engender liability or raise significant legal issues for the data collector, who then incorporates that data into a dataset.

4. Can NLP researchers train a language model on borrowed library books? E.g., from the Internet Archive or other online book lender?

See A.2, as the analysis is very similar.

Note, however, that there may be no legal issues for books that are no longer copyright-protected and have entered the public domain – these are free to be used by anyone for any (legal) purpose, as the *Copyright Act* no longer applies to protect the copyright owners of these works. Liability for infringing the copyright owners' exclusive right to perform their works and/or publish translations of same may only arise for works still protected under the *Copyright Act*.

5. How do rights on the source data (e.g., copyright) transfer to the trained model?

Applicable legislation:

- Copyright Act, R.S.C., 1985, c. C-42

Analysis:

Assuming that "source data" refers to the training dataset, there is no "transfer" of the copyright from the dataset to the trained model *per se*. Rather, a new copyright, separate that subsisting in the datasets, will imbue the trained model, which is arguably a derivative work of the dataset. The formation of the new copyright protection in the derivative trained model does not depend on where training occurs or where the data in the dataset is gathered.

That said, in order for copyright to exist in the trained model, the model must be a copyrightable subject matter (likely satisfied as a computer program/software), and have sufficient expression, fixation, originality and authorship.

Conclusion:

The copyright in the dataset does not transfer to the trained model; rather, a new copyright forms in the trained model upon its creation, so long as the requirements for copyright protection are satisfied.

B. Licensing Questions

1. Are datasets licensable? Are models licensable?

Applicable legislation:

- *Copyright Act*, RSC, 1985, c C-42

Analysis:

Subsection 13(4) of the *Copyright Act* provides for the assignment and licensing of the copyright in a work:

Assignments and licences

(4) The owner of the copyright in any work may assign the right, either wholly or partially, and either generally or subject to limitations relating to territory, medium or sector of the market or other limitations relating to the scope of the assignment, and either for the whole term of the copyright or for any other part thereof, and may grant any interest in the right by licence, but no assignment or grant is valid unless it is in writing signed by the owner of the right in respect of which the assignment or grant is made, or by the owner's duly authorized agent.

Ownership in case of partial assignment

(5) Where, under any partial assignment of copyright, the assignee becomes entitled to any right comprised in copyright, the assignee, with respect to the rights so assigned, and the assignor, with respect to the rights not assigned, shall be treated for the purposes of this Act as the owner of the copyright, and this Act has effect accordingly.

Assignment of right of action

(6) For greater certainty, it is deemed always to have been the law that a right of action for infringement of copyright may be assigned in association with the assignment of the copyright or the grant of an interest in the copyright by licence.

Exclusive licence

(7) For greater certainty, it is deemed always to have been the law that a grant of an exclusive licence in a copyright constitutes the grant of an interest in the copyright by licence.

Because copyright will likely subsist in both the training dataset and model, the foregoing provisions apply to them, rendering them assignable or licensable by written agreement "signed by the owner of the right in respect of which the assignment or grant is made, or by the owner's duly authorized agent."

As previously noted in A.2, the House of Commons' Standing Committee on Industry, Science and Technology issued, in June 2019, a report setting out a number of recommendations related to AI, one of which was amendment of the *Copyright Act* to facilitate the use of a work or other subject matter for the purpose of informational analysis. The Government has not identified a timeline for introducing copyright reform legislation in Parliament, but there is a growing understanding (and, arguably, concern) that Canada runs the risk of falling behind other countries (e.g., the US, Japan and the EU), which have implemented copyright regimes that allow for information analysis of works without a separate license.

Note that licenses granting use of the datasets or models should outline the specific rights that the licensee can enjoy and the obligations that they must fulfill in order to enjoy those rights.

Failure to observe the terms of the license (e.g., if a the dataset/model is used in a way that is not countenanced by the license or if the licensee's obligations are not satisfied), then the potential violation may amount to a breach of contract.

Conclusion:

Both datasets and models are likely licensable.

2. Overview of dataset licenses

Applicable legislation: N/A

Analysis:

Publicly available datasets are being widely used to build AI software. The usage of these datasets is governed by the license associated with a particular dataset. Such licenses outline the rights to which licensees are entitled and the obligations they must fulfill in return to enjoy these rights.

In the context of AI software (e.g., language model) development, the rights set out in a dataset license determines the specific way(s) that the dataset can be used (e.g., for building commercial applications, redistribution to third parties, etc.). The obligations under a dataset license can be construed as software requirements that need to be captured and
traced throughout the development and distribution of the AI software if the dataset is used (i.e., the AI software must be distributed under the same license governing the dataset upon which the software is based). Therefore, to ensure that the use of a dataset is legal, observing the rights and obligations outlined in the dataset's license is crucial.

Identifying the rights and obligations enumerated in the license associated with a publicly available dataset can be challenging for several reasons, first of which is the difficulty in locating and identifying the complete and correct license(s) applicable to a given dataset. Second is the difficulty in verifying the validity of the license associated with a given dataset. Many publicly available datasets are created by combining data from multiple data sources, each of which may have different licenses whose terms are inconsistent or contradictory. As well, creators of publicly available datasets rarely document the different licenses associated with the different data sources used, and routinely fail to consider the impact that these different licenses may have on the license that is created and imposed in relation to the aggregated dataset, potentially rendering the resulting license invalid. Finally, the licenses of publicly available datasets are typically ambiguous and do not clearly identify the intended rights and obligations that govern the usage of datasets, making it difficult in practice to use the datasets to build AI software without the risk of license violation.

Reviewing and understanding the license of a publicly available dataset is important for the purposes of determining if the given dataset can be used in a specific way (e.g., for model training) while ensuring compliance with the terms of the license. When an AI software uses a publicly available dataset, the owners of the AI software implicitly enter into an agreement with the

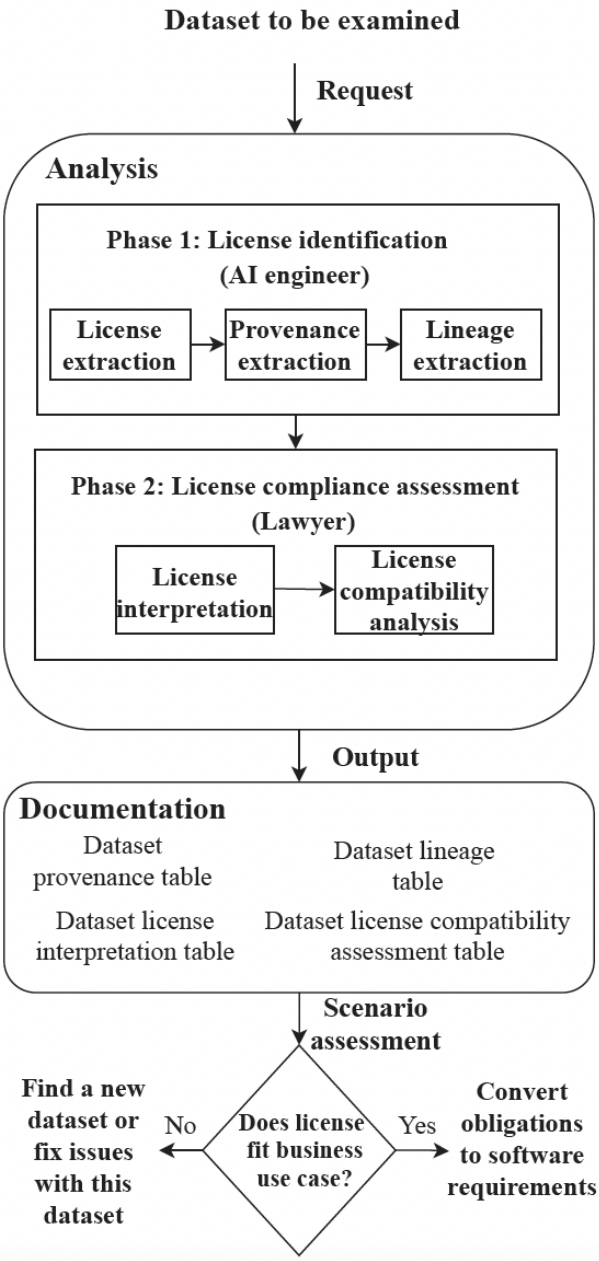
copyright owner(s) of the dataset, whereby it is agreed that the AI software is to only use the dataset pursuant the rights granted in the license and that the obligations also outlined in the license will be satisfied. If there is a failure on either front, AI software and its creator will likely have violated the dataset's license and may be liable for breach of contract.

As noted above, licenses for publicly available datasets typically do not clearly state the rights and obligations associated with the usage of the datasets. Where requirements are ambiguous, it is vital that NLP researchers/AI engineers do their due diligence or take due care to ensure that their software and systems are compliant with those requirements – that is, they should make reasonable efforts to ensure that their AI software or model is not in breach of licenses associated with the data/works on which the software is based or trained. Additionally, AI engineers should be able to justify that their interpretation of a license's rights and obligations is consistent with how they implemented those requirements in the constructed AI software or model. Accordingly, a rigorous approach to license compliance which ascertains the rights and obligations associated with a publicly available dataset is crucial, especially given that identifying a dataset's provenance and lineage is often convoluted.

Of the few dataset-specific licenses that are available, they can be categorized as either *copyleft* and *permissive* licenses. Copyleft dataset licenses typically require the AI software using the dataset to be licensed under the same license under which the dataset was provided (alternatively, they do not clearly mention how the software's license should read or be structured). Permissive licenses allow the AI software that uses the datasets to be licensed under any license that the AI software's engineer chooses.

For more on a potential method for assessing and anticipating the potential license compliance violations that may occur when using a given publicly available dataset to build AI software, see "Can I use this publicly available dataset to build commercial AI software? Most likely not" at https://arxiv.org/abs/2111.02374. The method described in this paper is summarized in the following chart:

Figure 1: An overview of our approach.



The researchers of this article concluded that publicly available datasets may not be suitable for building commercial AI software, reasoning, in part, that even if a dataset's license permits all rights to the dataset (even commercial use) so long as certain obligations are fulfilled, the licenses of the data sources may be more restrictive, creating potential risk of license breach if the data contained in the dataset is used in a way not permitted by the data source licenses. Such precluded uses could include using the data to train an AI model, modifying or distributing the data, etc.

The paper also sets out some general recommendations for the AI engineers who use publicly available datasets to create AI software:

Recommendation #1: Employ caution while using publicly available datasets to build commercial AI software. Just because a dataset is publicly available, it doesn't imply that it could be used to create commercial AI software. Therefore, it is pivotal for the AI engineers to use our approach to verify if the publicly available dataset that they are interested in using can indeed be used in the commercial context that they intend to use.

Recommendation #2: While assessing the license compliance of datasets, it is essential to use systematic approaches and clear documentations to demonstrate due diligence. We recommend the AI engineers to use our proposed approach and the schema to document various aspects (e.g., provenance, lineage and the rights/obligations) of a dataset and its license(s).

Recommendation #3: Share the knowledge regarding the risks associated with using a given publicly available dataset commercially. We recommend that AI Engineers to openly share and maintain the use cases where a given dataset can or cannot be used for the following three reasons: (1) to advocate the importance of this area of work, (2) to minimize the duplicated effort while analyzing the license compliance for the same datasets, and (3) to further validate the analysis via crowd sourcing.

Quick answers to pertinent questions/concerns based on Analysis:

- Q: What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?

  A: To determine the validity of a *dataset* license, researchers should determine whether it contradicts/is inconsistent with any of the provisions in the *data source* license. If there is such contradiction/inconsistency, then the dataset license may be found to be invalid and not legally binding.

  As to the actual terms of dataset (or data source) licenses, there is no definitive answer as to which terms and on what grounds each term may be found invalid, although the general answer would be that licenses are governed by contract law as established through case law. There may be terms in a license that are deemed invalid for a multitude of reasons, each of which may have its own extensive case law/jurisprudence. To determine the validity of a term, it would be best to consult a lawyer who would be able to conduct the appropriate research pertaining to that specific term and provide a legal opinion on its validity.

- Q: What if the users download or copy their own data and then provide it to NLP researchers directly?

  A: As best practice, the NLP researchers/AI engineers should still systematically trace and identify all applicable licenses, and determine the rights and obligations contained therein as they apply to the researcher/engineer's use of the data or dataset.

- Q: Does the license that the dataset is shared under override the license of the data source?

  A: No, the license of the data source overrides the license for dataset – see B.2 for a detailed explanation.

- Q: Can NLP researchers use licensed copyrighted data, to which they legally have access, to train a language model? If so, how does this affect the publication and distribution of the model?

  A: Hypothetically, yes, so long as such use (whether it is training a language model, publishing the model, or distributing the model) is granted in the license of the copyrighted data, and the use is in compliance with all of the relevant license terms and conditions.

- Q: Can data gathered by text data mining be released by means of a license? What about a model trained on the collected data and the training dataset (i.e., the compilation of the data structured in a specific way)?

  A: Data gathered by text data mining (i.e., a dataset of the aggregated data) can be released by means of a license; however, it is advisable that the terms of the dataset license be consistent with - or otherwise not contradict - those set out in the licenses of data sources in order to avoid potentially invalidating the dataset license. As for the model trained on the collected data and the training dataset, it can also likely be released via a license, provided that the terms of the license for the model is consistent with - or otherwise not contradict - the data source licenses and the dataset license.

3. What are the relationships between laws, licenses, and terms of use? Which is more binding?

Applicable legislation: N/A

Analysis & Conclusion:

Laws/legislation (i.e., case law and jurisprudence; federal or provincial/territorial statutes) take precedent and are the ultimate forms of legal authority. Jurisprudence in contract and related areas of law interpret and otherwise govern licenses. Terms of use may be construed as a license or some other subset of contracts. Licenses and terms of use are less binding than legislation.

C. Text Data Mining and Fair Use Questions

1. What are types of legally permitted text data mining?

Applicable legislation:

- *Copyright Act*, RSC, 1985, c C-42

Analysis:

It would appear that text data mining can be conducted freely and without restrictions on works that fall within the public domain (i.e., are not protected by copyright). However, for copyright-protected works, what constitutes legally permissible text data mining is governed by contracts in the form of data mining policies, terms of use, terms and conditions, etc. The copyright owner of a work to be mined may wish to grant mining rights in this way because the dataset that is created from compiling the information/data extracted from the mine could arguably constitute a derivative work, the production of which the copyright owner presumably retains an exclusive right.

For instance, Canadian Science Publishing ("CSP"), Canada's largest publisher of international scientific journals, has a "Text and Data Mining Policy" which "grants subscribers and other lawful users ("Users") the right to text and data mine (TDM) online content for non-commercial purposes according to the terms and conditions of this policy." The policy provides, in part (the full policy can be found at https://cdnsciencepub.com/about/policies/text-and-data-mining):

CSP grants Users the non-exclusive, non-transferrable right to text and data mine CSP content for the purposes of non-commercial, scholarly research related to specific projects. TDM and TDM Output must not be used for direct or indirect commercial purposes.

For commercial use, the policy advises that Copyright Clearance Center's RightFind® XML for Mining be contacted.

It would appear that the copyright owner of the work to be mined must grant (in the foregoing way or otherwise) an explicit right to text data mine the work because the *Copyright Act*, as noted elsewhere in this Playbook, does not currently permit an exception for informational analysis, which has been described as "encompassing the use of processing techniques to obtain and process text, images, sound, video, and other forms of data in order to generate new facts, discover patterns, and analyse relationships."

It has been noted by industry and legal experts that current uncertainties in the *Copyright Act* curtail the ability for Canadian companies to "access a basic, necessary resource to train their algorithms" and that this could be resolved by amending the Act to include a fair dealing exception for informational analysis, which would ensure that copying a lawfully accessed work for the purpose of information analysis is not infringing on the copyright that subsists in the work being accessed and mined.

That said, until AI-focused copyright reforms (i.e., such as the enactment of a targeted informational analysis fair dealing exception) occur, where copyrighted works have been text data mined and there is no policy, terms of use, etc. permitting such use, the fair dealing defense can be invoked to avoid liability for copyright infringement. The defense may be particularly successful where the use is non-commercial and for purposes of scholarly research. For a more detailed discussion of the availability of the fair dealing defense, see A.1.

Conclusion:

Works that are not protected by copyright can be subject to text data mining, freely and without restrictions. For copyright-protected works, NLP researchers should look for text data mining policies, terms of use, etc. which explicitly grant the right to use the particular work for the purpose of text data mining to ensure that such usage of the work is permitted. Pursuant to such terms of agreement, NLP researchers are legally authorized by the copyright owners to text data mine the latters' works. However, absent such grant of authorization, NLP researchers may proceed to use the copyrighted works for the purposes of text data mining and then raise a fair dealing defense, should a copyright infringement claim be brought against them by the copyright owners.

2. What are the limits regarding using this data for training a language model? Can NLP researchers redistribute the data afterwards?

Applicable legislation:

- *Copyright Act*, RSC, 1985, c C-42

Analysis:

Use of the data, including for training a language model and redistribution, may be constrained by the terms of use or other similar agreement. For instance, Canadian Science Publishing's Text and Data Mining Policy sets out restrictions for how its works may be used:

Restrictions

Except as expressly stated in this policy or otherwise permitted in writing by CSP, Users may ***NOT*** *carry out the following:*

- *perform systematic or substantive extracting for the purposes of creating a product or service for use by third parties, or that has the potential to substitute and (or) replicate any other existing CSP product, service, and(or) solution;*
- *create any form of central repository containing CSP content except as described above for the purpose of specific [text and data mine ("TDM")] projects;*
- *make the results of any TDM Output available on an externally facing server or website, except as permitted by an approved API service;*
- *allow a third party to harvest any TDM Output to an internal server;*
- *extract, develop, or use CSP content in any direct or indirect commercial activity;*
- *abridge, modify, translate, or create any derivative work based on CSP content, except to the extent necessary to make it perceptible on a computer screen to the User for research purposes;*
- *remove, obscure, or modify in any way any copyright notices, other notices or disclaimers as they appear in CSP content;*
- *use any robots, spiders, crawlers, or other automated downloading programs, algorithms, or devices to continuously and automatically search, scrape, extract, deep link, index, or disrupt the working of CSP's content hosting platform;*

- *reproduce any illustrations, including photographs, figures, and line drawings, in the TDM Output without the consent of the rights holder (unless permitted under the article-level license).*

If NLP researchers choose not to adhere to a policy which prohibits using the data for language model training and/or distribution and proceed to use the data in those ways, they may be committing copyright infringement. That said, liability could be avoided via the fair dealing defense.

Conclusion:

The specific uses of mined data, including for training a language model and redistribution, may be explicitly permitted or prohibited by the copyright owner via a policy, terms of use, etc. Absent an express grant of permission, the use of the data to train a language model and/or for redistribution may constitute copyright infringement; that said, liability may be neutralized by a successful fair dealing defense.

3. Do the terms and conditions of Twitter, Facebook, Youtube, etc., tell us whether NLP researchers can collect data from them for a project such as BigScience? What are some of the risks raised by directly collecting data from these social media platforms?

Applicable legislation:

- *Copyright Act*, RSC, 1985, c C-42

Analysis:

For the purpose of answering the above question, Instagram's Terms of Use and Data Policy are considered. They are used as examples because Instagram is a major social media platform whose policies can be assumed to set a best practice standard and serve as a template upon which other smaller platforms may base their policies, such that their data privacy policies and related terms of use read very similarly to Instagram's.

Instagram's Terms of Use addresses copyright ownership in its users' content as follows:

We do not claim ownership of your content, but you grant us a license to use it.
Nothing is changing about your rights in your content. We do not claim ownership of your content that you post on or through the Service and you are free to share your content with anyone else, wherever you want. However, we need certain legal permissions from you (known as a "license") to provide the Service. When you share, post, or upload content that is covered by intellectual property rights (like photos or videos) on or in connection with our Service, you hereby grant to us a non-exclusive, royalty-free, transferable, sub-licensable, worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of your content (consistent with your privacy and application settings). This license will end when your content is deleted from our systems. You can delete content individually or all at once by deleting your account. To learn more about how we use information, and how to control or delete your content, review the Data Policy and visit the Instagram Help Center.

In short, while users retain copyright ownership in their content, they permit Instagram (and likely other social media platforms) to, among other things, distribute this content. This suggests that NLP researchers can collect users' content from platforms directly, without first obtaining permission from each individual user.

That user content can be acquired directly from Instagram, bypassing users' consent, is supported by Instagram's Data Policy, which states:

Sharing with Third-Party Partners

We work with third-party partners who help us provide and improve our Products or who use Meta Business Tools to grow their businesses, which makes it possible to operate our companies and provide free services to people around the world. We don't sell any of your information to anyone, and we never will. We also impose strict restrictions on how our partners can use and disclose the data we provide. Here are the types of third parties we share information with:

. . .

Researchers and academics.
We also provide information and content to research partners and academics to conduct research that advances scholarship and innovation that support our business or mission, and enhances discovery and innovation on topics of general social welfare, technological advancement, public interest, health and well-being.

Assuming that NLP researchers fall within Instagram's definition of "research partners and academics," they will be able to obtain user content directly from Instagram for the purposes of text data mining and developing a language model.

However, receiving content directly from social media platforms poses its own risks for NLP researchers because the content and data collected by platforms, all of which could potentially be provided to the NLP researchers without first being culled for relevance to the NLP researchers' reasons for requesting user content in the first place, are extensive and could be perceived as an invasion of privacy. For instance, the types of information collected, as outlined in Instagram's Data Policy, include "Things you and others do and provide"; "Device Information"; and "Information from partners." "Things you and others do and provide" is particularly pertinent in the context of NLP research, and its most relevant subcategories are provided below:

Things you and others do and provide.

- *Information and content you provide. We collect the content, communications and other information you provide when you use our Products, including when you sign up for an account, create or share content, and message or communicate with others. This can include information in or about the content you provide (like metadata), such as the location of a photo or the date a file was created. . . .*

- *Things others do and information they provide about you. We also receive and analyze content, communications and information that other people provide when they use our Products. This can include information about you, such as when others share or comment on a photo of you, send a message to you, or upload, sync or import your contact information.*

Privacy concerns on the part of NLP researchers can be reduced by collecting content/information that has been made public by users. Further reading of the Data Policy suggests that users' public information is open to access and use by external third parties:

III. How is this information shared?

Your information is shared with others in the following ways:

Sharing on Meta Products

People and accounts you share and communicate with
When you share and communicate using our Products, you choose the audience for what you share. For example, when you post on Facebook, you select the audience for the post, such as a group, all of your friends, the public, or a customized list of people. . . .

Public information can be seen by anyone, on or off our Products, including if they don't have an account. This includes your Instagram username; any information you share with a public audience; information in your public profile on Facebook; and content you share on a Facebook Page, public Instagram account or any other public forum, such as Facebook Marketplace. You, other people using Facebook and Instagram, and we can provide access to or send public information to anyone on or off our Products, including in other Meta Company Products, in search results, or through tools and APIs. Public information can also be seen, accessed, reshared or downloaded through third-party services such as search engines, APIs, and offline media such as TV, and by apps, websites and other services that integrate with our Products.

. . .

Content others share or reshare about you
You should consider who you choose to share with, because people who can see your activity on our Products can choose to share it with others on and off our Products, including people and businesses outside the audience you shared with. For example, when you share a post or send a message to specific friends or accounts, they can download, screenshot, or reshare that content to others across or off our Products, in person or in virtual reality experiences such as Facebook Spaces. Also, when you comment on someone else's post or react to their content, your comment or reaction is visible to anyone who can see the other person's content, and that person can change the audience later.

(Underlines added.)

Given that users have constructively been notified about the accessibility of public content, they have no legitimate claim to the right of privacy over the information that they have made (whether intentionally or inadvertently) public.

Notwithstanding that NLP researchers may be able to gain access to social media content, whether they can (arguably) replicate it as a derivative work of the social media user/content creator without infringing the user/creator's copyright in the content is up for debate. Recall that Instagram's Terms of Use permitted Instagram to "host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of [its users'] content" – this grant of rights does not extend or transfer to NLP researchers simply because user content is provided to them pursuant to the Data Policy or by virtue of the fact that the content/information is public. However, should a user bring a copyright infringement claim against the NLP researchers for the use of their content for the purposes of training and developing a language model, a fair dealing defense may be raised to ward off liability.

Conclusion:

Based on a review of one social media service provider's (i.e., Instagram) Terms of Use and Data Policy, it appears that NLP researchers may be able to access social media users' content for the purposes of their work. That said, obtaining user content/information directly from service providers may raise privacy concerns, given the abundance of content and data collected by social media platforms and potentially provided to NLP researchers upon request. Moreover, users retain copyright ownership in their content and use by NLP researchers for the purposes of developing a language model may give rise to a copyright infringement claim. If such a claim is brought, the fair dealing defense may help to shield NLP researchers from liability.

D. Privacy Questions

1. Overview of relevant privacy issues

Applicable legislation:

- *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5
- Bill C-11: *Digital Charter Implementation Act, 2020*
- Alberta: *Personal Information Protection Act*, SA 2003, c P-6.5
- British Columbia: *Personal Information Protection Act*, SBC 2003, c 63
- Quebec: *Act respecting the Protection of Personal Information in the Private Sector*, CQLR, c P-39.1

Analysis:

Personal Information Protection and Electronic Documents Act ("PIPEDA") & Provincial Statutes

PIPEDA applies to private-sector organizations in Canada that collect, use or disclose personal information in the course of a commercial activity (paragraph 4(1)(a) of PIPEDA). Subsection 2(1)

of PIPEDA defines “commercial activity” as “any particular transaction, act or conduct or any regular course of conduct that is of a commercial character, including the selling, bartering or leasing of donor, membership or other fundraising lists.”

Notably, case law indicates that not-for-profit businesses/organizations are not automatically exempt from PIPEDA. The not-for-profit status of a business/organization for purposes of taxation is not determinative of whether its collection, use or disclosure of personal information is carried out in the course of commercial activity (*Rodgers v Calvert*, 2004 ON SC (CanLII)). However, an organization's collection, use or disclosure of personal information solely for journalistic, artistic or literary purposes is exempt from the application of PIPEDA.

All businesses that operate in Canada and deal with personal information that crosses provincial or national borders in the course of commercial activities are subject to PIPEDA, regardless of the province or territory in which they are based, including provinces with substantially similar legislation that would typically override PIPEDA in relation to *intraprovincial* privacy matters. These provinces are Alberta (*Personal Information Protection Act*, SA 2003, c P-6.5), British Columbia (*Personal Information Protection Act*, SBC 2003, c 63), and Quebec (*Act respecting the Protection of Personal Information in the Private Sector*, CQLR, c P-39.1). Like PIPEDA, these provinces’ privacy laws apply to the collection, use, and disclosure of personal information in the public sector. Unlike PIPEDA, these statutes apply regardless of whether an activity is commercial in nature.

Consent, Reasonable Purpose, Anonymization

Canada’s privacy legislation is founded on meaningful consent and reasonable purpose: processing information about an identifiable individual requires meaningful, informed consent, but even with such consent, the collection, use or disclosure of personal information must be for a reasonable purpose.

Consent must be obtained prior to the collection, use, and disclosure of personal information. Consent may be express or implied, depending on the circumstances; the intended collections, uses, and disclosures; and the level of sensitivity of the information. Implied consent is generally not appropriate and insufficient for sensitive personal information (e.g., health and financial information). Consent is only valid if it is reasonable to expect that the individual to whom the personal information relates would understand the nature, purpose, and consequences of the collection, use, or disclosure of the personal information to which they are consenting.

Consent is generally not required for the collection, use, and disclosure of certain publicly available information (e.g., published information, court decisions, etc.), so long as the collection, use, or disclosure is related to the purpose for which the information is publicly available (see PIPEDA Regulations).

Personal information collected by an organization may only be used for the purposes for which it was collected. If an organization intends to use it for another purpose, they must obtain consent from the person to whom the information pertains in relation to the new intended use.

Because obtaining meaningful consent and satisfying the reasonable purpose requirement is arguably difficult in the AI context, organizations in Canada are increasingly seeking to limit the application of privacy laws via "anonymization" of the data that their AI programs/systems require. There is, however, a question as to when true anonymity is achieved.

Safeguards

Organizations are required to use security safeguards to protect personal information from loss or theft, as well as to prevent unauthorized access, disclosure, copying, use or modification. The applicable legislation does not identify specific security safeguards that must be used, requiring instead that the nature of the safeguards be appropriate to the degree of sensitivity of the information that has been collected; the amount, distribution, and format of the information; and the method of storage.

Organizations must also educate their employees about the importance of maintaining the confidentiality of personal information, and ensure that care is used in the disposal or destruction of personal information to prevent unauthorized parties from obtaining access to the information.

Transfers

Organizations are responsible for personal information in their possession or custody, including information that has been transferred to an external third-party for processing. Principle 1 of PIPEDA states that "[a]n organization is responsible for personal information in its possession or custody, including information that has been transferred to a third party for processing. The organization shall use contractual or other means to provide a comparable level of protection while the information is being processed by a third party." This has been interpreted as requiring, among other things, a contract which includes a requirement for the third-party processor to have privacy policies and processes in place, including training for its staff and effective security measures; a requirement that the information be properly safeguarded at all times; and a right to audit and inspect how the third-party processor handles and stores personal information.

There has been suggestion from the Office of the Privacy Commissioner of Canada that consent may be required for transfers to service providers. This view is controversial, and no legislative action effecting this position has taken place to date. That said, consent - or, at minimum, notice to the individual to whom the personal information pertains - may be a best practice that organizations choose to implement to reduce risk and avoid liability.

Data Subjects' Rights

As previously noted, Canadian private-sector privacy laws generally require the knowledge and consent of the individual whose personal information is being collected and/or used. Organizations must be open and transparent about their practices, and inform individuals about the information collected, used, and disclosed, as well as the purposes for such activities, among other requirements.

Individuals have a general right to obtain access to their personal information held by organizations, and to require organizations to amend inaccurate/incomplete information. Individuals also have the right to submit complaints to organizations, to withdraw consent (subject to some limitations), and to file complaints with the body overseeing and enforcing privacy legislation.

Bill C-11: Digital Charter Implementation Act, 2020

Recently, the Government of Canada introduced new privacy legislation for the private sector: Bill C-11, referred to as the *Digital Charter Implementation Act, 2020*, which in turn proposes the enactment of new privacy statutes. One such statute is the *Consumer Privacy Protection Act* ("CPPA"), which would replace the privacy provisions of the PIPEDA.

As applicable to AI and the work of NLP researchers, the proposed legislation proposes as follows:

- Allow an organization to use an individual's personal information without their knowledge or consent for the organization's internal research and development purposes, so long as the information has been de-identified prior to use.
- Create a limited exception permitting the disclosure of de-identified information for socially beneficial purposes.
- Make it an offence to use de-identified information alone or in combination with other information to identify an individual, except for the purposes of testing the efficacy of security safeguards that the organization has implemented to protect the information.

2. An NLP researcher's privacy risk – quick questions and answers

For each of the following questions, it is assumed that an NLP researcher's Canadian branch/office will not be based in either Alberta, British Columbia or Quebec. Accordingly, all answers are based on PIPEDA as the relevant source of law.

- Q: Are there legal concerns around accessing web content that have personal information? What about datasets?

  A: It is difficult to assess whether an NLP researcher is subject to PIPEDA (and any succeeding/amending legislation), as NLP researchers/AI engineers would not directly be soliciting personal information from individuals. Rather, they intend to access and use web content or publicly available datasets that contain personal information that has been collected by a third party or that has been provided online voluntarily by the individuals to whom the personal information pertains. Moreover, PIPEDA applicability to an NLP researcher is contingent on the latter's use of personal information being deemed "in course of a commercial activity." Whether its use/work constitutes "commercial activity" is not clear.

- Q: What about distributing such datasets, including across borders?

A: If an NLP researcher's use goes beyond accessing personal information for the purposes of creating a dataset to include distribution and such distribution is construed as a "commercial activity," as a business that operates in Canada and deals with personal information that crosses provincial or national borders, an NLP researcher would be subject to PIPEDA.

- Q: Are there legal concerns around publishing a pre-trained language model that is trained on a subset of common crawl, a published dataset based on a crawl of the web in terms of personal information?

  A: If the publishing is considered use of personal information in the course of commercial activity, then an NLP researcher would be required to observe PIPEDA and any succeeding/amending legislation.

- Q: Are there concerns around distributing models that have personal information stored within it? That can expose such personal information?

  A: If distribution of the models is considered use of personal information in the course of commercial activity, then an NLP researcher would be required to observe PIPEDA and any succeeding/amending legislation.

- Q: What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exceptions for research purposes?

- A: NLP researchers may wish to implement anonymization of personal information in all stages of developing their language model, particularly as it is unclear whether an NLP researcher's various uses throughout model construction/development constitutes a "commercial activity" such that PIPEDA applies. It would appear that a research purpose must be strictly non-commercial in nature to avoid the application of PIPEDA.

- Q: How does consent of the individuals affect what an NLP organization can and cannot do?

  A: Under PIPEDA, the processing of personal information requires meaningful, informed consent; even where consent is obtained, the collection, use or disclosure of personal information must be for a "reasonable purpose." As such, if an NLP researcher's uses are considered commercial activities, an NLP researcher would still be subject to the PIPEDA and required to meet all of the requirements outlined therein, despite having obtained consent.

- Q: What are the privacy risks related to data collection directly from persons? E.g., when you interview people or they donate data, etc.

A: In such a scenario, an NLP researcher would be directly subject to PIPEDA, if its uses of the personal information collected are deemed to be “commercial activities,” and must meet all requirements set out therein.

CHINA

Prepared by: Chenxi Zhou, Dian Yu

Yifan Xu, Yingxin Xu

Zhe Tan, Zifan Ye

The Introduction of the Chinese Legal System

The Chinese Legal System has several components: 1. the Constitution, 2. the laws and judicial interpretations, 3. the regulations, and 4. judicial decisions. For the purpose of this playbook, China only includes Mainland China without the Hong Kong SAR, Macau SAR, and Taiwan Region.

1. The Constitution

The Constitution is the core of the Chinese Legal System, and it is the country's fundamental law. The Constitution has supreme legal authority in the Chinese Legal System. All laws, administrative and local regulations must be made in accordance with the Constitution and must not contravene the Constitution. The Constitution stipulates that the National People's Congress (NPC) and National People's Congress Standing Committee (NPCSC) exercise the state's legislative power.

Currently, there is no provision in the Constitution directly related to data or artificial intelligence.

2. The Laws and Judicial Interpretations

The laws enacted by the NPC and NPCSC establish the important and basic legal systems in the construction of the nation's economic, political, cultural, social, and ecological civilization. They constitute the main body of the Chinese Legal System. The laws enacted by the NPC and NPCSC must not be contravened by regulations.

There are many provisions in laws related to the topic. For example, Article 253 (i) of the *Criminal Code* incriminates illegally obtaining personal information. Article 1034-39 of the *Civil Code* generally regulates obtaining and using personal information.

Many laws were also recently enacted or amended related to the topic of this playbook. The *Cybersecurity Law*, effective on June 1, 2017, is about the construction, operation, maintenance, and use of the network as well as the supervision and administration of cybersecurity in China. In particular, Article 37 of the Cybersecurity Law has a data local storage requirement for Critical Information Infrastructure (CII) operators and prohibits the unauthorized cross-border transmission of these data. The *Copyright Law*, the latest amendment effective on June 1, 2021, is about the limitations of copyright and its protection in China. The *Data Security Law*, effective on Sep. 1, 2021, is about regulating the utilization of data. Particularly, Article 31 of the *Data Security Law* authorized the relevant department to formulate regulations on cross-border transmission of some important data. The *Personal Information Protection Law*, effective on Nov. 1, 2021, is about protecting personal information and regulating the utilization of personal information. Especially, Chapter 3 of the *Personal Information Protection Law* includes the rules related to the cross-border provision of personal information.

Judicial interpretation is also an important source of law in China. The laws grant the Supreme People's Court and the Supreme People's Procuratorate the power to interpret the specific governing laws and procuratorial work. Judicial interpretation has universal judicial

effects, and courts can directly quote judicial interpretations as the basis for judgment. Meanwhile, it shall not contravene the Constitution and the laws.

There are some judicial interpretations related to this topic. For example, the Supreme People's Court amended the *Provisions of the Supreme People's Court on Several Issues concerning the Application of Law in the Trial of Cases involving Civil Disputes over Infringements upon Personal Rights and Interests through Information Networks* in 2020, to address the issues related to personal rights online infringement cases. Also, the Supreme People's Court issued the *Provisions of the Supreme People's Court on Several Issues concerning the Application of Law in the Trial of Civil Cases Relating to Processing of Personal Information by Using the Facial Recognition Technology* in 2021, to provide guidance for cases related to the processing of personal information by using the facial recognition technology.

3. The Regulations

There are four different kinds of regulations in China, 1) administrative regulations, 2) ministerial regulations, 3) local regulations, and 4) other regulations.

1) Administrative Regulations

The State Council, the central government of China, is the highest administrative organ of the state. It formulates administrative regulations in accordance with the Constitution and laws. Administrative regulations may regulate matters concerning the implementation of the provisions of the laws and the performance of the administrative functions and powers of the State Council. For matters to be governed by laws formulated by the NPC and NPCSC, the State Council may enact administrative regulations first in its place with authorization from the NPC and NPCSC.

For now, only one administrative regulation has been formulated in this area. Based on the provisions in the *Cybersecurity Law*, the State Council issued the *Regulation on Protecting the Security of Critical Information Infrastructure* to implement relevant provisions in the *Data Security Law* in 2021.

2) Ministerial Regulations

The ministries and commissions and directly affiliated institutions with the administrative functions of the State Council, the People's Bank of China, and the National Audit Office may formulate regulations within the scope of their functions and powers and in accordance with the law and the administrative regulations.

Many ministerial regulations have been or will be formulated related to this topic. As early as 2013, the Ministry of Industry and Information Technology issued the *Provisions on Protecting the Personal Information of Telecommunications and Internet Users* to protect personal information. In 2019, the Cyberspace Administration issued the *Provisions on the Cyber Protection of Children's Personal Information* to further strengthen the protection of children's personal information. Currently, the drafts of the *Regulations on Network Data Security Management* and the *Measures for Data Security Management* have been published recently for public comments.

They will serve as the implementing regulations for the *Cybersecurity Law* and the *Data Security Law*.

3) Local Regulations

In accordance with the Constitution and laws, the people's congresses and their standing committees of the provinces, autonomous regions, municipalities, cities with subordinate districts, and autonomous prefectures may formulate local regulations. The people's congresses of the ethnic autonomous areas have the power to formulate autonomous and separate regulations on the basis of the political, economic, and cultural characteristics of the local ethnic group(s). The people's congresses and their standing committees of the provinces and cities where special economic zones are located may, upon authorization by the NPC, formulate and enforce regulations within limits allowed by the special economic zones.

2021 witnessed two newly issued local regulations related to data, the *Shanghai Data Provisions* and the *Shenzhen Special Economic Zone Data Provisions*. They contain specific rules related to personal information data and data transactions.

4) Other Regulations

The people's governments of provinces, autonomous regions, municipalities, cities with subordinate districts, and autonomous prefectures may formulate regulations in accordance with the law, the administrative regulations, and applicable local regulations.

4. Judicial Decisions

Although China is not a country that practices case law, some judicial decisions nevertheless have significance as guides for judicial practice, such as a) the guiding cases published by the Supreme People's Court, b) decisions issued by the Supreme People's Court, c) decisions issued by the High People's Court of provinces, autonomous regions, municipalities, and d) decisions issued by courts of higher level. They provide important references for judges dealing with similar cases, and therefore, provide guidance for future legal issues.

A. IP Questions

1. Are the data training sets and models protected by IP rights and if so which IP rights?

Data training sets are made up of data. Whether data itself can be protected by copyright is debatable in China. According to the Article 3 of the *Copyright Law*[2], works shall be original to be copyrightable. Most data cannot meet the originality requirement, and therefore cannot be protected by copyright. Similarly, whether data training sets can be protected by copyright is also controversial in China. Data training sets is the compilation of the data. According to Article 15 of the *Copyright Law*, a work created by compilation of data shall be copyrightable if the choice or layout of the contents of which embodies the original creation. It's debatable whether the choice or layout of the data training sets are original to meet the threshold of copyright law. Therefore, it may depend on the fact of a specific case.

Data training sets can also be protected by trade secrets. According to the Article 9 of the *Anti-Unfair Competition Law*[3], "trade secret" means 1) technical, operational or other commercial information unknown to the public and 2) is of commercial value for which the right holder has taken corresponding confidentiality measures, and a business shall not infringe others' trade secrets. As long as the data training sets are unknown to the public, they can meet the other requirements of trade secrets. It's important to keep the data training sets as confidential to enjoy the protection offered by the *Anti-Unfair Competition Law*.

Data training models may be protected by copyright. According to Article 5 of the *Regulation on Computers Software Protection*[4], computer software is protected by copyright. Particularly, Chinese citizens, legal entities or other organizations enjoy copyright in the software which they have developed, **whether published or not**; in comparison, foreigners or stateless persons' software needs to be **first published in China** enjoy copyright, unless otherwise protected by international treaties.

Also, according to Article 2 of the *Patent Law*[5], the term "invention" refers to any new technical solution relating to a product, a process, or an improvement thereof. Therefore, particular methods used in the data training models can be protected by patent law. However, algorithm, by itself, are unlikely to be protected by patents because of exclusions of scientific discoveries or rules and methods for mental activities[6].

[2] PRC Copyright Law: 中华人民共和国著作权法
[3] PRC Anti-Unfair Competition Law: 中华人民共和国反不正当竞争法
[4] Regulation on Computer Software Protection: 计算机软件保护条例

[5] PRC Patent Law: 中华人民共和国专利法
[6] Article 25.1-2 of PRC Patent law

2. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?

See A1, publishing pre-trained language models based on the dataset does not infringe copyright if the dataset is not protected by copyright. On the other hand, if the data set is protected by copyright, the pre-trained language model may infringe on technical measures, right of adaptation, right of translation, right of reproduction, right of dissemination, etc. See A11 for details. However, one may argue fair use of copyrighted works under *Article 24.6 of the PRC Copyright law*, provided that such use shall not conflict with a normal exploitation of the work, or unreasonably prejudice the legitimate rights and interests of the copyright owner.

3. Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g. HTML tags or document structure)?

Publishing datasets that contain HTML tags or document structure might get involved with infringement of the right to communicate works to the public over information networks, especially when such HTML tags insert pictures or hyperlink into the text.

The concept of hyperlink is not defined in any international treaty, and there has been much discussion in China about whether hyperlink in text would infringe the right to communicate works to the public over information networks.

There are two types of hyperlinks. One is surface link, which can directly link to the homepage of the linked website. The other is deep link, which can be further classified as two categories. One is the normal deep link, which can link to the secondary page of the linked website without changing the content of the linked website. The controversial one is called special deep link, through which, internet users may be able to view the content of other website (e.g., videos, goods) without jumping to another website, resulting in the user mistakenly thinking that the content of the page is self-linked when viewing the content, which usually exists as a media file inserted in the web page.

The majority view of scholars is that surface links and normal deep links are permitted within the PRC copyright legal framework, while special deep links may cause direct infringement of the right to communicate works to the public over information networks.[7]

4. What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?

Please refer to the answer of question A3.

5. Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR)

[7] *See* Liu Yinliang, *An Inquiry Into The Legal Nature of Deep Links Within The Framework of The Right of Communicate Works to The Public Over Information Networks*, Global Law Review, Vol.03, 2018.

Article 2 of the *Regulation on Computers Software Protection* (“RCSP”) states that “computer software” means computer programs and the relevant documentation. Article 3 of the RCSP especially provides that “computer program” refers to coded instructional sequences or those symbolic instructional sequences or numeric language sequences which can be automatically converted into coded instructional sequences—which are for the purpose of obtaining a certain result and which are operated on information processing equipment such as computers. The source code program of a piece of computer software and its object code program should be regarded as one work.

In the broad sense, C4 or OSCAR text could be understood as such a computer program as referred to in the RCSP. Therefore, to publish a dataset that refers to locations in another dataset, even with restricted availability (e.g., C4/OSCAR) should note the potential infringement of protected right over computer software.

6. What are the IP risks related to data collection directly from persons? For example when you interview people or they donate data etc.

The main IP risk is copyright risk in this scenario.

In the interview example, the data collector may violate the copyright of the interviewee.

According to Article 3 of the *Copyright Law*, oral work is copyrightable. Therefore, the working product, such as the interview transcript or the recording, has copyright.

Depending on the different levels of input from the interviewer and the interviewee, the ownership of the copyright may be different. For example, if the interview transcript only copies what the interviewee said, the interviewee will likely be the sole owner of the oral work. If the interview transcript copies both the interviewer and the interviewee, they will be likely to share the ownership of the oral work.

In the scenario of sole ownership by the interviewee, the subsequent use of the data is likely to infringe the interviewee’s copyright. In this case, the interviewee may have the right to sue the infringement for applicable damage or injunction.

In the scenario of shared ownership, it’s more complicated. According to Article 14 of the *Copyright Law*, the copyright of a cooperative work shall be exercised by co-authors upon consensus. Where no consensus has been reached, and there is no justified reason, no party shall prevent another party from exercising rights other than transferring and permitting others’ exclusive use and pledging, but the proceeds obtained shall be reasonably distributed to all co-authors. Therefore, the subsequent use may not be prohibited, but the interviewee may have legal title to proceed, if any. In this case, the interviewee may have the right to sue for the proceedings.

The best way for the data collector to avoid any subsequent dispute is to obtain consent from the interviewee that the interviewee voluntarily waives his copyright or transfers his rights to the data collector, and the data collector shall pay adequate consideration to the interviewee.

In the data donation example, the data collector shall make sure that the donor has legal title to the data. The best way is to conduct proper legal due diligence and sign an agreement with relevant representation.

Sometimes, NLP researchers can choose to use an online agreement to have some protection. It works sometimes, and it will be better than nothing, but it may not be the best option. An agreement can offer only ex post protection against the risks mentioned in the question. However, if the donor does not have legal title to the data, NLP researchers will be the ones infringing other's rights, and they will be sued directly (most likely). Sure they have the option to recover their losses afterward based on the online agreement, but it will still a long process, with significant uncertainty. It will be a huge deal for the researchers. The point of a legal due diligence is to make sure that such risk will not happen at all. Therefore, if the legal risk is significant, I think an agreement is not sufficient.

On the other hand, professional legal due diligence may cost a lot, and the risk, in some cases, may not worth the money. Therefore, please consider the risk to determine if legal due diligence is needed.

Meanwhile, please refer to the answer of question A1 that the copyright of data is still debatable.

7. Can NLP researchers train a language model on borrowed library books - e.g., from the Internet Archive or other online book lender?

Most likely, no.

A person does not enjoy the copyright to the borrowed book and, therefore, has no right to copy its contents. The NLP process is likely to include copying and therefore prohibited by copyright law.

One possible exception is that one can use a published work for the purposes of their own private study, research, or self-entertainment. However, for organization researchers, there is no such exception, and such behavior will constitute infringement.

8. How do rights on the source data (e.g. copyright) transfer to the trained model?

Please refer to the answer of question A10.

9. Are there any other IP-related rights for data? e.g., trade secret.

Business data may be protected by trade secret rules.

Under Article 9 of the Anti-Unfair Competition Law of the People's Republic of China (2019 Amendment), For the purpose of this Law, "trade secret" means technical, operational, or other commercial information unknown to the public and is of commercial value for which the right holder has taken corresponding confidentiality measures. According to Article 9 of the interpretation of the Supreme People's Court on Several Issues concerning the application of law

in the trial of civil cases of unfair competition (implemented on January 1, 2021), if the relevant information is not generally known and easily available to relevant personnel in their field, it shall be recognized as "not known to the public" as stipulated in paragraph 3 of Article 10 of the anti-unfair competition law.

Under any of the following circumstances, the relevant information may be deemed not to be unknown to the public:

(1) The information is the general knowledge or industry practice of the person in the technical or economic field to which it belongs;

(2) This information only involves the size, structure, materials, simple combination of components, and other contents of the product. After entering the market, the relevant public can directly obtain it by observing the product.

(3) The information has been publicly disclosed in public publications or other media;

(4) The information has been disclosed through public reports, exhibitions, etc;

(5) The information is available from other public sources;

(6) This information is easily available at no cost.

Article 10: if the relevant information has real or potential commercial value and can bring competitive advantage to the obligee, it shall be recognized as "capable of bringing economic benefits to the obligee and practical" as stipulated in paragraph 3 of Article 10 of the anti-unfair competition law.

Article 11: the reasonable protective measures taken by the obligee to prevent information leakage in accordance with its commercial value and other specific circumstances shall be recognized as "confidentiality measures" stipulated in paragraph 3 of Article 10 of the anti-unfair competition law.

The people's court shall determine whether the obligee has taken confidentiality measures according to the characteristics of the information carrier involved, the obligee's willingness to keep confidential, the identifiable degree of confidentiality measures, the difficulty of others obtaining them through legitimate means and other factors.

Under any of the following circumstances, the obligee shall be deemed to have taken confidentiality measures if it is sufficient to prevent the disclosure of confidential information under normal circumstances:

(1) Limit the scope of knowledge of confidential information and only inform relevant personnel who must know its contents;

(2) Take locking and other preventive measures for confidential information carriers;

(3) The carrier of confidential information is marked with confidentiality marks;

(4) Adopt password or code for confidential information;

(5) Sign confidentiality agreement;

(6) Restrict visitors or put forward confidentiality requirements for confidential machines, factories, workshops and other places;

(7) Other reasonable measures to ensure the confidentiality of information.

To sum up, data can enjoy trade secret rights if it can meet the three requirements stipulated by the Anti-unfair competition Law and relevant judicial interpretations: unknown to the public, of commercial value, and the right holder has taken corresponding confidentiality measures. According to Article 9 of the Anti-Unfair Competition Law of the People's Republic of China (2019 Amendment), a business shall not commit the following acts of infringing upon trade secrets:

(1) Acquiring a trade secret from the right holder by theft, bribery, fraud, coercion, electronic intrusion, or any other illicit means.

(2) Disclosing, using, or allowing another person to use a trade secret acquired from the right holder by any means as specified in the preceding subparagraph.

(3) Disclosing, using, or allowing another person to use a trade secret in its possession, in violation of its confidentiality obligation or the requirements of the right holder for keeping the trade secret confidential.

(4) Abetting a person, or tempting, or aiding a person into or in acquiring, disclosing, using, or allowing another person to use the trade secret of the right holder in violation of his or her non-disclosure obligation or the requirements of the right holder for keeping the trade secret confidential.

An illegal act as set forth in the preceding paragraph committed by a natural person, legal person, or unincorporated organization other than a business shall be treated as infringement of the trade secret.

Where a third party knows or should have known that an employee or a former employee of the right holder of a trade secret or any other entity or individual has committed an illegal act as specified in paragraph 1 of this Article but still acquires, discloses, uses, or allows another person to use the trade secret, the third party shall be deemed to have infringed upon the trade secret.

10. How is the ownership attribution of database rights?

The attribution of data-based rights should be discussed categorically: (1) personal data, (2) government data, and (3) corporate data.

First, in the context of personal data, relevant fundamental data rights belong to individuals. Article 44 of the PRC Personal Information Protection Law[8] specifies that individuals have the

[8] PRC Personal Information Protection Law: 中华人民共和国个人信息保护法

right to know and the right to decide relating to their personal information, and have the right to limit or refuse the handling of their personal information by others. In addition, according to the protective mechanism provided by the PRC Civil Code[9], rights of personal data include property rights and spiritual rights.

Second, in the context of government data, relevant data rights come naturally with public goods attributes. Government data is data collected by government authorities in the process of performing their duties (such as approval, filing, etc.). Such data may include population data, market regulation data, financial data, social governance data and so on. Government data is mainly collected by using state funds, and therefore has the attributes of public goods. The policy of using government data is to promote use to the public good. However, such use may be restricted to some extent. For the fair use of public data, please refer to question C.14.

Third, as for corporate data (here, “corporate data” means data collected for business purpose, that is to say, if a research organization created a dataset for business use, such dataset will be treated as “corporate data”), there is no black letter law on the ownership attribution of such data rights so far. In the practice in China, courts attempt to recognize and protect business interests in corporate data under the legal framework of competition law: (a) when a corporate has property rights of the data it invested in building a database, such a corporate enjoys property interests in data derivative products and database, (b) when the data service provided by a corporate meet the relevant needs of the public and increase consumer welfare, which is essentially a competitive right, the other corporations are prohibited from using the data to the extent of causing substantial alternative consequences, which is considered as infringement of the legitimate commercial interests of the corporate who builds the database, (c) when the data is shared and utilized between companies, such share and utilization shall, on the basis of protecting the individual rights of users, abide by the spirit of autonomous contracts and comply with the agreements between companies.

11. Does data mining infringe intellectual property rights?

China's existing copyright system cannot provide an exemption for text data mining, so text data mining is likely to infringe copyright and other related property rights.

Article 24 and Article 25 of the Copyright Law of the People's Republic of China stipulate the use of works without the permission of copyright owners and payment of remuneration, and text data mining is not one of them:

Article 24 In the following cases, a work may be exploited without the permission from, and without payment of remuneration to, the copyright owner, provided that the name or designation of the author and the title of the work are mentioned and the normal use of the work, or unreasonably damage the lawful rights and interests of the copyright owner shall not be affected:

(1) use of a published work for the purposes of the user's own private study, research or self-entertainment;

[9] PRC Civil Code: 中华人民共和国民法典

(2) appropriate quotation from a published work in one's own work for the purposes of introduction of, or comment on, a work, or demonstration of a point;

(3) inevitable reappearance or citation of a published work in newspapers, periodicals, radio stations, television stations or other media for the purpose of reporting news;

(4) reprinting by newspapers or periodicals or other media, or rebroadcasting by radio stations or television stations or other media, of the current event articles on the issues of politics, economy and religion, which have been published by other newspapers, periodicals, radio stations or television stations or other media, except where the copyright owner has declared that publication or broadcasting is not permitted;

(5) publication in newspapers or periodicals or other media, or broadcasting by radio stations or television stations or other media, of a speech delivered at a public assembly, except where the author has declared that publication or broadcasting is not permitted;

(6) translation, adaptation, compilation, and broadcasting or reproduction, in a small quality of copies, of a published work for use by teachers or scientific researchers in classroom teaching or scientific research, provided that the translation or reproduction is not published or distributed;

(7) use of a published work by a State organ within the reasonable scope for the purpose of fulfilling its official duties;

(8) reproduction of a work in its collections by a library, archive, memorial hall, museum, art gallery, art museum or similar institution, for the purpose of the display or preservation of a copy of the work;

(9) free of charge performance of a published work, that is, with respect to the performance, neither fees are charged from the public nor the remuneration is paid to the performers, nor the performance is for profit;

(10) copying, drawing, photographing, or video recording of an artistic work located or on display in a public place;

(11) translation of a work published by a Chinese citizen, legal entity or unincorporated organization, which is created in the national common language and characters, into a minority nationality language for publication and distribution within the country.

(12) providing published works for dyslexics in a barrier-free way through which they can perceive.

(13) other circumstances prescribed by laws and administrative regulations.

The provisions of the preceding paragraph shall apply to restrictions on copyright-related rights.

Article 25: Those who compile and publish textbooks for the purpose of implementing compulsory education or educational planning of the state may, without permission of copyright

owners, compile published fragments of works, short written works, musical works, or single art works, photographic works, or graphic works in the textbooks, however, they shall pay remunerations to copyright owners according to the provisions, and designate the names or designations of authors, and titles of works, and shall not infringe upon other rights enjoyed by copyright owners in accordance with this Law.

The provisions of the preceding paragraph shall apply to restrictions on copyright-related rights.

Therefore, text data mining may infringe the following copyrights.

First, text data mining may violate the technical measures of "contact-based control" by avoiding the account password of data set in the contact link of the sample. Article 49 of the Copyright Law stipulates that in order to protect copyright and copyright-related rights, the right holder may take technical measures. Without permission of the right holder, no organization or individual shall deliberately avoid or destroy the technical measures, manufacture, import or provide relevant devices or components for the public for the purposes of avoiding or destroying the technical measures, or deliberately provide technical services for others' avoidance or destruction of the technical measures, except under the circumstances under which avoidance is allowed as prescribed in the laws and administrative regulations. For the purpose of this Law, ‘technical measures' means effective technologies, devices or components used to prevent or restrict browsing or appreciation of works, performance, and audio and video recordings, or provision of works, performance, and audio and video recordings for the public via information networks without permission of the right holder. The so-called control contacts technical measures refer to the measures mentioned in the above law to prevent and restrict browsing and appreciation of works, performances, audio, and video recordings without the permission of the right holder. Database copyright owners often use contact-based technical measures such as account and password to protect databases, while text data mining may need to evade account and password in sample contact links, which may infringe technical measures.

Secondly, in the process of sample collection, the right to copy works may be infringed by evading the "control and utilization" of technical measures to copy works. The so-called control and utilization technical measures refer to the technical measures to prevent and restrict the provision of works, performances, audio, and video products to the public through the information network without the permission of the obligee as specified in the above-mentioned law. The copyright owner of the data set is likely to set up "control and utilization" technical measures such as monitoring abnormal traffic and Restricting Private downloading. In order to avoid such technical measures, text data mining may require indirect and short-term data collection. However, according to paragraph 5 of Article 10 of the copyright law, the right of reproduction, that is, the right to produce one or more copies of the work by means of printing, Xeroxing, rubbing, sound recording, video recording, duplicating, re-shooting, or digital way, etc. In other words, the right to copy includes the digitization of works, so text data mining is likely to infringe the right to copy.

Third, in the process of sample processing, the right of interpretation of works may be infringed due to "transcoding". The right of performing arts mentioned here mainly includes the right of adaptation and the right of translation. According to Article 10 of the copyright law, the

right of adaptation is the right to change work and create a new work with originality; The right of translation is the right to convert a work from one language to another. After collecting samples in text data mining, only when users "transcode" the samples into machine-readable structured data, can the computer system recognize, count, and analyze them. From the external manifestation of "transcoding" behavior, this behavior changes the manifestation of the original sample and will eventually be presented in new manifestations such as analysis results. The adaptation right of works emphasizes the innovation of the form of expression of the original works on the premise of maintaining the consistency of substantive content. It can be seen that the "transcoding" behavior belongs to the category of adaptation right. From the internal mechanism of the "transcoding" behavior, the generated structured data comes from the original sample, "transcoding" does not change the internal expression of the creative thought of the sample. This behavior is a special translation behavior and may be regarded as a translated "deductive work". Therefore, from the external performance and internal mechanism of the behavior, the "transcoding" behavior of TDM samples is homogeneous with the "adaptation" behavior and "translation" behavior of works. Therefore, the right of adaptation and translation of works may be violated in the process of dealing with TDM samples.

Finally, in the link of sample output, the right of dissemination of works may be infringed due to the public release of analysis results. Article 26 of the Regulations on the Protection of the Right to Information Network Communication[10] stipulates that the right to information network communication refers to the right to provide works, performances, or audio and video products to the public by wired or wireless means so that the public can obtain works, performances or audio and video products at the time and place selected by themselves. In the sample output link, text data mining may enable the public to obtain the analysis results containing the contents protected by copyright law, thus infringing the right of communication.[11]

In addition, there is no exception for organizational research. According to Article 24 of the PRC Copyright Law, the exception is for personal study or research only.

In conclusion, text data mining is likely to infringe copyright.

Other than copyright infringement, there are often restrictions based on contract laws. For instance, the terms of use of a website may prohibit any use any robot, bot, script, spider, scraper, crawler, data mining, gathering, or extraction tools, or other automated means to access or copy any content. So even if you may argue “fair use” to avoid any copyright infringement, please keep in mind that you should always read the Terms of Use, and make sure that your text data mining does not breach any terms.

The *draft Data Security Management Measures*[12] released on November 14, 2021, requires the data handlers to assess the effect on online services brought about by their

[10] Regulations on the Protection of the Right to Information Network Communication: 信息网络传播权保护条例

[11] See Ma Zhiguo, Zhao Long, Text and Data Mining's Impact on Copyright Exception System and Countermeasures, Journal of Northwest Normal University (Social Sciences), Vol.58, 2021.

[12] http://www.cac.gov.cn/2021-11/14/c_1638501991577898.htm

nature and functions and not to interfere with the regular functioning of online services when they adopt automated tools to access or collect data. (*Article 17*)

B. Licensing Questions

1. Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?

Under the current legal framework of the People's Republic of China ("PRC", for the purpose of this document, excluding Hong Kong, Macao and Taiwan), both data sets and models are generally licensable as there is no specific restriction. In other words, the owner of the data sets and models can contractually license others to use the data sets and models. Similar to the license of IP, restrictions listed as below may be considered in the license contract.

1) Purpose of use;

2) Scope of use;

3) Compliance with laws;

4) Exclusivity; and

5) Sublicense.

2. What are the relationships between laws, licenses, and terms of use? Which is more binding?

In summary, laws are enacted and enforced by the state and everyone must abide by them, while licenses and terms of use are contracts agreed upon by civil subjects such as persons and corporations per autonomy of will. Laws are more binding because violation of law will absolutely result in liability and punishment. However, licenses and terms of use may be deemed as void or voidable if certain provisions violate the law. Detailed definition and illustration are as follows:

Laws are legal documents that are enacted, revised and promulgated in accordance with legal procedures by the legislative organs with legislative powers, and enforced by the coercive force of the state. In China, laws can be classified into: (1) Constitution (2) laws (3) administrative regulations (4) local regulations (5) autonomous regulations and separate regulations.

Licenses are agreements where the owner authorizes others to use something within a certain period and scope without changing the ownership. According to the scope of licensing, they can be classified into: (1) proprietary license (2) exclusive license (3) ordinary license.

Terms of use are agreements between service providers and users regarding the service provided and rights and obligations of both parties. For example, users need to read and consent

to the terms of use before they use certain online services such as apps, websites and other e-platforms.

Licenses and terms of use are less binding than laws, because they may be deemed as void or voidable if certain provisions violate the law. According to PRC Civil Code Article 144, 146, 153, 154, 156, contracts are void in situations: (1) The civil juristic acts performed by a person who has no capacity for civil conduct (2) The civil juristic acts by persons of civil conduct and counterparties under false expression of intent (3) A civil juristic act that violates the mandatory provisions of laws and administrative regulations (4) Where an actor colludes with another party to perform a civil juristic act that damages others' legitimate rights and interests. According to PRC Civil Code Article 147, 148, 149, 150, 151, contracts are voidable in situations: (1) A civil juristic act that is performed based on a substantial misunderstanding (2) A civil juristic act that is performed by a party against his or her real intention as a result of fraud committed by another party (3) A civil juristic act is performed by a party against his or her real intention as a result of fraud committed by the third party under the circumstance that the other party knows or might know about the fraud (4) A civil juristic act that is performed by a party against his or her real intention as a result of coercion by another party or a third party (5) Where a civil juristic act is obviously unfair when instituted by a party making use of another party's dangerous or unfavorable position or lack of judgment.

Especially, terms of use are likely to be construed as standard form contracts. Therefore, they shall abide by PRC Law on Protection of Consumer Rights and Interests[13] Article 24: Business operators may not, through format contracts, notices, announcements, entrance hall bulletins and so on, impose unfair or unreasonable rules on consumers or reduce or escape their civil liability for their infringement of the legitimate rights and interests of consumers. Format contracts, notices, announcements, entrance hall bulletins and so on with contents mentioned in the preceding paragraph shall be invalid.

3. What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?

As mentioned in Question b1, data sets and models are generally contractually licensable. Therefore, whether a license is valid depends on the validity of the license contract. According to the Civil Code of the PRC, a legally executed contract is generally valid, except for the below circumstances:

1) The contract is executed by a person who has no capacity for performing civil juristic acts; (Article 144)

2) The contract is executed by a person and another person based on a false expression of intent; (Article 146)

3) The contract is executed based on serious misunderstanding; (Article 147)

[13] PRC Law on Protection of Consumer Rights and Interests: 中华人民共和国消费者权益保护法_中国人大网 (npc.gov.cn)

4) The contract is executed by fraudulent means; (Article 148)

5) A party knows or should have known that a contract is made by the other party is based on a third person's fraudulent act and is against the other party's true intention; (Article 149)

6) A party makes the contract against its true intention owing to duress of the other party or a third person; (Article 150)

7) One party takes advantage of the other party that is in a desperate situation or lacks the ability of making judgment, and as a result the contract is obviously unfair; (Article 151)

8) The contract is in violation of the mandatory provisions of laws or administrative regulations, or public order or good morals; (Article 153)

9) The contract is executed through malicious collusion between a person who performs the act and a counterparty thereof and thus harms the lawful rights and interests of another person; (Article 154)

In practice, to avoid potential dispute on the authenticity of the contract, parties to a contract will usually have the contract notarized with the notary office.

4. What about if the users download or copy their own data and then provide it to NLP researchers directly?

Users have right to obtain and copy their own personal information, and arguably, voluntary submission of their personal data directly to the NLP researchers constitute a consent. NLP researchers may use or process such personal data based on consent pursuant to *Article 13.1 of the PIPL*.

Article 14 of the PIPL provides that where personal information is processed based on an individual's consent, such consent shall be voluntarily and explicitly given by the individual on a fully informed basis. So it is recommended that NLP researches make a clear, truthful, accurate and complete notice in a conspicuous way, notifying the users its processing purpose and method, retention period etc. Please note that consent can be withdrawn according to *Article 15 of the PIPL*.

However, if the terms of use agreed between users and processors clearly states that the users can't share the data with third-parties absent the platform's consent, they may not do this as long as the terms of use is valid. Please refer to answer to question B5 for detailed explanation.

5. [simplify the question into] Does the dataset license override the terms of use?

There are different legal relationships in the license and terms of use. We understand that the license is between the data subject and NLP researchers where the NLP researchers are

authorized to use the data subjects' data, while the terms of use are between the platform and all its users. Since there's no clear regulation on the ownership of data yet, practically the right to use data can be set up in contracts like terms of use. Terms of use often stipulate that the dataset generated on the platform can't be shared with third-party without the platform's consent. Absent the void or voidable conditions as discussed in question B2, the terms of use are valid. In this situation, the data subjects agreed and shall bind by the terms. If they license others to use the data, they will incur the liability of breach of contract. Note that the platform may not have a claim against the third-party data recipient because it's not the party of the terms of use.

6. Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?

Likely, yes. NLP researchers can use licensed/copyrighted data to which they legally have access to train a language model as long as NLP researchers are licensed to copy the contents (or other activities required in the NLP research process), not merely licensed to access the data.

As mentioned in A7, one big problem is that NLP researchers may violate copyright of the data. If those data is licensed to the NLP researchers so that NLP researchers can legally use the data, NLP researchers can use the data.

According to Article 13 of the *Copyright Law*, where a work is created by arrangement of a pre-existing work, the copyright in the work thus created shall be enjoyed by the arranger, provided that the copyright in the original work is not infringed upon. Therefore, NLP researchers shall has the copyright of the model, and with the license, NLP researchers will not infringe the copyright in the original work.

Also, according to Article 16 of the *Copyright Law*, the publication, performance, and production of audio and video works using works produced by adapting, translating, annotating, arranging or compiling existing work, shall obtain the permission of those works' copyright holders and of the copyright holders for the original work, and pay remunerations. Therefore, in order to comply with Article 16, it's better to acquire the consent in the license and stipulate the remuneration in the license.

Meanwhile, please pay attention to A1 that the copyright of data is still debatable.

7. Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?

As mentioned in Question b1, data sets and models are generally contractually licensable. Under the PRC legal framework, parties to the contract can agree on the amount of the royalty of the license of the data sets or models. Royalty free license is permitted as long as the parties agree.

8. Is there any supporting policy for what we are doing in this jurisdiction?

There are many supporting policy in China concerning big data and artificial intelligence.

The most significant national policy documents in the areas include:

1) *Notice of the State Council on Issuing the Action Outline for Promoting the Development of Big Data*: Create a new innovation-driven pattern of business startups and innovations from all walks of life.

2) *Guiding Opinions of the State Council on Vigorously Advancing the "Internet Plus" Action*: Include AI as an important mission.

3) *Notice of the State Council on Issuing the Development Plan on the New Generation of Artificial Intelligence*: Include AI in the national strategy.

4) *Outline of the 14th Five-Year Plan (2021-2025) for National Economic and Social Development and Vision 2035 of the People's Republic of China*: Emerging digital industries including artificial intelligence, big data, blockchain, cloud computing, and cybersecurity will be grown stronger. … Enterprises will be encouraged to provide open access to search, e-commerce, social, and other data.

5) *Three-year action plan for the development of new data centers*: promote the development of AI and data center.

6) *Data Security Law*: Article 7 The state shall … encourage the lawful, reasonable and effective utilization of data, safeguard the orderly and free flow of data in accordance with the law, and promote the development of a digital economy with data as a key factor. Article 11 The state shall proactively engage in international exchanges and cooperation in data security governance, data development and utilization, and other fields, participate in the formulation of international rules and standards related to data security, and promote the secure and free cross-border flow of data. Article 16 The state shall support the technological research on data security and data development and utilization, encourage technological promotion and commercial innovation in the fields of data development and utilization and data security, among others, and cultivate and develop data development and utilization and data security product and industry systems.

There are also many provincial policy document in many provinces, autonomous regions, and municipalities.

C. Text Data Mining and Fair Use Questions

1. What are types of legally permitted text data mining?

According to different classification standards, data mining can be divided into different types, and Chinese laws do not distinguish between different types of data mining. Therefore, it

is suggested to change the original question "what types of text data mining are legally permitted" into "legal risks in the process of text data mining".

1) Intellectual property risk

This has been explained in the IP section.

2) Personal privacy risk

This will be explained in the privacy section. It is suggested to use data desensitization and anonymization technology to protect personal data security and privacy.

In conclusion, the key to the legality of text data mining is not the type, but the specific implementation method and technical operation. As the above technical operations are difficult to avoid in the process of text data mining, such risks may be unavoidable before the revision of relevant laws at this stage.

2. Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question?

NLP researchers can get their own data from the Internet without problems and without regional limitations. According to Article 7 of the Data Security Law of the People's Republic of China[14], the state shall protect the data-related rights and interests of individuals and organizations, encourage the lawful, reasonable, and effective utilization of data, safeguard the orderly and free flow of data in accordance with the law, and promote the development of a digital economy with data as a key factor. However, Article 8 of the Data Security Law stipulates that when conducting data processing activities, one shall comply with laws and regulations, respect social norms and ethics, observe business and professional ethics, act in good faith, perform data security protection obligations, and undertake social responsibilities, and shall neither compromise national security and public interest nor harm the lawful rights and interests of any organization or individual. Therefore, researchers should pay attention not to violate the corresponding laws and regulations.

First, crawler data capturing data containing citizens' personal information may constitute the crime of infringing citizens' personal information according to Article 253 of the Criminal Law[15], which is described in the section on privacy.

Second, when the web crawler crawls the original data content, it is subject to the intellectual property law and criminal laws, which are described in the intellectual property section.

Third, web crawler behavior may violate the anti-unfair competition law. Article 2 of the Anti-Unfair Competition Law of the People's Republic of China[16] stipulates that Businesses shall, in their production and distribution activities, adhere to the free will, equality, fairness, and good faith

[14] PRC Data Security Law, available at http://www.npc.gov.cn/npc/c30834/202106/7c9af12f51334a73b56d7938f99a788a.shtml
[15] PRC Criminal Law, available at http://www.npc.gov.cn/wxzl/wxzl/2000-12/17/content_4680.htm
[16] PRC Anti-Unfair Competition Law, available at http://www.npc.gov.cn/wxzl/wxzl/2000-12/05/content_4600.htm

principles, and abide by laws and business ethics. The Anti-unfair competition Law has no special provisions on crawler behavior, but crawlers may still violate the commercial ethics stipulated in this article and be identified as violating the Anti-unfair competition law. In the case of Sina v. Maimai, the court held that Maimai violated the robot agreement, and without the authorization of Sina, the market used by the crawler data directly competed with the original data holder and was likely to substantially replace the market economic value of the original market subject, thus violating the anti-unfair competition law.

Fourthly, crawler data may constitute an illegal invasion of computer information systems and other related crimes. Article 285 of the Criminal law stipulates the crime of trespassing into the computer information system: Whoever violates state regulations and intrudes into computer systems with information concerning state affairs, construction of defense facilities, and sophisticated science and technology is be sentenced to not more than three years of fixed-term imprisonment or limited incarceration.

Whoever, in violation of the state provisions, intrudes into a computer information system other than that prescribed in the preceding paragraph or uses other technical means to obtain the data stored, processed or transmitted in the said computer information system or exercise illegal control over the said computer information system shall, if the circumstances are serious, be sentenced to fixed-term imprisonment not more than three years or limited incarceration, and/or be fined; or if the circumstances are extremely serious, shall be sentenced to fixed-term imprisonment not less than three years but not more than seven years, and be fined.

Whoever provides special programs or tools specially used for intruding into or illegally controlling computer information systems, or whoever knows that any other person is committing the criminal act of intruding into or illegally controlling a computer information system and still provides programs or tools for such a person shall if the circumstances are serious, be punished under the preceding paragraph.

Where an entity commits any crime as provided for in the preceding three paragraphs, the entity shall be sentenced to a fine, and its directly responsible person in charge and other directly liable persons shall be punished according to the provisions of the applicable paragraph.

Meanwhile, Article 286 of the Criminal Law stipulates the crime of destroying the computer information system:

Whoever violates states regulations and deletes, alters, adds, and interferes in computer information systems, causing abnormal operations of the systems and grave consequences, is to be sentenced to not more than five years of fixed-term imprisonment or limited incarceration; when the consequences are particularly serious, the sentence is to be not less than five years of fixed-term imprisonment.

Whoever violates state regulations and deletes, alters, or adds the data or application programs installed in or processed and transmitted by the computer systems, and causes grave consequences, is to be punished according to the preceding paragraph.

Whoever deliberately creates and propagates computer viruses and other programs which sabotage the normal operation of the computer system and cause grave consequences is to be punished according to the first paragraph.

Where an entity commits any crime as provided for in the preceding three paragraphs, the entity shall be sentenced to a fine, and its directly responsible person in charge and other directly liable persons shall be punished according to the provisions of paragraph 1.

To sum up, crawler behavior itself is feasible and has no regional restrictions, but relevant laws and regulations need to be paid attention to.

3. If NLP researchers can crawl data ourselves, what are the limits regarding using this data for training a language model? Can NLP researchers redistribute the data afterwards? If so, under which license and in which geographical regions?

In addition to complying with the relevant regulations described in the previous question, the use of data training language models and the redistribution of data in China requires attention to intellectual property issues, as described in section A.

4. If NLP researchers use parts of Common Crawl, OSCAR or C4, can NLP researchers redistribute the data later? If yes under which conditions and in which countries?

When obtaining the relevant database information, it should follow the principle of legality, legitimacy, and necessary limits. It can make use of the database information for the purpose of achieving positive effects to a certain extent but can't substitute the original database developer's service substantially by improper means. Researchers can therefore redistribute data taken from other databases but be aware of the legal risks mentioned in the previous two questions, including intellectual property rights, personal privacy, unfair competition, and computer system crime.

5. What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify?

There is currently no provision regulating internet scraping behavior. In practice, courts generally applied Article 2 of Anti-Unfair Competition Law of the People’s Republic of China (Amended and effective since April 2019) in adjudicating disputes between the data scraper and data controller, Article 2 stipulates that “[b]usinesses shall, in their production and distribution activities, adhere to the free will, equality, fairness, and good faith principles, and abide by laws and business ethics.”

The courts usually analyze such cases by examining the following aspects:

1) Whether the legitimate rights and interests of other operators have been actually damaged due to such behavior.

2) Whether such behavior violates the principle of good faith and recognized business ethics.

3) Whether the technical means harm the interests of consumers, such as restricting consumers' right to choose independently, failing to protect consumers' right to know, infringing on consumers' right to privacy, etc.

4) Whether such behavior destroys the open and fair market competition order in the Internet environment and leads to or has the possibility of vicious competition.

5) Internet cases also follows the "triple authorization principle" established in Sina Weibo v. MaiMai[17], i.e. the scraping company shall have user's authorization to data holding company, user's authorization to data scraping company and the authorization given by the date holding company to the data scraping company to conduct the scraping behavior.

Neither is there any provision specifying the scope and definition of "social media content". However, some instructions on how social media content is defined can be found in the Terms of Use of Sina Weibo. Social media content is impliedly defined in a catch all provision in Section 4.12 of the Terms of Use, including but not limited to "user position, words and remarks, pictures, audios, videos, trademarks, patents and publications." It is fair to say that almost everything posted qualifies as social media content, obviously comment on news falls within the scope.

6. For example, do the terms and conditions of Twitter, Facebook, Youtube, etc., tell us whether NLP researchers can collect data from them for a project such as BigScience?

Given the status quo that services of social media platforms such as Twitter and Facebook are not available in China, the more relevant social media platform to study is Sina Weibo, the most prominent social media platform with highest popularity. The answer to these series of questions will be based on the Terms of Use and practices of Sina Weibo.

Section 1.3.2 of the Terms of Use[18] stipulates that "Without the prior written consent of Weibo Operator, the user shall not **illegally scrape** or authorize or assist any third party to scrape the social media content. "**Illegal scraping**" refers to the behavior of obtaining content data by using programs or abnormal browsing and other technical means."

Section 4.5 prohibits **"any person"** from conducting "automatic behavior", which is defined in Section 4.5.1 as "the behavior that users themselves or authorize or assist a third party to publish contents or … **scrape** data by using automated means or means obviously different from that available to ordinary people, which is much higher than the frequency of normal users" without the consent of Weibo Operator.

In conclusion, according to the terms and conditions of social media platform such as Sina Weibo, NLP researchers are generally not allowed to collect data unless with the consent of operators.

[17] Sina Weibo v. Maimai, available at https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=49854fde619a47d7b772a71d000fcf00

[18] Terms of Use of Sina Weibo, available at https://weibo.com/signup/v5/protocol

7. What are some of the risks raised by collecting data from these social media directly?

In case of manual replication, the scale and real-time feature of data obtained cannot be compared with that obtained through technical means. As there is no technical means that meet the basic application threshold of Article 12 Section 2 of the Anti-Unfair Competition Law(amended and effective since April 2019) which set limits on data collecting business using technical means, and because this behavior is essentially the same as the user's behavior of copying content when browsing the web page, it is difficult for court to determine it as violating the principle of good faith and recognized business ethics. Therefore, such behavior itself is difficult to be evaluated as unfair competition alone.

However, for cases lacking obvious illegality in the data collection stage, the courts tends to focus on the legitimacy of the subsequent use and render judgement applying Article 2 of the Anti-Unfair Competition Law (amended and effective since April 2019). In *Yachang v. Xunbao*[19], both companies operate auction websites providing users with auction display and online auction services etc. The defendant copied auction information and pictures from the plaintiff. In this case, the court focused on the substitutability effect of the accused infringement act and finally determined that the act involved was illegitimate according to Article 2.

This case shed light to us that the court looked into substitutability in applying Article 2. Therefore, if NLP researchers' subsequent behavior does not have substitutability with the social media platform, mere data collecting will not be regarded in violation of Article 2 of the Anti-Unfair Competition Law.

8. What changes if NLP researchers get direct consent from the users concerned?

As provided in Section 4.5.1 of the Terms of Use, scraping social media content is included in the definition of automatic behavior, and is so generally forbidden. Direct consent of users does not change the results. According to the relevant sections, scraping of social media content is only permitted with the consent of Weibo Operator.

9. Does the consent override the Terms of Use?

According to current regulation, there is no provision indicating whether user's consent would override the Terms of Use. However, as indicated in Answer to C5, courts tend to adopt the "triple authorization" principle set in *Weibo v. MaiMai*, which means the scraping company shall have user's authorization to data holding company, user's authorization to data scraping company and the authorization given by the date holding company to the data scraping company to conduct the scraping behavior. Generally, consent from the users' side does not override the Terms of Use.

10. Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can't be scrapped?

[19] *Yachang v. Xunbao*, available at https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=cebd2abd9e49457d9450ad5d000b446d

Still, as forums such as Reddit is now not available in mainland China, answer to this question will focus on a more relevant forum, Douban. According to Section 2.1 of its Legal Declaration[20], unless otherwise specified by law, other rights and interests of "all **information and data**" belong to Douban. And social media content posted by users are included in "information and data".

Douban's Legal Declaration also put restriction on use of its data and data scraping. Section 2.2 stipulates that "[u]nless otherwise provided by law, no subject shall use any public information or data of Douban in any form for any purpose without obtaining Douban's written consent, including but not limited to: (1) commercial use…(3) derivative utilization of Douban content, including but not limited to any plug-ins, software, applications and websites developed based on or using Douban content…(5) data collection, web crawler or similar data collection and data extraction that infringe or may infringe the rights and interests of Douban…" Therefore, data scraping or direct use of Douban's data must meet the prerequisite of obtaining the platform's written consent.

The terms and conditions set by Douban were similar to that of Weibo, and the social media companies are negative alike in data scraping. Generally speaking, there is little meaningful legal difference between content-based social media and forums on what can be scrapped as scrapping of almost any social media content is not allowed without acquiring the platform's consent beforehand. But in practice, different platforms may have different attitude on granting such consent.

11. How is automated decision making regulated in this jurisdiction?

"Automated decision-making" in China, according to PIPL Article 73, is defined as the activities of automatically analyzing and assessing individuals' behavioral habits, hobbies, or financial, health and credit status through computer programs and making decisions thereon.

The major concerns when using automated decision making are fair transaction and privacy. Relevant laws are as follows:

First, automated decision making shall not be used to discriminate among consumers. For instance, to provide different prices according to the consumption level through analyzing consumers' order history. The recently enacted regulation Internet Information Service Algorithm Recommendation Management Regulations[21] (effective in March 1st 2022) Article 21 states that when an algorithm recommendation service provider sells goods or provides services to consumers, it shall protect consumers' right to fair transactions, and shall not use algorithms on transaction conditions such as transaction prices based on consumer preferences and transaction habits to carry out illegal acts such as unreasonable differential treatment. Similarly, PIPL Article 24 holds that where a personal information processor conducts automated decision-making by using personal information, it or he shall ensure the transparency of the decision-making and the

[20] Legal Declaration of Douban, available at https://www.douban.com/about/legal

[21] Internet Information Service Algorithm Recommendation Management Regulations: 互联网信息服务算法推荐管理规定-中共中央网络安全和信息化委员会办公室 (cac.gov.cn)

fairness and impartiality of the result, and shall not give unreasonable differential treatment to individuals in terms of trading price or other trading conditions.

Second, when personal information is involved to conduct commercial marketing such as targeted advertising, the processors must provide the option not specific to persons, and individuals have the right to refuse. According to PIPL Article 24, where information push or commercial marketing to individuals is conducted by means of automated decision-making, options not specific to individuals' characteristics shall be provided simultaneously, or convenient ways to refuse shall be provided to individuals. Where a decision that has a major impact on an individual's rights and interests is made by means of automated decision-making, the individual shall have the right to request the personal information processor to make explanations and to refuse to accept that the personal information processor makes decisions solely by means of automated decision-making.

12. Does the newly enacted Data Security Law change the landscape of data scraping?

Yes. Although the Data Security Law enacted in June 2021 does specifically mention the term "data scraping", it establishes some basic principles for data collection. Article 8 of the Data Security Law stipulates that "the carrying out of data handling activities shall obey laws and regulations, respect social mores and ethics, comply with commercial ethics and professional ethics, be honest and trustworthy, perform obligations to protect data security, and undertake social responsibility; it must not endanger national security, the public interest, or individuals" and organizations" lawful rights and interests"[22]. As such, data scraping shall now follow the above principles.

In addition, a derivative administrative regulation draft, the Regulations on Network Data Security Management (draft)[23], is now open to the public for comment in November, 2021. This regulation contains specific provisions regarding the data scraping. Article 16 provides that "network operators shall not, when using automatic means to access or collect website data, interfere with the normal operation of their websites. If such acts seriously affect the operation of websites (e.g., if the traffic of automatic visits or data collection exceeds one third of the average traffic of the website) and the website requests the network operator to cease such automatic access and collection, the network operator shall cease such practice." Thus, data scraping will be subject to the above restrictions if the regulation is promulgated.

13. What is the fair use of public data?

Based on originating sources, data can be divided into public data, commercial data and personal data.

Public data refers to data resources that administrative authorities collect, produce, or obtain in the course of performing their functions in accordance with legal provisions, and record and

[22] Data Security Law of the PRC: 中华人民共和国数据安全法

[23] Regulations on Network Data Security Management (draft): 《网络数据安全管理条例（征求意见稿）》

preserve them through certain forms. The process of forming public data determines that it already has a public attribute in essence.

Unlike personal data, which contains personal information, the use of public data does not require the consent of the original data sources. However, the relevant use still needs to comply with the basic duty of care to prevent the improper use of the data from causing damage to the interests of the original data source.

Compared with other data types, public data is characterized by open sharing and encouraging use. In the collection and release of public information, the following basic principles shall be adhered to:

The first is the principle of legal data sources. In the process of capturing and collecting data, big data analysis companies shall use information that has been disclosed by government departments or other management entities in accordance with law, and must not collect data that is prohibited by laws or administrative regulations. The act of collecting data and information from the public domain is lawful and does not require the consent of the information sources, but the security of the information shall be ensured for non-public information involving personal privacy, corporate trade secrets, etc.

The second is to pay attention to the principle of information timeliness. Timeliness includes two aspects, namely, the timeliness of information updates and the accuracy of information changes.

The third principle is ensuring information quality. Big data platforms shall employ reasonable measures to ensure the accuracy of the information they provide. The quality of information shall include the truthfulness, accuracy and completeness of the published information.

The fourth is the principle of sensitive information verification. The duty of care with respect to data quality should be differentiated according to the type of data. For non-sensitive data, when data deviations occur, big data analysis enterprises should be allowed to correct them through ex post facto relief; for sensitive data, especially information involving enterprise liquidation and bankruptcy, big data analysis enterprises shall establish differentiated technical processing principles, and improve the quality of data by improving algorithm technology, data review, cross-verification and other means, so as to avoid bringing major negative impacts to data subjects.

14. What are the risks of data scraping other platforms other than social media content?

Per answer to question C2, web scraping may violate Article 2 of the PRC Unfair Competition Law. The answer illustrates the Sina v. Maimai case, which is a social media platform example and the data scraped is personal information. We want to add here that web scraping other types of data from other kinds of platforms may also be construed as unfair competition.

There are two interesting cases in China. One is Dianping v. Baidu, which decided that web scraping user comments from the customer review platform is unfair competition. Another is Gumi

v. Yuanguang, which decided that web scraping public transportation data from authorized private service provider is unfair competition. More detailed introduction is as follows:

In Dianping v. Baidu[24], Baidu map and Baidu Q&A scraped the customer reviews from Dianping.com and displayed these information on their own service platforms without consent. In our country's practice, there are 3 elements to identify unfair competition: 1) the relationship between the plaintiff and the defendant shall be competitive 2) the defendant's conduct harmed the plaintiff 3) the conduct is unlawful. The court finally decided that Baidu's behavior is unfair competition because 1) Although Dianping and Baidu provide different services, they target the same group of customers, so they could be viewed as competitors. 2) Baidu's conduct enabled users to see the customer reviews without visiting Dianping.com. So it would cause loss of user visits and potential business opportunities for Dianping. 3) Since Dianping has invested a large amount of time and effort to collect, classify and display these customer review information, Baidu's conduct is "free-riding" and not in good faith.

In Gumi v. Yuanguang[25], Gumi cooperates with bus companies to obtain real-time bus location data, sorts and provides it to the public for free searching through their App "Kumike". However, the defendant Yuanguang cracked into the "Kumike" App encryption system, crawled its data, and used it directly for providing the same searching services to the public through its App "Chelaile". The court also decided that Gumi collected and arranged the data with time and efforts, thus enjoys competitive interest in these data which it has right to claim.

15. Does engaging in text data mining lead to data security protection obligations?

Text data mining needs to undertake data protection obligations, including improving the whole process of data security management system, organizing and carrying out data security education and training, etc.

According to Article 27 of the Data Security Law of the People's Republic of China, in conducting data processing activities, one shall, in accordance with the provisions of laws and regulations, establish and improve a whole-process data security management system, organize data security education and training, and take corresponding technical measures and other necessary measures to safeguard data security. In conducting data processing activities by using the Internet or any other information network, one shall perform the above data security protection obligations on the basis of the hierarchical cybersecurity protection system.

A processor of important data shall specify a person in charge of data security and a data security management body and enforce the responsibility for data security protection.

Under Article 28, conducting data processing activities and research and development of new data technologies shall be conducive to promoting economic and social development and enhancing the well-being of the people, and comply with social norms and ethics.

[24] Dianping v. Baidu: 上海市浦东新区人民法院发布10个互联网不正当竞争典型案例之六：大众点评网数据信息不正当竞争纠纷案——数据信息使用行为是否构成不正当竞争的司法认定 - Google Docs

[25] Gumi v. Yuanguang: 深圳市谷米科技有限公司与武汉元光科技有限公司、邵凌霜等不正当竞争纠纷 - 世界互联网法治论坛 (court.gov.cn)

According to Article 29, in conducting data processing activities, one shall strengthen risk monitoring, and when any data security defect, vulnerability, or other risk is discovered, immediately take remedial measures; and when a data security event occurs, immediately take disposition measures, and notify users and report to the appropriate department in a timely manner as required.

And under Article 30,a processor of important data shall conduct regular risk assessments with respect to its data processing activities as required, and submit risk assessment reports to the appropriate department.

The risk assessment report shall include the type and quantity of important data processed, information on data processing activities, the data security risks faced, and countermeasures.

In conclusion, text data mining may need to undertake the data security protection obligations stipulated in the above clauses.

However, what constitutes Important Data are still an educated guess, other than in automotive sector.

D. Privacy Questions

1. Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?

The legal concern around accessing web content and datasets that have PII are mainly provided in the PRC Personal Information Protection Law (“PIPL”). The relevant provisions are as follows.

Article 6: Personal information handling shall have a clear and reasonable purpose, and shall be directly related to the handling purpose, using a method with the smallest influence on individual rights and interests. The collection of personal information shall be limited to the smallest scope for realizing the handling purpose, and excessive personal information collection is prohibited.

Article 10: No organization or individual may illegally collect, use, process, or transmit other persons' personal information, or illegally sell, buy, provide, or disclose other persons' personal information, or engage in personal information handling activities harming national security or the public interest.

Article 13: Personal information handlers may only handle personal information where they conform to one of the following circumstances: 1. Obtaining individuals’ consent; 2. Where necessary to conclude or fulfill a contract in which the individual is an interested party, or where necessary to conduct human resources management according to lawfully formulated labor rules and regulations and lawfully concluded contracts; 3. Where necessary to fulfill statutory duties and responsibilities or statutory obligations; 4. Where necessary to respond to sudden public health incidents or protect natural persons’ lives and health, or the security of their property, under

emergency conditions; 5. Handling personal information within a reasonable scope to implement news reporting, public opinion supervision, and other such activities for the public interest; 6. When handling personal information already disclosed by persons themselves or otherwise lawfully disclosed, within a reasonable scope in accordance with the provisions of this Law; 7. Other circumstances provided in laws and administrative regulations. In accordance with other relevant provisions of this Law, when handling personal information, individual consent shall be obtained. However, obtaining individual consent is not required under conditions in items 2 through 7 above.

Article 27: Personal information handlers may, within a reasonable scope, handle personal information that has already been disclosed by the person themselves or otherwise lawfully disclosed, except where the person clearly refuses. Personal information handlers handling already disclosed personal information, where there is a major influence on individual rights and interests, shall obtain personal consent in accordance with the provisions of this Law.

Article 28: Sensitive personal information means personal information that, once leaked or illegally used, may easily cause harm to the dignity of natural persons or grave harm to personal or property security, including information on biometric characteristics, religious beliefs, specially-designated status, medical health, financial accounts, individual location tracking, etc., as well as the personal information of minors under the age of 14. Only where there is a specific purpose and sufficient necessity, and under circumstances where strict protection measures are taken, may personal information handlers handle sensitive personal information.

PIPL also regulates cross-border distribution of PII, as provided in the articles below.

Article 38: Where personal information handlers need to provide personal information outside the territory of the People's Republic of China for business or other such requirements, they shall meet one of the following conditions: 1. Passing a security assessment organized by the national cyberspace authority according to Article 40 of this Law; 2. Undergoing personal information protection certification conducted by a specialized body according to provisions by the national cyberspace authority; 3. Concluding a contract with the overseas receiving party in accordance with a standard contract formulated by the national cyberspace authority, agreeing upon the rights and responsibilities of both sides; 4. Other conditions provided in laws or administrative regulations or by the national cyberspace authority; 5. Where treaties or international agreements that the People's Republic of China has concluded or acceded to contain provisions such as conditions on providing personal data outside the territory of the People's Republic of China, it is permitted to act according to those provisions. Personal information handlers shall adopt necessary measures to ensure that the personal information handling activities of overseas receiving parties reach the standard of personal information protection provided in this Law. As a practical matter, none of these mechanisms has existed as of today.

1.i. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of common crawl, a published dataset based on a crawl of the web in terms of PII?

Please refer to the answer of question C12 for relevant general regulations regarding data craw (data scraping). As publishing a model trained with content containing PII is also

a kind of “handling personal information”, provisions listed in the previous question are also applied.

2. Are there concerns around distributing models that have PII stored within it? That can expose such PII?

In summary, providing PII to other processors or making PII public by any means requires the processor to notify the data subject and obtain the separate consent of him. Distributing models is just one of the means. It’s not the means to disclose PII, but the fact that PII is disclosed that matters. Relevant laws are as follows:

According to PRC Personal Information Protection Law (“PIPL”) Article 25, personal information processors shall not disclose the personal information processed, except with the separate consent of the individuals.

According to PIPL Article 23, a personal information processor that provides any other personal information processor with the personal information it or he processes shall notify individuals of the recipient's name, contact information, purposes and methods of processing, and categories of personal information, and obtain the individuals' separate consent. The recipient shall process personal information within the scope of the aforementioned purposes and methods of processing, and categories of personal information, among others. Where the recipient changes the original purposes or methods of processing, it or he shall obtain individuals' consent anew in accordance with this Law.

3. What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exception for research purposes?

The takedown of personal information becomes a concern mostly when NLP researchers obtain personal information through means such as web scraping.

According to PIPL Article 27, a personal information processor may process within a reasonable scope the personal information that has been disclosed by an individual himself or herself or other personal information that has been legally disclosed, except that the individual has expressly refused. A personal information processor shall obtain consent from an individual in accordance with the provisions of this Law if the processing of the disclosed personal information of the individual has a major impact on the individual's rights and interests. Therefore, NLP researchers can process the already disclosed personal information within a reasonable scope without individual’s consent, except the individual has clearly refused or the processing has a great impact on the individual.

However, according to PIPL Article 44, individuals shall have the right to know and the right to decide on the processing of their personal information, and have the right to restrict or refuse the processing of their personal information by others. Thus if the individual contacts the NLP processor and refuses the processing, or the personal information published was not even authorized to be disclosed at the first place, NLP researchers had better delete those personal information immediately.

Although there's no clear legislation on the liability of special data processors like NLP researchers, I think the mechanisms NLP researchers need to maintain will be similar to Internet service providers (ISP), which can be summarized as the rule of "notify-delete". PRC Civil Code Article 1195 states that where a network user commits a tort through the network services, the right holder shall be entitled to notify the network service provider to take such necessary measures as deletion, block or disconnection. The notice shall include the prima facie evidence of the tort and the true identity information of the right holder. After receiving the notice, the network service provider shall, in a timely manner, forward the notice to the relevant network user and take necessary measures based on the prima facie evidence of the tort and type of service; and if the network service provider fails to take necessary measures in a timely manner, it shall be jointly and severally liable for any additional harm with the network user. Since NLP researchers didn't have bad faith when they scraped the published data, their obligations are likely to be delete the data timely upon data subjects' request.

According to the legitimate basis of collecting personal information in PIPL Article 13 discussed in question D5, there are exceptions such as news reporting purposes and public health emergencies purposes, but no exception for research purposes.

4. How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?

According to PRC Personal Information Protection Law Article 13 ("PIPL", effective on November 1, 2021), when handling personal information, individual consent shall be obtained. However, obtaining individual consent is not required under several exceptional cases, including (i) where necessary to conclude or fulfill a contract in which the individual is an interested party, or where necessary to conduct human resources management according to lawfully formulated labor rules and regulations and lawfully concluded contracts; (ii) where necessary to fulfill statutory duties and responsibilities or statutory obligations; (iii) where necessary to respond to sudden public health incidents or protect natural person's lives and health, or the security of their property, under emergency conditions; (iv) handling personal information within a reasonable scope to implement news reporting, public opinion supervision, and other such activities for the public interest; (v) when handling personal information already disclosed by persons themselves or otherwise lawfully disclosed, within a reasonable scope in accordance with the provisions of this law; and (vi) other circumstance provided in laws and administrative regulations.

It should be noted that Article 14 of the PIPL requires that such individual consent shall be given by individuals under the precondition of full knowledge, and in a voluntary and explicit statement.

5. What are the privacy risks related to data collection directly from persons? For example when you interview people or they donate data etc.

According to PRC Personal Information Protection Law ("PIPL") Article 13: A personal information processor may not process personal information unless: (1) the individual's consent has been obtained (2) the processing is necessary for the conclusion or performance of a contract to which the individual is a contracting party or for conducting human resource management under

the labor rules and regulations developed in accordance with the law and a collective contract signed in accordance with the law (3) the processing is necessary to fulfill statutory functions or statutory obligations (4) the processing is necessary to respond to public health emergencies or protect the life, health or property safety of natural persons under emergency circumstances (5) personal information is processed within a reasonable scope to conduct news reporting, public opinion-based supervision, or other activities in the public interest (6) the personal information that has been disclosed by the individuals themselves or other personal information that has been legally disclosed is processed within a reasonable scope in accordance with this Law (7) under any other circumstance as provided by any law or administrative regulation.

Therefore, if the data constitutes “personal information” which, according to PIPL Article 4, means all kinds of information related to identified or identifiable natural persons that are electronically or otherwise recorded, excluding information that has been anonymized, then it must fall within one of the legitimate bases above. For example, interviewing people may fall within “(5) news reporting purpose” and donating data may be voluntarily with consent or fall within “(6) disclosure by individuals themselves”, depending on the circumstances.

Specifically, if the personal information collection is based on data subject’s consent, according to PIPL Article 14, such consent shall be voluntarily and explicitly given by the individual on a fully informed basis. In order to achieve the “informed requirement”, PIPL Article 17 stipulates that a personal information processor shall, before processing personal information, truthfully, accurately and completely notify individuals of the following matters in a conspicuous way and in clear and easily understood language: (1) The name and contact information of the personal information processor. (2) Purposes and methods of processing of personal information, categories of personal information to be processed, and the retention periods. (3) Methods and procedures for individuals to exercise the rights provided in this Law. (4) Other matters that should be notified as provided by laws and administrative regulations. Where the personal information processor notifies such matters in the manner of developing personal information processing rules, the processing rules shall be disclosed and easy to consult and preserve. Therefore, the data processor shall obtain informed consent if no other legitimate basis exists.

6. What is the scope of the allowed personal information collection?

According to Article 6 of the PIPL, personal information handling shall have a clear and reasonable purpose, and shall be directly related to the handling purpose, using a method with the smallest influence on individual rights and interests. The collection of personal information shall be limited to the smallest scope for realizing the handling purpose, and excessive personal information collection is prohibited.

In addition, Article 14 of the PIPL provides that where a change occurs in the purpose of personal information handling, the handling method, or the categories of handled personal information, the individual’s consent shall be obtained again.

Note: For the purpose of the PIPL, personal information handling includes personal information collection, storage, use, processing, transmission, provision, publishing, deletion, etc.

7. Is there extra-territorial jurisdiction for PII?

Yes. The application scope of PIPL is extra-territorial. Apart from all the processing activities happening in China, the activities happening outside China shall comply with the PIPL if it provides products or services to people within China or analyzes their behaviors. Note that PII processing covers the whole life cycle of PII, including collection, storage and deletion etc. Relevant laws are as follows.

According to PIPL Article 3, This Law shall apply to the processing within the territory of the People's Republic of China of the personal information of natural persons. It shall also apply to the processing outside the territory of PRC of the personal information of natural persons located within the territory of PRC if the information is processed: (1) for the purpose of providing products or services to natural persons located within China (2) to analyze or assess the conduct of natural persons located within China (3) under any other circumstance as provided by any law or administrative regulation.

In addition, Article 4 states that Personal information processing includes, but is not limited to, the collection, storage, use, processing, transmission, provision, disclosure, and deletion of personal information.

8. What is the applicable remedy to infringement of personal information rights?

The Personal Information Protection Law of the People's Republic of China (effective since August 20, 2021) sets up legal liability and remedy to personal information rights infringement.

1) Administrative punishment

According to Article 66 of Personal Information Protection Law, departments fulfilling personal information protection duties and responsibilities have the right to order correction, issue a warning, confiscate unlawful income, and order the suspension or termination of service. When such correction is refused, a fine under 1 million CNY will be imposed, and the directly responsible person in charge and other directly responsible person will be fined from 10,000 to 100,000 CNY.

In grave circumstances, the fine for the entity can exceed 50 million CNY, or 5% of annual revenue. The directly responsible person in charge and other directly responsible person will be fined from 100,000 to 1 million CNY, and may also face some employment restrictions.

2) Civil liability

According to Article 69 of Personal Information Protection Law, infringers shall take responsibility for the infringement through compensation, the amount of which shall be determined according to the resulting loss to the individual or the resulting gains of the personal information handler. Where the loss to the individual and the gains to the personal information handler are difficult to determine, compensation shall be determined according to practical conditions.

3) Criminal liability

According to Article 71 of Personal Information Protection Law, when the infringement constitutes a crime, the infringers will face criminal liability according to the law.

4) Cases where the right of many individuals are infringed

According to Article 70 of Personal Information Protection Law, in case where the rights of many individuals are infringed, the People's Procuratorates, statutorily designated consumer organizations, and organizations designated by the national cyberspace authority may file a lawsuit according to the law.

9. Whether Cookie belongs to protective personal information?

There is possibility that Cookie may be deemed as PI in practice.

Article 76 of the Cybersecurity Law (effective since June 1, 2017) stipulates that PI refers to any information, recorded electronically or otherwise, that can be used alone or in combination with other information to identify a natural person or reflect the activities of a natural person, including but not limited to the name, date of birth, ID number, personal biometric information, residential address, contact information, communication records and content, user names and passwords, property information, credit information, records of whereabouts, accommodation information, health and physiological information, and transaction information.

To determine whether a particular piece of information is PI, the following two approaches shall be considered: First, information identifying individuals, that is, PI is the information that could help one identify a specific natural person through the specificity of the information; Second, information associated with individuals, that is, the information generated in the activities of a known natural person (such as the person's location information, call logs and browsing history) is PI. Information that meets either of the two criteria above shall be determined as PI.

Section 3.16 of Information Security Technology-Personal Information (PI) Security Specification (effective since Oct 1, 2020) included web browsing history as a subset of personal display.

Moreover, in the Guide to the Self-Assessment of Illegal Collection and Use of Personal Information by Apps (effective since July, 2020), Section 2.1(b) requires that if Cookies and similar technologies are used to collect personal information, the App shall make it clear to users the purpose and type of personal information collected.

In practice, to avoid being identified as PI, businesses usually specify in the use of Cookies disclaimers that "the information collected by using cookies does not include information that can identify specific individuals". However, personal browsing records do reflect personal interests, demand and other private information that can identify users to a certain extent. Such risk is now increasing as technology develops and Cookie may be recognized as PI in law enforcement practice if its use makes the users identifiable, satisfying PI's definition. In general, whether Cookie belongs to PI requires case-by-case analysis.

10. What kind of cyber protection is available for children's personal information?

Apart from the general protection offered in other laws and regulations, such as the *Civil Code* and the *Personal Information Protection Law*, there is a special regulation called the

Provisions on the Cyber Protection of Children's Personal Information (the "**Provisions**"), which was effective on Oct. 1, 2019. The following are some important articles in the Provisions:

1) Article 2: "child" means a minor under the age of 14.

2) Article 3: the Provisions shall apply to the collection, storage, use, transfer and disclosure of personal information from and about children through the Internet and other related activities within the territory of the China.

3) Article 8: a network operator shall develop specific rules and user agreements for the protection of children's personal information, and assign dedicated personnel responsible for protecting the children's personal information.

4) Article 9: a network operator collecting, using, transferring or disclosing any child's personal information shall notify the child's guardian in a conspicuous and clear manner, and obtain verified consent from the child's guardian for the collection, use, transfer or disclosure of personal information of the child.

5) Article 13: a network operator shall take measures such as encryption to store children's personal information, so as to ensure information security.

6) Article 20: where a child or his or her guardian requires a network operator to delete the child's personal information collected, stored, used or disclosed by it, the network operator shall take measures to delete such information in a timely manner.

11. What's the requirement for cross-border PII transfer?

In summary, the processor shall first inform the individual about the recipients' information and obtain his separate consent. Second, it shall satisfy one of the statutory conditions. Third, the provider must ensure the overseas recipient meet the protection standards under PIPL, maybe through signing warranty terms in contracts.

First, according to PIPL Article 39, Where a personal information processor provides personal information to any party outside the territory of the People's Republic of China, it or he shall notify individuals of the overseas recipient's name and contact information, purposes and methods of processing, categories of personal information, the methods and procedures for individuals' exercise of the rights provided in this Law over the overseas recipient, and other matters, and obtain individuals' separate consent.

Second, the statutory conditions are stipulated in PIPL Article 38 that where a personal information processor truly needs to provide personal information to any party outside the territory of the People's Republic of China for business or other needs, it or he shall meet any of the following conditions: (1) It or he has passed the security assessment organized by the national cyberspace administration in accordance with Article 40 of this Law. (2) It or he has been subject to the personal information protection certification by a specialized institution in accordance with the provisions issued by the national cyberspace administration. (3) It or he has entered into a contract with the overseas recipient in accordance with the model contract developed by the

national cyberspace administration, agreeing on both parties' rights and obligations. (4) It or he meets other conditions provided in laws or administrative regulations or by the national cyberspace administration.

Third, the personal information processor shall take necessary measures to ensure that personal information processing activities of the overseas recipient meet the personal information protection standards provided in this Law.

Last but not least, data protection impact assessment ("**DPIA**") shall be conducted pursuant to Article 55. Article 56 requires DPIA to include the following:

> *"(1) Whether the purposes and methods of processing of personal information, among others, are lawful, legitimate and necessary.*
>
> *(2) The impacts on individuals' rights and interests and security risks.*
>
> *(3) Whether the protection measures taken are lawful, effective, and commensurate with the degrees of risks."*

Such DIPA reports and records on processing shall be preserved for at least three years.

E. Prohibited content

1. What types of data may be prohibited from being text data mined?

As far as "data mining" is concerned, China has not yet established laws and regulations to restrict it. Therefore, there is no data that is absolutely prohibited by law from being mined. In other words, whatever data is mined, the fact that data mining itself has not yet triggered the law and does not require any liability.

However, there are still several aspects of legal risks that need to pay attention to.

First, data mining should not violate laws and regulations, social morality and ethics, business ethics, and professional ethics. Article 8 of the Data Security Law of the People's Republic of China (which came into force on September 1, 2021) stipulates that in carrying out data processing activities, it shall abide by laws and regulations, respect social morality and ethics, observe business ethics and professional ethics, be honest and trustworthy, fulfill data security protection obligations, undertake social responsibilities, and shall not endanger state security and public interests, or damage the lawful rights and interests of individuals and organizations.

Secondly, data mining cannot use theft and other illegal means. Article 32 rules that any organization or individual shall collect data in lawful and proper ways and shall not steal or obtain data by other illegal means.

Finally, data belonging to controlled items are subject to import and export controls. Article 25 stipulates that the State exercises export control in accordance with the law over data of controlled items related to safeguarding national security and interests and fulfilling international obligations. To sum up, from the existing legal system, there is no data that is absolutely prohibited from mining, but there are still legal risks in the process of data mining that need to be paid attention to.

2. What types of data may be prohibited from exportation?

According to Article 37 of the Cybersecurity Law of the People's Republic of China (effective on November 6, 2016), critical information infrastructure operators that gather or produce personal information or important data during operations within the mainland territory of the People's Republic of China, shall store it within mainland China. Where due to business requirements it is truly necessary to provide it outside the mainland, they shall follow the measures jointly formulated by the State cybersecurity and informatization departments and the relevant departments of the State Council to conduct a security assessment.

Article 2 of the Regulation on Protecting the Security of Critical Information Infrastructure gave specific definition to critical information infrastructure ("CII"), that CII means any of network facilities and information systems in important industries and fields—such as public communication and information services, energy, transportation, water conservancy, finance, public services, e-government, and science, technology and industry for national defense—that may seriously endanger national security, national economy and people's livelihood, and public interests in the event that they are damaged or lose their functions or their data are leaked.

3. What types of data may be prohibited from being generated?

Under the current legal framework of the PRC, it seems there is no specific prohibition against the generation of the data. However, please note that there are restrictions on the content of the data. Restrictions that are noteworthy are as below:

Article 12 of the Cybersecurity Law of the People's Republic of China ("Cybersecurity Law"): The state shall protect the rights of citizens, legal persons and other organizations to use the network in accordance with the law, promote the popularity of network access, provide better network services, provide the public with safe and convenient network services, and guarantee the orderly and free flow of network information in accordance with the law.

Any individual or organization using the network shall comply with the Constitution and laws, follow public order and respect social morality, shall not endanger cybersecurity, and shall not use the network to conduct any activity that endangers national security, honor and interest, incites to subvert the state power or overthrow the socialist system, incites to split the country or undermine national unity, advocates terrorism or extremism, propagates ethnic hatred or discrimination, spreads violent or pornographic information, fabricates or disseminates false

information to disrupt the economic and social order, or infringes upon the reputation, privacy, intellectual property rights or other lawful rights and interests of any other person.

Article 13 of the Cybersecurity Law: The state shall support the research and development of network products and services that are conducive to the healthy growth of minors, legally punish the activities that damage the physical and mental health of minors by using the network, and provide a safe and healthy network environment for minors.[26]

4. What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?

The prohibited content, as provided in the above questions is generally vague and open for interpretation in legal practices, serves as implications of social values including social morality and ethics, business ethics and the legal interests of other individuals, advocating honesty which is generally applicable in other areas of law. Specifically, the main policy concerns here are national security, individual privacy and protection of the lawful rights of minors in a paternalistic way.

The current regulation framework aims to prevent exploitation of personal information, especially that of minors who call for more protection, behaviors that might endanger national security and unfair business competition in the sense that some companies might exploit information of users in other internet platforms to gain unfair competitive advantages.

5. What types of licensing or other control mechanisms would be preferred under the applicable jurisdictions?

As mentioned in Question B1, data sets and models can be licensed by contract, thus the terms and conditions of the contract is the key to the control of the license or use of the data sets and models. When negotiating the terms and conditions of the license contract, special attention should be paid to the compliance issues in the PRC. Specifically, it is advisable for the owner of the data sets or models to add some restrictions on the license based on the laws discussed in the above question.

6. Is there any ethics requirement of AI?

There is. In June, 2021, the Ministry of Science and Technology issued the *Ethics Specifications of the New Generation of AI*. The document proposed six general principles, and 18 specific requirements applicable to the management, R&D, supply, and usage of AI technology. Among those specific requirements, it's specified that AI development shall not violate ethics, shall not violate IP rights, and shall not violate national security. However, those definitions are not explained in that document.

[26] Cybersecurity Law of the People's Republic of China: 中华人民共和国网络安全法

Conclusion

A. IP Questions

Original data are protected by copyright law. Similarly, as a compilation of data, the original data training set will be protected by copyright law. However, the data training set can be protected as a trade secret. On the contrary, the data training model can be protected by copyright. T The attribution of data-based rights should be discussed categorically: (1) personal data, (2) government data, and (3) corporate data.

Researchers may infringe the following intellectual property rights during data processing: first, publishing data sets containing HTML tags or document structures may involve violations of the right to disseminate works to the public through the information network, especially when these HTML tags insert pictures or hyperlinks in the text. Second, C4 or Oscar can be understood as computer programs. Therefore, when publishing a dataset that references a location in another dataset, even if the availability is limited, attention should be paid to the potential violation of the protected rights of computer software. Third, direct access to information from people may also infringe intellectual property rights. Depending on the different levels of input from the interviewer and the interview, the interview manuscript may be owned by the interviewee or jointly owned by the interviewer and the interviewee. The best way for the data collector to avoid any disputes in the future is to obtain the consent of the interviewee. Fourth, borrowed books are protected by copyright, so researchers using borrowed books to train language models may infringe intellectual property rights. Finally, China's existing copyright system cannot provide immunity for text data mining, so text data mining is likely to infringe copyright and other related property rights.

B. Licensing Questions

Data licenses are agreements where the data owner authorizes others to use the data within a certain period and scope without changing the ownership. Since data ownership has not been settled down in law, the owner can be data subject or data processor according to the agreement between the data subject and the processor. Both datasets and models are licensable through contracts, and the restriction can be set as terms such as purpose of use, scope of use and exclusivity. Licenses are contracts, which is less binding than law. Thus, they may be invalid under the void or voidable conditions stated in laws. The relationship between licenses and terms of use is a little tricky. In circumstances where the data subject signed the terms of use with platforms which stipulates that the data can’t be shared, and then licensed third-party to use the data, the platform can claim breach of contract against the data subject. Generally speaking, NLP researchers can use licensed data to train their language model, and they also enjoy the copyright of the model as long as they state clearly on the license the terms related to use. Furthermore, licenses can be royalty free according to agreements.

C. Text Data Mining and Fair Use Questions

Currently, there is no specific provision regulating data mining and data scraping. The recently enacted Data Security Law of PRC requires NLP researchers to comply with laws and regulations of China, respect social norms and ethics, and observe business and professional

ethics when collecting data. As the policy concern lies in intellectual property violation and the potential risk of intruding personal privacy, researchers should be aware of any relative violation in crawling data and abide by the IP laws and Personal Information Protection Law. In legal practice, courts generally apply Article 2 of Anti-Unfair Competition Law to regulate unlawful data mining behavior when it harms the legitimate rights and interests of other operators or that of consumers, when it violates the principle of good faith and recognized business ethics and when such behavior destroys the open and fair market competition order in the Internet environment. As social media platforms are another important stakeholder in this context, NLP researchers should note that most platforms, in its Terms of Use, do not allow data scraping of social media content without the consent of the operators even when the user gives his permit already. Moreover, by reading into the courts' opinion in *Sina Weibo v. Maimai,* courts tend to require triple authorization in this context. Generally, researchers will be insured in a safety belt if it obtains authorization of use from stakeholders before starting the data mining activities.

D. Privacy Questions

The legal concern around accessing web content and datasets that have PII are mainly provided in the PRC Personal Information Protection Law ("PIPL"). In general, when handling personal information, individual consent shall be obtained except in few case. Providing PII to other processors or making PII public by any means requires the processor to notify the data subject and obtain the separate consent of him. Such individual consent shall be given by individuals under the precondition of full knowledge, and in a voluntary and explicit statement.

According to Article 6 of the PIPL, personal information handling shall have a clear and reasonable purpose, and shall be directly related to the handling purpose, using a method with the smallest influence on individual rights and interests. The collection of personal information shall be limited to the smallest scope for realizing the handling purpose, and excessive personal information collection is prohibited.

The application scope of PIPL is extra-territorial. Apart from all the processing activities happening in China, the activities happening outside China shall comply with the PIPL if it provides products or services to people within China or analyzes their behaviors.

For special protection of children's personal information, apart from the general protection offered in other laws and regulations, such as the Civil Code and the Personal Information Protection Law, there is a special regulation called the Provisions on the Cyber Protection of Children's Personal Information, which states that a network operator collecting, using, transferring or disclosing any child's personal information shall notify the child's guardian in a conspicuous and clear manner, and obtain verified consent from the child's guardian for the collection, use, transfer or disclosure of personal information of the child.

The applicable remedy to infringement of personal information rights include administrative punishment, civil liability and criminal liability.

E. Prohibited content including state secrets, obscenity and child pornography

Although the legislation in data security is a relatively new area in China and the legal framework is just established with a series of newly adopted laws and regulations in recent years, there are some restrictions on handling data (including collection, storage, distribution of data, data mining, etc.) that need to be addressed. First, data handling shall not violate laws and regulations, social morality and ethics, business ethics, and professional ethics. Second, data exportation is subject to the State's control and shall follow certain measures formulated by the State, especially information related to personal identity, national security and critical infrastructure. Third, the laws provide more protection to minors by imposing more requirements of handling minors-related data. Fourth, terms and conditions of the license contract are the key to the control of licensing use of data sets and models. Last, the newly issued Ethics Specifications of the New Generation of AI sets out some ethical requirements on conducts related to AI.

SOUTH KOREA

Jihyun Kang

I. Introduction of South Korean legal system

➢ Overview of South Korean legal system

- Civil law system : codified legal system
  ☞ Statutes are the primary source of law in South Korea.
- Hierarchy of Acts & Subordinate Statutes

Reference : ABOUT KOREAN LAW

☞ The Constitution is the basic law of the country and articulates the law-making powers and procedures related to such powers(Young-Hee Kim, “Introduction to Korean Legal Materials”, Journal of Korean Law, Vol.2, No.1, 2002, p.139)

☞ Since the Acts and subordinate statutes form a certain hierarchy, subordinate statutes that are enacted under powers delegated by Acts or are enacted for the purpose of enforcing Acts are not permitted to contain details in conflict with such Acts(https://www.law.go.kr/LSW/eng/engAbout.do?menuId=3).

☞ Treaties duly concluded and promulgated under the Constitution and generally recognized rules of international law have the same effect as the domestic laws of South Korea(Constitution of South Korea Article 6).

- Legislative Power

  ☞ Both the National Assembly and the Executive Branches can make a bill, but only the National Assembly can make a law(Young-Hee Kim, “Introduction to Korean Legal Materials”, Journal of Korean Law, Vol.2, No.1, 2002, p.139).

  ☞ In the meanwhile, the President and the Ministries of the Executive Branch can promulgate secondary laws which are referred to as decrees or ordinances(Young-Hee Kim, “Introduction to Korean Legal Materials”, Journal of Korean Law, Vol.2, No.1, 2002, p.139).

- **No principle of stare decisis or precedent**

  ☞ The Korean legal system does not recognize the principle of *stare decisis*, meaning that precedents are not legally binding *per se.*
  (https://www.nyulawglobal.org/globalex/South_Korea1.html)

  ☞ A Higher court’s decision in the judicial hierarchy prevails over that of the lower court’s decision ‘only’ on the specific case concerned.
  (https://www.nyulawglobal.org/globalex/South_Korea1.html)

  ☞ However, as the lower courts tend to follow the legal interpretations ascertained by the Supreme Court of Korea in actual practice, the Supreme Court decisions are regarded as the secondary source of law in practice.
  (https://www.nyulawglobal.org/globalex/South_Korea1.html).

- **Statutory regulations related to personal data protection in South Korea**

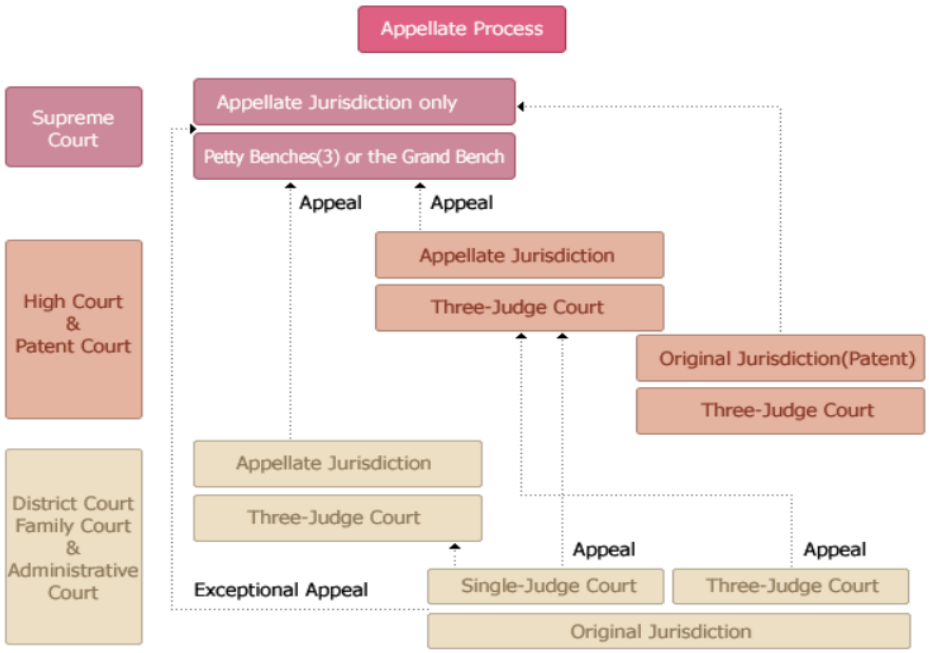
  ☞ South Korea has strict statutory regulations regarding personal data in Personal Information Protection Act.

➢ Introduction of South Korean Court System

(https://eng.scourt.go.kr/eng/judiciary/introduction.jsp)

- The Constitutional Law of Korea provides that the judiciary consists of the Supreme Court of Korea, the highest court of the nation and other courts.
- **Three-tiered system** : The judiciary in Korea is a three-tiered system, which is composed of district courts, the high courts and the Supreme Court. Supreme Court Justices are appointed by the President and the Supreme Court is known as the court of last resort.
- The first instance court is the district court with a single or a three-judge panel, the latter tries more serious, important, and high value claim cases. The court of appeal (also known as High Court) and the Supreme Court are appellate courts.

- Patent Court, Family Court and Administrative Court exercise specialized functions with the Patent Court positioned on the same level as the high courts and the family court and the administrative court positioned on the same level a the district courts.
- Except in military courts, adjudication including hearings and rendering judgment is presided by a judge. Case trials are presided either by a single judge or a panel of three judges.
- All cases are decided by judges, as no jury system exists in South Korea.
- In general, all hearings and rendering of judgments are open to the public.



The Judiciary > Introduction

- Outside of the normal judiciary chain, the Constitutional Court, as a separate constitutional organization, enjoys the same status as the Supreme Court vis a vis constitutional matters.
- The Constitutional Court handles constitutional issues such as the constitutionality of a law, impeachment, dissolution of a political party, constitutional petitions filed directly to the Constitutional Court, and jurisdictional conflicts involving State agencies and/or local governments.
- The Constitutional Court is comprised of 9 justices, also appointed by the President. This court hears constitutional issues under the country's Constitution.

➢ Useful information
- Korean laws in English

Korea Law Translation Center

ENGLISH | Easy to Find, Practical Law

☞ You can research Korean law in English at the Korea Law Translation Center (KLT) of the Korea Legislation Research Institute and Practical Law website of the Ministry of Government Legislation.

※ The Ministry of Government Legislation, the administrative organization that oversees and coordinates government legislation.

※ The Korea Legislation Research Institute operates for the purpose of supporting the national legislative policies and promoting timely and accurate dissemination of legislative information as well as assisting general legislative activities by systematically collecting and managing legislative information and surveying or researching juristic and legislative issues with extensive expertise.

- Recently Passed Bills
  The National Assembly of The republic of Korea > LAWS & BILLS > Recently Passed Bills
  ☞ You can find recently passed bills at the National Assembly of the Republic of Korea at this website.
- Introduction to Korean Legislative System and Procedures
  https://www.law.go.kr/LSW/eng/engAbout.do?menuId=3
  ☞ You can find brief information about Korean Legislative System and Procedures at Korean Law Information Center of the Ministry of Government Legislation.
- Supreme Court Decisions
  Supreme Court Library
  ☞ Supreme Court decisions are published by the Supreme Court Library of Korea. You can research leading or latest Supreme Court decisions at this website.
- Constitutional Court Decisions
  https://search.ccourt.go.kr/ths/ep/selectThsEp0101List.do
  ☞ You can research leading or latest Constitutional Court decisions at this website.

- Useful website
  http://koreanlii.or.kr/w/index.php/Main_Page?ckattempt=1
  ☞ You can find useful information about Korean law in English and useful links of website.

A. IP Questions

1. Are the data training sets and models protected by IP rights and if so which IP rights?

Executive Summary

Yes, If data training sets and models express human thoughts and emotions and have a creative nature, it can be protected by copyrighted works of the Copyright Act in South Korea.

Rule

Copyright Act
Article 2 (Definitions)
The terms used in this Act shall be defined as follows: <Amended on Apr. 22, 2009; Jun. 30, 2011; Dec. 2, 2011; Mar. 22, 2016>
1. The term “work” means a creative production that expresses human thoughts and emotions;

Article 4 (Examples of Works)
(1) The following shall be the examples of works referred to in this Act:
1. Novels, poems, theses, lectures, speeches, plays and other literary works;
2. Musical works;
3. Theatrical works including dramas, choreographies, pantomimes, etc.;
4. Paintings, calligraphic works, sculptures, printmaking, crafts, works of applied art, and other works of art;
5. Architectural works including buildings, architectural models and design drawings;
6. Photographic works (including those produced by similar methods);
7. Cinematographic works;
8. Maps, charts, design drawings, sketches, models and other diagrammatic works;
9. Computer program works.
(2) Deleted. <Apr. 22, 2009>

Analysis

Article 2(Definitions) of Copyright Act in South Korea define that ‘work’ means a creative production that expresses human thoughts and emotion. Article 4(Examples of Works) just suggests an example of work, so copyright works never confined that example. Therefore, if data training sets and models contain human thoughts and emotion, and they have a creative nature, it can be protected by the Copyright Act of South Korea.

Conclusion

So, data training sets and models express human thoughts and emotions and have a creative nature. It can be protected by copyrighted works of the Copyright Act in South Korea.

2. What are the IP risks related to data collection directly from persons? For example when you interview people or they donate data etc.

Executive Summary

If the person providing the information is not the owner of IP rights or the owner cannot provide the data because of contractual restriction with a third party, it may result in aiding in IP infringement.

Rule

Civil Act of South Korea Article 760(Liability of Joint Tort-feasors)

(1) If two or more persons have by their joint unlawful acts caused damages to another, they shall be jointly and severally liable to make compensation for such damages.
(2) The provisions of paragraph (1) shall also apply if it is impossible to ascertain which of the participants, albeit not joint, has caused the damages.
(3) Instigators and accessories shall be deemed to act jointly.

Analysis

If the person providing the information is not the owner of IP rights or the owner cannot provide the data because of contractual restriction with a third party, it may result in aiding in IP infringement.

Also, it is not exactly the IP risk, there is a risk that errors in the training data collected directly from persons can cause errors. This problem is similar to the so-called overfitting problem. To solve this problem, verification data is required to verify the accuracy of the collected data.

Conclusion

It can be possible that data mining directly from a person consists of torts as an accessory if the person does not have a right to the data.

B. Licensing Question

1. **Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?**

Executive Summary

I think it is possible because there is no legal restriction on licensing data sets in South Korea.

Rule

License means the right to use intellectual property rights as well as commercially protected rights. There is no legal restriction on the kind of licenable rights. There are clauses related to licensing of patents, but it is not the restriction on the kind of rights.

Analysis

A license means the right to use intellectual property as well as commercially protected rights. There is no legal restriction on the kind of licensable rights in South Korea. License contract is just one of a kind commercial agreement. According to the license agreement, the licensor grants the licensee the right to use its exclusive rights, and the licensee forms a consideration relationship in which the licensee pays royalties. The content of the license agreement can be freely determined by the parties in accordance with the principle of freedom of contract. Therefore, data sets can be licensed from a legal perspective.

Conclusion

I think it is possible because there is no legal restriction on licensing data sets in South Korea.

2. **What are the relationships between laws, licenses, and terms of use? Which is more binding?**

Executive Summary

If related rules are overriding mandatory rules, laws are more binding than licenses and terms of use in South Korea. However, If related rules are default rules, it can be possible licenses or terms of use override the law. In the meanwhile, licenses can override terms of use by stipulating prevalence over terms and conditions.

Rule

Article 4 (Precedence of Individual Agreement)
If a business person and a customer agree on a matter in a manner that is different from the manner stipulated in the terms and conditions, the agreement shall prevail over such terms and conditions.

Analysis

If related rules are overriding mandatory rules, laws are more binding than licenses and terms of use in South Korea. However, If related rules are default rules, it can be possible licenses or terms of use override the law. In the meanwhile, licenses can override terms of use by stipulating prevalence over terms and conditions. License contract is one of a kind commercial contract, and Act on the Regulation of Terms and Conditions of South Korea Article 4 stipulate “If a business person and a customer agree on a matter in a manner that is different from the manner stipulated in the terms and conditions, the agreement shall prevail over such terms and conditions.”. Therefore, if a license contract stipulating it prevails over terms and conditions, the license contract overrides terms and use.

Conclusion

If related rules are overriding mandatory rules, laws are more binding than licenses and terms of use in South Korea. However, If related rules are default rules, it can be possible licenses or terms of use override the law. In the meanwhile, licenses can override terms of use by stipulating prevalence over terms and conditions.

C. Text Data Mining and Fair Use Questions

1. What are types of legally permitted text data mining?

Executive Summary

If text data mining does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, it would fall under the 'Fair Use' of Copyright Act in South Korea.

Rules

Copyright Act
Article 35-5 (Fair Use of Works)
(1) Except as provided in Articles 23 through 35-4 and 101-3 through 101-5, where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, he or she is entitled to use such works. <Amended on Mar. 22, 2016; Nov. 26, 2019>
(2) In determining whether an act of using works falls under paragraph (1), the following matters shall be considered: <Amended on Mar. 22, 2016>
1. Purposes and characteristics of use;
2. Types and purposes of works;
3. Amount and substantiality of portion used in relation to the whole works;
4. Effect of the use of works on the existing or potential market for the works or current or potential value thereof.
[This Article Newly Inserted on Dec. 2, 2011]
[Moved from Article 35-3 <Nov. 26, 2019>]

Analysis

Article 35-5(Fair Use of Works) stipulates that where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, it can be possible to use such works. As a result, if text data mining does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, it would fall under the 'Fair Use' of Copyright Act in South Korea.

Conclusion

Therefore, under the current Copyright Act, text data mining should meet the fair use requirements to fall under fair use.

2. Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can't be scrapped?

Executive Summary

No. There is no meaningful legal difference between kinds of social media.

Rule

Personal Information Protection Act

Article 15 (Collection and Use of Personal Information)
(1) A personal information controller may collect personal information in any of the following circumstances, and use it with the scope of the purpose of collection:
1. Where consent is obtained from a data subject;
2. Where special provisions exist in other laws or it is inevitable to observe legal obligations;
3. Where it is inevitable for a public institution's performance of its duties under its jurisdiction as prescribed by statutes, etc.;
4. Where it is inevitably necessary to execute and perform a contract with a data subject;
5. Where it is deemed manifestly necessary for the protection of life, bodily or property interests of the data subject or third party from imminent danger where the data subject or his or her legal representative is not in a position to express intention, or prior consent cannot be obtained owing to unknown addresses, etc.;
6. Where it is necessary to attain the justifiable interest of a personal information controller, which such interest is manifestly superior to the rights of the data subject. In such cases, processing shall be allowed only to the extent the processing is substantially related to the justifiable interest of the personal information controller and does not go beyond a reasonable scope.
(2) A personal information controller shall inform a data subject of the following matters when it obtains consent under paragraph (1) 1. The same shall apply when any of the following is modified.
1. The purpose of the collection and use of personal information;
2. Particulars of personal information to be collected;
3. The period for retaining and using personal information;
4. The fact that the data subject is entitled to deny consent, and disadvantages, if any, resulting from the denial of consent.
(3) A personal information controller may use personal information without the consent of a data subject within the scope reasonably related to the initial purpose of the collection as prescribed by Presidential Decree, in consideration whether disadvantages have been caused to the data subject and whether necessary measures have been taken to secure such as encryption, etc. < This Article Newly Inserted by Act No. 16930, 4. February, 2020 >

Copyright Act

Article 35-5 (Fair Use of Works)
(1) Except as provided in Articles 23 through 35-4 and 101-3 through 101-5, where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, he or she is entitled to use such works. <Amended on Mar. 22, 2016; Nov. 26, 2019>
(2) In determining whether an act of using works falls under paragraph (1), the following matters shall be considered: <Amended on Mar. 22, 2016>
1. Purposes and characteristics of use;
2. Types and purposes of works;
3. Amount and substantiality of portion used in relation to the whole works;
4. Effect of the use of works on the existing or potential market for the works or current or potential value thereof.
[This Article Newly Inserted on Dec. 2, 2011]
[Moved from Article 35-3 <Nov. 26, 2019>]

Military Secret Protection Act

Article 12 (Leakage)
(1) If any person who has detected or collected military secrets leaks them to others, he/she shall be punished by imprisonment with labor for a limited term of not less than one year.
(2) If any person who has come to know or possess military secrets by chance leaks them to others despite knowledge that they are the military secrets, he/she shall be punished by imprisonment with labor for not more than five years or by a fine not exceeding 50 million won. <Amended by Act No. 12556, May 9, 2014>

Act On Promotion Of Information And Communications Network Utilization And Information Protection

Article 44-7 (Prohibition on Circulation of Unlawful Information)
(1) No one may circulate any of the following information through an information and communications network: <Amended on Sep. 15, 2011; Mar. 22, 2016; Jun. 12, 2018>
1. Information with obscene content distributed, sold, rented, or displayed openly in the form of code, words, sound, images, or motion picture;
2. Information with content that defames other persons by divulging a fact or false information, openly and with intent to disparage the person's reputation;
3. Information with content that arouses fear or apprehension by reaching other persons repeatedly in the form of code, words, sound, image, or motion picture;
4. Information with content that compromises, destroys, alters, or forges an information and communications system, data, a program, or similar or that interferes with the operation of such system, data, program, or similar without good cause;
5. Information with content that amounts to a media product harmful to youths under the Youth Protection Act and that is provided for profit without fulfilling the duties and obligations under the relevant statutes and regulations, including the duty to verify the subject's age and the duty of labeling;

6. Information with content that amounts to speculative activities prohibited by statutes and regulations;
6-2. Information with content of transactions of personal information in violation of this Act or any other statute or regulation regarding the protection of personal information;
6-3. Information regarding methods, drawings, etc. for manufacturing guns or explosives (including things with a yield that may expose people to risk of life or bodily injury);
7. Information with content that divulges a secret classified under statutes and regulations or any other State secret;
8. Information with content that violates the National Security Act;
9. Other information with content that attempts to commit, aids, or abets a crime.

Act On The Protection Of Children And Youth Against Sex Offenses

Article 11 (Production or Distribution of Child or Youth Sexual Exploitation Materials)
(1) Any person who produces, imports, or exports child or youth sexual exploitation materials shall be punished by imprisonment with labor for an indefinite term or for a limited term of at least five years. <Amended on Jun. 2, 2020>
(2) Any person who sells, lends, distributes, or provides child or youth sexual exploitation materials for commercial purposes, or possesses, transports, advertises or introduce them for any of such purposes, or publicly exhibits or displays them shall be punished by imprisonment with labor for not more than five years. <Amended on Jun. 2, 2020>
(3) Any person who distributes or provides child or youth sexual exploitation materials, advertises or introduces them for any of such purposes, or publicly exhibits or displays them shall be punished by imprisonment with labor for not more than three years. <Amended on Jun. 2, 2020>
(4) Any person who procures a child or youth for a child or youth sexual exploitation materials producer, knowing that he or she is to be used for producing child or youth child or youth sexual exploitation materials, shall be punished by imprisonment with labor for at least three years. <Amended on Jun. 2, 2020>
(5) Any person who purchases child or youth sexual exploitation materials or possesses or views them with the knowledge that it is a child or youth sexual exploitation materials, shall be punished by imprisonment with labor for at least one year. <Amended on Jun. 2, 2020>
(6) Any person who attempts to commit an offense prescribed in paragraph (1) shall be punished.
(7) Any person who habitually commits offenses referred to in paragraph (1) shall be subject to an aggravated punishment by up to 1/2 of the penalty stipulated for such offense. <Newly Inserted on Jun. 2, 2020>
[Title of This Article Amended on Jul. 2, 2020]

Analysis

Because related regulations such as Personal Information Protection Act, Copyright Act, Military Secret Protection Act, Act On Promotion Of Information And Communications Network Utilization And Information Protection, Act On The Protection Of Children And Youth Against Sex Offenses are all care about content and nature of the content. Social media is just a tool to convey

expression and it is not necessarily related to the nature of the content. Therefore, type of social media is not important from a legal perspective.

Conclusion

There is no meaningful legal difference between kinds of social media.

3. Additional Question : Does the copyright law allow so-called non-contractual use of data without the consent of the right holder?

Executive Summary

It can be possible if non-contractual use falls under the fair use doctrine under the current Copyright Act of South Korea. However, there are arguments that the Copyright Act should be revised to make a separate provision for data mining that can be possible without the consent of the right holder.

Rules

Copyright Act
Article 35-5 (Fair Use of Works)
(1) Except as provided in Articles 23 through 35-4 and 101-3 through 101-5, where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, he or she is entitled to use such works. <Amended on Mar. 22, 2016; Nov. 26, 2019>
(2) In determining whether an act of using works falls under paragraph (1), the following matters shall be considered: <Amended on Mar. 22, 2016>
1. Purposes and characteristics of use;
2. Types and purposes of works;
3. Amount and substantiality of portion used in relation to the whole works;
4. Effect of the use of works on the existing or potential market for the works or current or potential value thereof.
[This Article Newly Inserted on Dec. 2, 2011]
[Moved from Article 35-3 <Nov. 26, 2019>]

Analysis

In principle, in order to secure the legality of the use of data, it is needed to obtain the consent of the right holder. However, It is practically impossible to identify the object of copyright among numerous individual data and obtain the consent of the right holder. Under the current Copyright Act allow non-contractual use when it falls under the fair use doctrine. Article 35-5(Fair Use of Works) stipulates that where a person does not unreasonably undermine an author’s legitimate interest without conflicting with the normal exploitation of works, it can be possible to use such works. As a way to solve this problem, ‘fair use’ is suggested as a useful legal doctrine. However, in order to improve predictability, a separate provision for data mining needs to be prepared. Many countries including the United Kingdom, Germany, the European Union, the United States, and

Japan have already introduced such provisions, and a revised copyright law has been submitted in Korea for this purpose.

However, there are many arguments and legislative attempts to allow data mining without the consent of the right holder in separate provision for data mining in order to improve predictability. Many countries including United Kingdom, Germany, the European Union have already introduced such provisions, and there are legislative attempts in South Korea.

※ Related scholarly article : Sang Yong Lee, “Non-Contractual Use of Data - Focusing on Limitations to Copyright for Text and Data Mining”, Kangwon Law Review 65(2021. 11.)

Conclusion

Therefore, under the current Copyright Act allow non-contractual use when it falls under the fair use doctrine, but there are arguments that the Copyright Act should be revised to make a separate provision for data mining can be possible without the consent of the right holder.

4. Additional Question : Should it be limited to academic and non-commercial use?

Executive Summary

Should not be limited to academic and non-commercial use to fall under the fair-use doctrine.

Rule

Copyright Act

Article 35-5 (Fair Use of Works)
(1) Except as provided in Articles 23 through 35-4 and 101-3 through 101-5, where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, he or she is entitled to use such works. <Amended on Mar. 22, 2016; Nov. 26, 2019>
(2) In determining whether an act of using works falls under paragraph (1), the following matters shall be considered: <Amended on Mar. 22, 2016>
1. Purposes and characteristics of use;
2. Types and purposes of works;
3. Amount and substantiality of portion used in relation to the whole works;
4. Effect of the use of works on the existing or potential market for the works or current or potential value thereof.
[This Article Newly Inserted on Dec. 2, 2011]
[Moved from Article 35-3 <Nov. 26, 2019>]

Analysis

There is fierce debate over whether to grant copyright restrictions for data mining, even for commercial purposes in South Korea. Proponents argue that there is a need to develop the data industry. Opponents say that restricting copyright, even in the case of commercial research, could harm the legitimate interests of the copyright holder. In the case of creation and use of datasets for artificial intelligence learning, unless intermediate products are directly provided for transactions, they are transformative or non-expressive uses, which meet the public interest request for new technology development, so there is considerable room for fair use.

Conclusion

Should not be limited to academic and non-commercial use to fall under the fair-use doctrine.

D. Privacy Questions

1. Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?

Executive Summary

To access web content that has PII(Personal Identifiable Information) legally, it is required to comply with the Personal Information Protection Act of South Korea.

Rules

Personal Information Protection Act
Article 15 (Collection and Use of Personal Information)
(1) A personal information controller may collect personal information in any of the following circumstances, and use it with the scope of the purpose of collection:
1. Where consent is obtained from a data subject;
2. Where special provisions exist in other laws or it is inevitable to observe legal obligations;
3. Where it is inevitable for a public institution's performance of its duties under its jurisdiction as prescribed by statutes, etc.;
4. Where it is inevitably necessary to execute and perform a contract with a data subject;
5. Where it is deemed manifestly necessary for the protection of life, bodily or property interests of the data subject or third party from imminent danger where the data subject or his or her legal representative is not in a position to express intention, or prior consent cannot be obtained owing to unknown addresses, etc.;
6. Where it is necessary to attain the justifiable interest of a personal information controller, which such interest is manifestly superior to the rights of the data subject. In such cases, processing shall be allowed only to the extent the processing is substantially related to the justifiable interest of the personal information controller and does not go beyond a reasonable scope.

Analysis

There are strict requirements to collect personal information in South Korea. Personal Information Protection Act Article 15(Collection and Use of Personal Information) stipulates the requirements to collect personal information such as consent of the data subject.

[Leading Case] “Iruda” case

On April 28, 2021, the Personal Information Protection Commission of South Korea(“PIPC”) judged that the Scatter Lab Corporation’s act of using KakaoTalk conversations during the development and operation of an AI chatbot service “Iruda” violated the Personal Information Protection Act. PIPC imposed penalty surcharges of 55,500,000 won and administrative fines of 47,800,000 won on Scatter Lab.

Scatter lap Corporation uses KakaoTalk conversation sentences of about 600,000 users without deleting or encrypting personal information such as names, mobile phone numbers, and addresses included in KakaoTalk conversations for the purpose of learning the algorithm for developing the Iruda AI model. About 9.4 billion cases were used, and 100 million KakaoTalk conversation sentences of women in their 20s were constructed as a response DB for Iruda service operation so that Iruda could select and utter one of the sentences above.

KakaoTalk conversations are characterized by a high probability of including personal information such as real names and mobile phone numbers, and the possibility of identifying individuals through conversations that can infer human relationships and affiliations in addition to identification information, social login IDs, etc. Since member information such as social login ID and KakaoTalk conversations are collected and used together, KakaoTalk conversations are combined with the above member information and personal information included in the conversations to personal information that can identify the user who uttered a specific conversational sentence. For this reason, PIPC decided that the collected information corresponds to personal information, so personal information should have been deleted or encrypted under the Personal Information Protection Act.

PIPC believes that Kakao Talk conversations entered by one party to the conversation can be collected only with the consent of one party to the conversation, unless the member information of the other party is also collected. However, PIPC judged that it was in violation of Article 22 of the Personal Information Protection Act, as consent was not explicitly notified to the information subject while collecting conversations.

In addition, PIPC was judged that by including the development of new services in the personal information processing policy, the user consented by logging in did not correspond to obtaining the user's consent. Nevertheless, Scatter lab corporation used personal information to develop new services.The use of personal information beyond the scope of consent is prohibited by Article 18 of the Personal Information Protection Act.

A personal information controller may process pseudonymized information without the consent of data subjects for statistical purposes, scientific research purposes, and archiving purposes in the public interest, etc by Article 28-2 of the Personal Information Protection Act. However, PIPC judged that KakaoTalk conversations were not treated as pseudonyms, and that providing an external conversation service using KakaoTalk conversations for the operation of 'Iruda' does not fall under the 'scientific research' of Article 28-2 of the Personal Information Protection Act.

According to PIPC, the act of sharing a user's KakaoTalk conversation text with a third party is to provide pseudonymous information, including 'information that can be used to identify a specific individual', to an unspecified number of people. It was judged to be in violation of Article 28 of the Personal Information Protection Act Article 2 Paragraph 2 and ordered to take corrective action.

Reference : https://m.lawtimes.co.kr/Content/LawFirm-NewsLetter?serial=170076
https://www.pipc.go.kr/np/default/agenda.do?op=view&mCode=E030010000&page=37&isPre=&mrtlCd=&idxId=2021-0257&schStr=&fromDt=&toDt=&insttDivCdNm=&insttNms=&processCdNm=

Personal Information Protection Act
Article 18 (Limitation to Out-of-Purpose Use and Provision of Personal Information)
(1) A personal information controller shall not use personal information beyond the scope provided for in Articles 15 (1) and 39-3 (1) and (2), or provide it to any third party beyond the scope provided for in Article 17 (1) and (3). <Amended by Act No. 16930, Feb. 4, 2020>

Article 22 (Methods of Obtaining Consent)
(1) Where a personal information controller intends to obtain the consent of the data subject (including his or her legal representative as stated in paragraph (6): hereafter in this Article the same applies) to the processing of his or her personal information, the personal information controller shall present the request for consent to the data subject in a clearly recognizable manner where each matter requiring consent is distinctly presented, and obtain his or her consent thereto, respectively. <Amended by Act No. 14765, Apr. 18, 2017>

Article 28-2 (Processing of Pseudonymous Data)
(1) A personal information controller may process pseudonymized information without the consent of data subjects for statistical purposes, scientific research purposes, and archiving purposes in the public interest, etc.
(2) A personal information controller shall not include information that may be used to identify a certain individual when providing pseudonymized information to a third party according to paragraph (1).
[This Article Newly Inserted by Act No. 16930, Feb. 4, 2020]

Conclusion

Therefore, accessing web content that has PII needs to get a consent of the data subject. In South Korea, when collecting personal information, we use a method that requires users to agree to the terms and conditions. Notifying the user of the subject of personal information collection and purpose

of use, asking them to confirm that they have read the terms and conditions, and then asking them to agree or not to the terms and conditions and checking the checkbox.

2. Are there concerns around distributing models that have PII stored within it? That can expose such PII?

Executive Summary

If distributing models causes providing PII to third parties, a consent of the data subject is required in South Korea under Personal Information Protection Act in principle.

Rule

Personal Information Protection Act

Article 17 (Provision of Personal Information)

(1) A personal information controller may provide (or share; hereinafter the same shall apply) the personal information of a data subject to a third party in any of the following circumstances: <Amended by Act No 16930, February. 4, 2020>

1. Where the consent is obtained from the data subject;
2. Where the personal information is provided within the scope of purposes for which it is collected pursuant to Articles 15 (1) 2, 3 and 5 and 39-3 (2) 2 and 3.

(2) A personal information controller shall inform a data subject of the following matters when it obtains the consent under paragraph (1) 1. The same shall apply when any of the following is modified:

1. The recipient of personal information;
2. The purpose for which the recipient of personal information uses such information;
3. Particulars of personal information to be provided;
4. The period during which the recipient retains and uses personal information;
5. The fact that the data subject is entitled to deny consent, and disadvantages, if any, resulting from the denial of consent.

(3) A personal information controller shall inform a data subject of the matters provided for in paragraph (2), and obtain the consent from the data subject in order to provide personal information to a third party overseas; and shall not enter into a contract for the cross-border transfer of personal information in violation of this Act.

(4) A personal information controller may provide personal information without the consent of a data subject within the scope reasonably related to the purposes for which the personal information was initially collected, in accordance with the matters prescribed by Presidential Decree taking into consideration whether disadvantages are caused to the data subject, whether necessary measures to secure safety, such as encryption, have been taken, etc. .<Newly Inserted by Act No. 16930, 4. February, 2020 >

Analysis

Personal Information Protection Act of South Korea Article 17 requires consent of the data subject in principle whenever anyone provides personal information to a third party in principle.

Conclusion

If distributing models causes providing PII to third parties, a consent of the data subject is required in South Korea.

3. How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?

Executive Summary

As the possibility of collection and availability of personal information differs depending on whether or not consent is given to the collection of information. Therefore, consent of the individuals affects the possibility of using personal data.

Rule

Personal Information Protection Act

Article 15 (Collection and Use of Personal Information)
(1) A personal information controller may collect personal information in any of the following circumstances, and use it with the scope of the purpose of collection:
1. Where consent is obtained from a data subject;
2. Where special provisions exist in other laws or it is inevitable to observe legal obligations;
3. Where it is inevitable for a public institution's performance of its duties under its jurisdiction as prescribed by statutes, etc.;
4. Where it is inevitably necessary to execute and perform a contract with a data subject;
5. Where it is deemed manifestly necessary for the protection of life, bodily or property interests of the data subject or third party from imminent danger where the data subject or his or her legal representative is not in a position to express intention, or prior consent cannot be obtained owing to unknown addresses, etc.;
6. Where it is necessary to attain the justifiable interest of a personal information controller, which such interest is manifestly superior to the rights of the data subject. In such cases, processing shall be allowed only to the extent the processing is substantially related to the justifiable interest of the personal information controller and does not go beyond a reasonable scope.
(2) A personal information controller shall inform a data subject of the following matters when it obtains consent under paragraph (1) 1. The same shall apply when any of the following is modified.
1. The purpose of the collection and use of personal information;
2. Particulars of personal information to be collected;
3. The period for retaining and using personal information;

4. The fact that the data subject is entitled to deny consent, and disadvantages, if any, resulting from the denial of consent.
(3) A personal information controller may use personal information without the consent of a data subject within the scope reasonably related to the initial purpose of the collection as prescribed by Presidential Decree, in consideration whether disadvantages have been caused to the data subject and whether necessary measures have been taken to secure such as encryption, etc. < This Article Newly Inserted by Act No. 16930, 4. February, 2020 >

Analysis

The possibility of collection and availability of personal information differs depending on whether or not consent is given to the collection of information in principle under the Personal Information Protection Act of South Korea. Therefore, consent of the individuals affects the possibility of using personal data. When obtaining consent for the collection and use of personal information, the following matters must be accurately notified and consent must be obtained.

Conclusion

As the possibility of collection and availability of personal information differs depending on whether or not consent is given to the collection of information. Therefore, consent of the individuals affects the possibility of using personal data. The purpose of the collection and use of personal information; Particulars of personal information to be collected; The period for retaining and using personal information; The fact that the data subject is entitled to deny consent, and disadvantages, if any, resulting from the denial of consent.

4. Additional Question : Does the applicable law change depending on whether the content of the collected data is relevant to the individual?

Executive Summary

Yes it does. The applicable law will change depending on whether the content of the collected data is relevant to the individual.

Rule
Article 2 (Definitions)
The terms used in this Act shall be defined as follows: <Amended by Act No. 12504, Mar. 24, 2014; Act No. 16930, Feb. 4, 2020>
1. The term "personal information" means any of the following information relating to a living individual:
(a) Information that identifies a particular individual by his or her full name, resident registration number, image, etc.;
(b) Information which, even if it by itself does not identify a particular individual, may be easily combined with other information to identify a particular individual. In such cases, whether or not there is ease of combination shall be determined by reasonably considering the time, cost,

technology, etc. used to identify the individual such as likelihood that the other information can be procured;
(c) Information under items (a) or (b) above that is pseudonymized in accordance with subparagraph 1-2 below and thereby becomes incapable of identifying a particular individual without the use or combination of information for restoration to the original state (hereinafter referred to as “pseudonymized information”);
1-2. The term “pseudonymization” means a procedure to process personal information so that the information cannot identify a particular individual without additional information, by deleting in part, or replacing in whole or in part, such information;
2. The term “processing” means the collection, generation, connecting, interlocking, recording, storage, retention, value-added processing, editing, searching, output, correction, recovery, use, provision, disclosure, and destruction of personal information and other similar activities;
3. The term “data subject” means an individual who is identifiable through the information processed and is the subject of that information;
4. The term “personal information file” means a set or sets of personal information arranged or organized in a systematic manner based on a certain rule for easy search of the personal information;
5. The term “personal information controller” means a public institution, legal person, organization, individual, etc. that processes personal information directly or indirectly to operate the personal information files as part of its activities;
6. The term "public institution" means any of the following institutions:
(a) The administrative bodies of the National Assembly, the Courts, the Constitutional Court, and the National Election Commission; the central administrative agencies (including agencies under the Presidential Office and the Prime Minister’s Office) and their affiliated entities; and local governments;
(b) Other national agencies and public entities prescribed by Presidential Decree;
7. The term "visual data processing devices" means the devices prescribed by Presidential Decree, which are continuously installed at a certain place to take pictures of persons or images of things, or transmit such pictures or images via wired or wireless networks.
8. The term “scientific research” means research that applies scientific methods, such as technological development and demonstration, fundamental research, applied research and privately funded research.

Analysis

According to Korean law, the applicable laws differ depending on whether data contains personal information, de-identified personal information, or pseudonymous information. If the data contains personal or de-identified personal or pseudonymous information, the Personal Information Protection Act applies. However, the Personal Information Protection Act does not apply to data that is not related to an individual. The Personal Information Protection Act of South Korea will apply when the content of the collected data contains personal information. If the data is not relevant to the individual, the Act is not applicable. Article 2(Definitions) of the Act defines the term ‘personal information’ as follows: (a) Information that identifies a particular individual by his or her full name, resident registration number, image, etc.; (b) Information which, even if it by itself does not identify a particular individual, may be easily combined with other information to identify

a particular individual. In such cases, whether or not there is ease of combination shall be determined by reasonably considering the time, cost, technology, etc. used to identify the individual such as likelihood that the other information can be procured; (c) Information under items (a) or (b) above that is pseudonymized in accordance with subparagraph 1-2 below and thereby becomes incapable of identifying a particular individual without the use or combination of information for restoration to the original state (hereinafter referred to as “pseudonymized information”).

※ Related scholarly article : Sang-yook Cha, “Legal Issues on Protection of Trainable Datasets for Artificial Intelligence under Copyright Law - Focusing on The Copyright Exception for Text and Data Mining(TDM) -”, Journal of Business Administration & Law. Oct 31, 2021 32(1).

Conclusion

Therefore, the applicable law will be changed if the collected data contains personal information.

5. Additional Question : What are the mechanisms by which data mining has not caused a privacy infringement?

Executive Summary

Pseudonymization is one of the best options to prevent personal information infringement.

Rule

Personal Information Protection Act
Article 2 (Definitions)
The terms used in this Act shall be defined as follows: <Amended by Act No. 12504, Mar. 24, 2014; Act No. 16930, Feb. 4, 2020>
1-2. The term “pseudonymization” means a procedure to process personal information so that the information cannot identify a particular individual without additional information, by deleting in part, or replacing in whole or in part, such information;

Article 28-2 (Processing of Pseudonymous Data)
(1) A personal information controller may process pseudonymized information without the consent of data subjects for statistical purposes, scientific research purposes, and archiving purposes in the public interest, etc.
(2) A personal information controller shall not include information that may be used to identify a certain individual when providing pseudonymized information to a third party according to paragraph (1).
[This Article Newly Inserted by Act No. 16930, Feb. 4, 2020]

Analysis

Personal Information Protection Act of South Korea Article 28-2 allows process personal information without a consent of the data subject if pseudonymization has been carried out. Pseudonymization means a procedure to process personal information so that the information cannot identify a particular individual without additional information by deleting in part or replacing in whole or in part such information. Therefore, pseudonized information does not make a problem of privacy infringement. As a result, pseudonymization is one of the best options to prevent personal information infringement in South Korea.

Conclusion

Pseudonymization is one of the best options to prevent personal information infringement.

E. Prohibited content

1. What types of data may be prohibited from being stored or distributed?

Executive Summary

Circulation of data which contain obscene content, defame other persons, harmful to youths under the Youth Protection Act, and speculative activities etc. may be prohibited. Also, distributing, or providing child or youth sexual exploitation materials for commercial purposes, or possessing, transporting child or youth sexual exploitation materials is prohibited.

Rule

Act On Promotion Of Information And Communications Network Utilization And Information Protection

Article 44-7 (Prohibition on Circulation of Unlawful Information)
(1) No one may circulate any of the following information through an information and communications network: <Amended on Sep. 15, 2011; Mar. 22, 2016; Jun. 12, 2018>
1. Information with obscene content distributed, sold, rented, or displayed openly in the form of code, words, sound, images, or motion picture;
2. Information with content that defames other persons by divulging a fact or false information, openly and with intent to disparage the person's reputation;
3. Information with content that arouses fear or apprehension by reaching other persons repeatedly in the form of code, words, sound, image, or motion picture;
4. Information with content that compromises, destroys, alters, or forges an information and communications system, data, a program, or similar or that interferes with the operation of such system, data, program, or similar without good cause;
5. Information with content that amounts to a media product harmful to youths under the Youth Protection Act and that is provided for profit without fulfilling the duties and obligations

under the relevant statutes and regulations, including the duty to verify the subject's age and the duty of labeling;
6. Information with content that amounts to speculative activities prohibited by statutes and regulations;
6-2. Information with content of transactions of personal information in violation of this Act or any other statute or regulation regarding the protection of personal information;
6-3. Information regarding methods, drawings, etc. for manufacturing guns or explosives (including things with a yield that may expose people to risk of life or bodily injury);
7. Information with content that divulges a secret classified under statutes and regulations or any other State secret;
8. Information with content that violates the National Security Act;
9. Other information with content that attempts to commit, aids, or abets a crime.

Act On The Protection Of Children And Youth Against Sex Offenses

Article 11 (Production or Distribution of Child or Youth Sexual Exploitation Materials)
(1) Any person who produces, imports, or exports child or youth sexual exploitation materials shall be punished by imprisonment with labor for an indefinite term or for a limited term of at least five years. <Amended on Jun. 2, 2020>
(2) Any person who sells, lends, distributes, or provides child or youth sexual exploitation materials for commercial purposes, or possesses, transports, advertises or introduce them for any of such purposes, or publicly exhibits or displays them shall be punished by imprisonment with labor for not more than five years. <Amended on Jun. 2, 2020>
(3) Any person who distributes or provides child or youth sexual exploitation materials, advertises or introduces them for any of such purposes, or publicly exhibits or displays them shall be punished by imprisonment with labor for not more than three years. <Amended on Jun. 2, 2020>
(4) Any person who procures a child or youth for a child or youth sexual exploitation materials producer, knowing that he or she is to be used for producing child or youth child or youth sexual exploitation materials, shall be punished by imprisonment with labor for at least three years. <Amended on Jun. 2, 2020>
(5) Any person who purchases child or youth sexual exploitation materials or possesses or views them with the knowledge that it is a child or youth sexual exploitation materials, shall be punished by imprisonment with labor for at least one year. <Amended on Jun. 2, 2020>
(6) Any person who attempts to commit an offense prescribed in paragraph (1) shall be punished.
(7) Any person who habitually commits offenses referred to in paragraph (1) shall be subject to an aggravated punishment by up to 1/2 of the penalty stipulated for such offense. <Newly Inserted on Jun. 2, 2020>
[Title of This Article Amended on Jul. 2, 2020]

Analysis

Act On Promotion Of Information And Communications Network Utilization And Information Protection Article 44-7 prohibit circulating any of the information with obscene content, defames other people, harmful to youths under the Youth Protection Act, and speculative activities etc. In the meanwhile, Act On The Protection Of Children And Youth Against Sex Offenses Article 11 prohibit distributing, or providing child or youth sexual exploitation materials for commercial purposes, or possessing, transporting them for any of such purposes etc.

Conclusion

Prohibited data type which contain obscene content, defame other persons, harmful to youths under the Youth Protection Act, and speculative activities etc. may be prohibited. Also, distributing, or providing child or youth sexual exploitation materials for commercial purposes, or possessing, transporting child or youth sexual exploitation materials is prohibited.

2. Is there a legal restriction on the distribution of data for national security reasons?

Executive Summary

Distributing data containing military secrets is illegal in South Korea.

Rule

Military Secret Protection Act

Article 12 (Leakage)
(1) If any person who has detected or collected military secrets leaks them to others, he/she shall be punished by imprisonment with labor for a limited term of not less than one year.
(2) If any person who has come to know or possess military secrets by chance leaks them to others despite knowledge that they are the military secrets, he/she shall be punished by imprisonment with labor for not more than five years or by a fine not exceeding 50 million won. <Amended by Act No. 12556, May 9, 2014>
[This Article Wholly Amended by Act No. 10792, Jun. 9, 2011]

Analysis

Military Secret Protection Act of South Korea Article 12 prohibits leaking military secrets. If a person who detects or collects military secrets divulges them to another person, he or she may be punished by imprisonment with labor for not less than one year, and if a person who accidentally becomes aware of military secrets leaks them to another person even though he knows they are military secrets, he shall be punished by imprisonment for not more than five years, or A fine of not more than 50 million won may be imposed. Therefore, distributing data containing military secrets is illegal and could be convicted in South Korea.

Conclusion

Distributing data containing military secrets is illegal and could be convicted guilty in South Korea.

JAPAN

Maiko Takeuchi and Sinee Sang-aroonsiri

Introduction

Legal system

The Japanese legal system is based on the civil law system. Although under the Constitution, bills can be submitted by both lawmakers (National Diet [Congress] members) and the Cabinet, approximately 80% of the legislatures are drafted by the government agencies and submitted by the Cabinet.

Recent Development of the field relevant to this projects (e.g. data protection, personal information protection and intangible technology transfer protection)

Recently, relevant laws and regulations related to the IP, PII, and technology transfer have been developed. For example, the Copyright Act (revised in 2018, 2019 and 2020), Act on the Protection of Personal Information (revised in 2020) and Foreign Exchange and Foreign Trade Control Law (revised in 2019 and 2021) are revised to address the newly emerged data protection/use-related issues.

A. IP Questions

1. **Are the data training sets and models protected by IP rights and if so which IP rights?**

Data/ Databases

There is no definition of data ownership in Japan's legal (hard law) regime. Data is not the subject of rights under the Civil Code such as ownership or possession, usufruct, or security interest. There is also no concept similar to EU's sui generis or database rights in Japan.

Sometimes data and/or databases might be protected under IPRs, i.e. **copyrights, patents, or trade secrets** if it meets the requirement of each concept. However, it is rather unlikely for data and databases to meet the legal requirements of copyrights and patents. Lastly, legal precedents in Tokyo District Court and IP High Court demonstrate that, despite not being protected under IPRs, the database owner is likely to sue a competitor who exploits the database without permission for tortious liability.

Copyright

Under the Copyright Act, each data or database may be protected if it is "creatively expressed" and falls within the literary, academic, artistic, or musical domain (Article 2(1)(i)). Particularly, the database whose information selection or its structural composition is creative could be protected as copyrighted work (Article 12-2). Nevertheless, since the data and the database are usually mechanically generated, it might be challenging to find them "creatively expressed."[27]

[27] Ministry of Economy, Trade and Industry, "Contract Guidelines on Utilization of AI and Data, Data Section" https://www.meti.go.jp/press/2019/04/20190404001/20190404001-1.pdf pp.9

More importantly, the Copyright Act's amendment in 2018 has expanded the scope of free use of copyrighted work. Specifically, Article 30-4 permits the use of a copyrighted work in data analysis, text mining, and machine learning, to the extent considered necessary, provided that: (i) such copyrighted work is used without any purpose of having the expression of the work perceived by any person, and (ii) the use does not unreasonably injure the copyright owner's interest in light of the kind and nature the work, and the circumstances of such work's exploitation. Consequently, similar to the fair use doctrine in the US, using the dataset for computational purposes (without disclosing it as-is) should not constitute copyright infringement, unless the use injures the interest of copyright holders.

In addition to the rule in Article 30-4 of the Copyright Act, Article 47-4 and Article 47-5 explicitly clarifying that the reproduction of copyrighted works for the machine learning process does not violate the Copyright Act. Specifically, Article 47-4 permits electronic incidental copies of works, recognizing that this process is necessary to carry out machine learning activities but does not harm copyright owners. Article 47-5 allows the use of copyrighted works for data verification when conducting research, recognizing that such use is important to researchers and is not detrimental to rights holders. As a result, the article enables searchable databases, which are necessary to carry out data verification of the results and insights obtained through text data mining.

Patent

In terms of patent right, even though the method employed to produce, process, and/or analyze data might be novel and inventive (See Article 2(1), Article 29(1), and Article 66(1) of the Patent Act), to find that the data and/or database alone is inventive and/or novel might be difficult.

Trade Secrets

Therefore, under the Japanese IPRs regime, trade secrets might have the highest possibility to give protection to data and/or databases. Information will be considered a trade secret if it is (i) managed and as a secret, (ii) having utility, and (iii) is not a public domain (Unfair Competition Prevention Act (UCPA) - 不正競争防止法 Article1(6)). Nevertheless, if datasets are created by scraping publicly available information on websites, it might be difficult to claim that the datasets itself are trade secrets. However, recently METI has been trying to make trade secrets more enforceable, such as amending UCPA to implement more effective civil remedies[28].

The term "data ownership"

Nevertheless, the term "data ownership" is sometimes used in data contracts. In this case, data ownership usually refers to "the de facto position of being able to access and control data, or a contractual status in cases where an undertaking has been entered into by contract regarding the authority to use data."[29]

Tortious Liability: an alternative for database owner to have his database protected

Even though database might not be protected by copyright, Tokyo District Court and IP High Court have found in several cases that the unauthorized use or duplication of databases may trigger

[28] See e.g. https://blogs.orrick.com/trade-secrets-watch/2016/04/18/were-not-gonna-take-it-significant-changes-to-japans-trade-secret-protection-law/

[29] Supra note 1

tortious liability (Civil Code Article 709, see e.g., Tsubasa System case (Tokyo District Court, Judgment, May 25, 2001), Office Caster case (Tokyo District Court, Judgment, February 2, 2002), and Yomiuri Online Headline case (IP High Court, Judgment, October 6, 2005)). The court usually rules so if: (i) the database producer invests a considerable amount of effort and expense to build such a database, and (ii) the defendant is a competitor using the database against the producer's interest. If both prongs are fulfilled, the defendant's action will be considered a violation of the producer's "legally protected interest" (法律上保護される利益), since it interferes the principle of fair and free competition (公正かつ自由な競争原理).

One leading case on the tortious liability arising from the misuse of database by a competitor is Tsubasa System Case （翼システム事件）which is as follows. This case is an excellent case portraying the difficulty of a database to be considered a copyrighted work under Japanese Copyright Law and an alternative legal strategy to have a database protected under tort law.

Fact:
The plaintiff (Super Frontman Co., Ltd., hereinafter "S") and the defendant (Tom Cat Co., Ltd., hereinafter "T") were competitors in automobile maintenance service providers' support systems. In other words, they both produced the systems that helped automobile maintenance service providers create transaction documents such as quotations, orders, invoices, and delivery statements and manage customer relationships. To help users input information more quickly, both companies built their own four-wheeled automobile databases that included each car model's information, e.g. model number, specification, and manufacturer's name.

Here, S sued T asserting that T had copied their database and filed for injunction relief. S reasoned that T infringed either its copyright over the databases under the Copyright Act, or its "legally protected interest" (法律上保護される利益) triggering the tortious liability under Article 709 of Civil Code. The fact is settled that the T copied S's database.

Rules:

Copyright Act

Article 2 (1)
(x)-3 "database" means an aggregate of data such as articles, numerical values, or diagrams, which is systematically constructed so that such data can be searched with a computer

Article 12-2
(1) A database that, by reason of **the selection or systematic construction** of information contained therein, constitutes a creation is protected as a copyrighted work.

(2) The provisions of the preceding paragraph do not affect the rights of the author of a copyrighted work that forms part of a database as referred to in that paragraph

Civil Code
Article 709
A person who has intentionally or negligently infringed any right of others, or legally protected interest of others, shall be liable to compensate any damages resulting in consequence.

Issue:

(1) Is S' database considered a copyrighted work which would cause T's action ?
(2) Does the T's action of copying the database result in tortious liability?

Holding:

(1) No, (2), Yes

Reasoning:

Issue 1: this database is not protected under copyright
S's database was considered a database under Article 2(1)(x)-3's definition, which could be copyrightable under Article 2 (1) (10) of the Copyright Act. However, the court found that this database lacks originality in selecting the information to be input into the database, pursuant to the rule in Article 12-2. Specifically, selecting which automobiles' to be input into the database was not original, as S just input all of the cars which were proving that they existed. Similarly, the method how S selected types of information relating to each car in the database were not also original since S just included the data required for producing the car inspection certificate (自動車検査証). This certificate is a document required to be issued to automobile owners or users under the Road Vehicles Act. Other competitors also possessed databases with similar types of information. As a result, this database is a copyrighted work under the Copyright Act and therefore T did not infringe S's copyright.

In addition to the selection, the court did not find how S constructed this database had originality either. Specifically, S only arranged the database here from the old automobile model to the newer one. Consequently, the court did not find there is originality in this aspect.

Issue 2: T's action constituted tortious liability
Even though S was not protected under the Copyright Act, S had a legally protected interest in the database and could make T tortiously liable for the unauthorized reproduction of the database. "**One might spend considerable expense and time to collect the information, build the database, and resell such a database as its business.** As a result, when the others who copied parts of such a database to build their own and resold them in the same territory competing against S, such actions could be considered unfair practice which infringed the original database producer, under the society where the fair and free competition principle is protected."

In this case, S spent development fees on the database for more than 500 million yen and maintenance fees for more than 40 million yen. As a result, the court concluded that T's action of

copying the database seriously deviated from the scope allowed under the fair and free competition doctrine and violated the legally protected interest of S. All in all, S should be liable for T because of this action.

Models

As models are compositions in programming languages, they will highly possibly be protected by the Copyright Act. Furthermore, when models themselves or the concept, process, or methods of how models are made are inventive and novel, the developer might also have their patents registered in Japan. Lastly, if the models have utility, are not a public domain, and have been protected as secrets, models' owners may protect the models as trade secrets under UCPA as well.

2. **Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?**

If the pre-trained language model in this question means the model which the training data could be seen as-is, it might results in copyright infringement if the data is creatively expressed (Article 30-4 of Copyright Act)

3. **Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g. HTML tags or document structure)?**

The answer might vary depending on the nature of such additional information. As a result, this must be considered on a case-by-case basis. In other words, in addition to the data/databases' legal issues as explained above, one must consider if additional information is protected under IPR or other legal regimes.

For example, regarding the HTML tag, Article 10 (3) of the Copyright Act explicitly states that programming languages, coding conventions, or algorithms are not copyrightable work. Simply mentioning <html>, <p>, and <head> separately, therefore, should not constitute a creatively expressed work and is less likely to result in copyright infringement. However, suppose the tags are selected discretely and accompanied by the other trained model's explanation. In that case, the whole explanation might be a copyrightable work, and the explanation's producer must be careful not to infringe the copyright of the other works.

4. **What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?**

Similar to data, metadata might possibly fall under protection of copyright, patent, and/or trade secret if its characteristics fall under the definition of each category. However, as explained above, the possibility of metadata being governed by copyright and patent is limited. If a metadata's developer wants to control his metadata's utilization by a third party, he may protect the metadata in a manner that makes it a trade secret. However, this might not be an option when the metadata is publicly available information. Therefore, due to the lack of default rules, the developer should

pay attention to the terms and conditions in the metadata's licensing agreement so that the developer could maintain control over metadata (e.g. Is the metadata sub-licensable? If so, does the licensee need to acquire permission prior to the sublicense? How long can the licensee use the metadata? If there are derived works of the metadata, who would own it?).

5. **Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR/Pile) (for example, such a dataset may contain information like: in the nth entry of C4, there is a "<b>" html tag after the mth character)?**

Intentionally Omitted

6. **What are the IP risks related to data collection directly from persons? For example when you interview people or they donate data etc.**

In the IPR regime, collecting data directly from a person should not constitute any IP risks, as such data is a fact. However, in terms of data privacy, a collector should notify a data subject on the purpose of collecting data and collect consent from the data subject.

7. **Can NLP researchers train a language model on borrowed library books - e.g., from the Internet Archive or other online book lender?**

Here, Article 30-4 of Copyright Law applies again. The datafication of information, creation of dataset, and the training of language model are allowed if the purpose of these actions are not to show the sentiments of the expression of the original works, and the actions do not unreasonably injure the copyright owner's interest in light of the kind and nature the work, and the circumstances of such work's exploitation.

8. **How do rights on the source data (e.g. copyright) transfer to the trained model?**
a. **Will this depend on where training occurs?**
b. **Will this depend on where data is gathered?**

Intentionally Omitted

B. Licensing Questions

1. **Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?**

Data/Datasets/Metadata

Even though no law directly governs data itself, unlike sui generis in the EU, data is licensable by contracts. However, please also be careful that if the dataset falls under the definition of copyright and/or patent, Copyright Act and/or Patent Act shall be the default rule regarding the ownership.

As explained above, since IPR law does not always govern mostly data, datasets, and metadata, most data does not automatically have licensing default rules. Therefore, it is crucial to clarify the terms of use and licensing conditions in an agreement. In case that the data, datasets, or metadata is also trade secrets, one should also set the terms and conditions which disclose such data to be protective enough to protect the confidentiality of the data.

Models

As models mostly are works of programming language, the models should be licensable under the default rule in the Copyright Act. Similar to data, the model may be occasionally inventive enough to have a patent registered or can be protected as trade secrets. In that case, one should also pay attention to the licensing conditions in the Patent Act or the measures to keep the model confidential so that it could be a trade secret under the Unfair Competition Prevention Act.

2. What are the relationships between laws, licenses, and terms of use? Which is more binding?

Assume that the terms of use here mean the terms of use of the website where a user could download datasets, and the datasets also have their own license.

Article 521 of Civil Code (民法), amended in 2017, codified the well-accepted freedom to contract doctrine (契約自由の原則) as follows:

> "(1) Except when there is a specific provision of the law, anyone can freely decide whether to enter an agreement
> (2) A party can freely decide the contents of the agreement within the restriction of the law."

Therefore, if not being restricted by the law which is "mandatory law" (jus cogen or 強行法規), generally parties to an agreement could negotiate and decide the contents of the contracts freely. Regarding the question of whether websites' terms of use or licenses will prevail, usually, the licenses attaching to the datasets shall prevail since they have more specific rules over such datasets.

However, sometimes the terms of uses and/or the licenses may explicitly stipulate that when there is a discrepancy between the term of use and the license, the terms of uses shall prevail. This usually happens when dataset owners desire to set the terms of use of the datasets' website to be a master agreement. With such provision, the terms of use might prevail. Consequently, to avoid liability datasets users should review the contents of licenses and terms of use before using the datasets.

More importantly, the validity of the terms of use and the licenses is how a dataset's user *accepts* them. I (a writer) assume that mostly when downloading datasets and/or training models, researchers might need to register themselves to the website/platform so that they could download. To be a member, therefore, registrants need to *expressly* accept the T/U. Plus, when downloading datasets or training models, the researchers might also be required to click "accept" that they will understand and comply with the licensing terms. In this case, all the T/U and license

will apply to the downloading researchers. In contrast, if the website does not explicitly require the researcher to accept the license and/or T/U, does not explicitly show the licenses and/or the T/U's contents for review, the downloader may have room to argue that he does not abide by the T/U and/or license. This will relate to the issue of the terms of use, even though not being The interpretation here could be tricky and depends on the T/U and license's wording and the manner in which T/U and licenses are represented. Consequently, the T/U and license might need to be reviewed on a case-by-case basis.

3. **What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?**

NLP researchers should review the scope of the license, which is usually explicitly provided in the license. In case of ambiguity, NLP researchers may inquire the dataset owners to confirm if the license applies to the dataset they wish to use.

4. **What about if the users download or copy their own data and then provide it to NLP researchers directly?**

Intentionally Omitted

5. **Does the license that the dataset is shared under override the terms and conditions?**

The answer here is provided in the answer to Question 2.

6. **Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?**

Question interpretation

Assume that NLP researchers here have legitimate access to the datasets which might be subject to copyright protection. For example, the datasets are publicly available on GitHub, so the researchers do not need to hack to obtain data. Even though one might be able to access datasets freely, the datasets' owner might set up licensing conditions and the dataset's terms of use.

This question will not include when researchers build up the dataset themselves from web-crawling

Copyright issue

As explained in Section A's Question 1, there might be little chance that datasets are protected under Japan's Copyright Act. However, for the sake of argument, let's suppose that the datasets here are copyrighted works.

Pursuant to Article 30-4 of the Copyright Act, as explained in Question 1 of Section A, NLP researchers may use the datasets for machine learning even though datasets' owner does not license such data to researchers, provided that (a) the researchers must use datasets without any

purpose of having the expression of the datasets perceived by any person, and (ii) the use does not unreasonably injure the dataset' copyright owner's interest in light of the kind and nature the work, and the circumstances of such work's exploitation. In such a case, if the model does not include the datasets with their as-is expression, a model developer may distribute the model freely. However, datasets should not be distributed per se.

7. **Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?**

Intentionally Omitted

C. Text Data Mining and Fair Use Questions

1. **What are types of legally permitted text data mining?**

Rules

In Japan, there is no regulation to prohibit text data mining per se. Text data could be protected under the Copyright Act. With certain exceptions, it is legally permitted to reproduce copyrighted works for personal or family use. Academic purpose is also permitted under the Copyright Act（著作権法）.

Furthermore, if a person produced a set of data using information extracted from the text data, distribution of the set of data could be a violation of the Act on the Protection of Personal Information（Privacy Act（個人情報保護法)).

The discussion hereafter is based on the precondition that the relevant actions are not violation of laws other than above-mentioned laws. e.g.the original text data was obtained with intrusion of databases（violation of Article 3 of Act on Prohibition of Unauthorized Computer Access (不正アクセス行為の禁止等に関する法律)) and the data scraping imposed the load and impeded the server access. (violation of Articles 233[Forcible Obstruction of Business] and 234[Obstruction of Business by Damaging a Computer] of the Penal Code （刑法）)

Application

Person/legal person 1) conducted text data mining and 2) then distributed it.

1) Text Data Mining
   Text data mining, without violating the relevant laws, is not restricted in Japan. The use of the copyrighted work which aims not to have its expression perceived and does not conflict with the interest of the copyright owner is allowed under the Copyright Act. Please refer to the answer in Section A's Question 1.

2) Distribution of the data

Data could be obtained through a) SNS information b) contents published and not protected under Copyright Act c) but protected under the Copyright Act.

Information categorized as c) publicly available information not protected under the Copyright Act could be a violation of the Privacy Act.

There are two relevant cases of publication of information of personal information of bankrupts. Name and address of bankrupts are published in the government gazette. The contents are openly available and not protected under the Copyright Act. In 2019, a group published a "bankrupt map," map with the address and names of the bankrupt. Personal Information Protection Commission（PPC, 個人情報保護法委員会）issued administrative guidance (行政指導 [gyosei-shido], a non-binding advice govern by an administrative agency) stating that publicagion of such contents could be violation of §18 (Notification of a Purpose of Use when Acquiring) and §23 (Restriction on Third Party Provision) and adequate action to be taken, and the organization followed the advice. After the closure, however, multiple entities published similar information on the internet. In July 2020, PPC ordered closure of the sites to two groups ordered pursuant to the Privacy Act. This is the first case of closure order issued by PPC.

This case the PPC weighed the privacy of the individuals to the availability of the information. Therefore it could be probable that publication or distribution of the similar sensitive personal information is considered to be a violation of the Privacy Act, and subject to the order to suspend the publication.

Conclusion
There is no regulation to prohibit text data mining per se. However, when the information could be protected by the relevant laws such as the Copyright Act and Privacy Act, then distribution of the information could be illegal.

1.1 Should it be limited to academic and non-commercial use?

Currently Japanese public opinion and the government policy is toward balancing between the protection of the contents and use of the data obtained through text data mining. Therefore, it is unlikely that the Japanese government will limit the use to academic and non-commercial rule. Plus, with the Article 30-4, 47-4, 47-5 of the Copyright Act altogether allow the use of copyrighted works for machine learning with no restriction to the academic or non-commercial use.

2. Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question?

To our knowledge, there is no particular restriction on the entity or individual that could crawl data on the internet in Japan.There is no requirement to register or to obtain it in order to do web crawling.

3. **If NLP researchers can crawl data ourselves, what are the limits regarding using this data for training a language model? Can NLP researchers redistribute the data afterwards? If so, under which license and in which geographical regions?**

The answer to this question is already provided in question 6 of Section B

4. **If NLP researchers use parts of the Pile, Common Crawl, OSCAR or C4, can NLP researchers redistribute the data later? If yes under which conditions and in which countries?**

Intentionally Omitted

5. **What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify? The edition conversations of Wikipedia?**

Social media content might involve personal data which will be subject to protection under Japan's Privacy Act.

Article 18 of Privacy Act requires one who uses a personal information database to notify or publicly announce the purpose of utilization when he acquires personal information. More importantly, Article 18 applies to any kind of personal information's acquisition, including collecting publicly available personal information. Similar to Bankrupt Map case in Question 1, where the map's drafter only collected personal information from the publicly available Gazette, collecting personal information from the publicly available web pages of social networks (e.g. LinkedIn's profile, some of which could be viewed without logging in) without informing data subject of the purpose could violate Article 18 of Privacy Act.

Furthermore, a data collector must not use personal data against the interest of the data subject. The famous case of Recruit Career, a recruiting platform, using the personal information of freshly graduated job seekers without their consent to analyze the potential of the job seekers declining job offers, is an excellent example.[30]

6. **For example, do the terms and conditions of Twitter, Facebook, Youtube, etc., tell us whether NLP researchers can collect data from them for a project such as BigScience?**

The question here may be rephrased to if the terms and condition of SNS bind NLP researchers. The change in Civil Law mig

This questions relates to the concept of "standard terms and conditions" (Teikei-Yakkan 定型約款, hereinafter "Standard T/C") which was newly introduced in 2018's Civil Code amendment. Standard T/C means "a collection of provisions prepared by a person with the purpose of applying them as the terms of a contract for a standard transaction," and apparently SNS's T/C is considered Standard T/C. According to Article 548-2 (1), to have a web T/C a valid agreement between a web provider and an user, the web provider must either (i) have **the user agree to**

[30] For further information please see https://mainichi.jp/english/articles/20190827/p2a/00m/0na/009000c

adopt the web T/C as the web service's terms and conditions , **or** (ii) **manifests to the user** that the web provider will use such web T/C as the web services's terms and conditions between them in advance.

Nevertheless, Article 548-2 (2) provides an exception of 548-2 (1). Specifically, a person will not be deemed accepting a provision in Standard T/C if (i) that provision restricts the rights or expand his duties, and (ii) that are found, in light of the manner and circumstances of the standard transaction, as well as the common sense in the transaction, to unilaterally prejudice his interests in violation of the good faith principle in the Civil Code's Article 1(2).

As a result, an important question here is if social networks' T/C explicitly barring web scraping and AI development based on the social networks' contents could override the Copyright Act's Article 30-4, 47-4, and 47-5. Similar to the US, if a web crawler crawls while logging in a particular social network, his action will potentially be deemed a breach of the T/C. Since a web crawler, expressly accepted the T/C of the social network to be a member and log in, such T/C is binding pursuant to Article 548-2 (1), and his action breaches the T/C's provision.

On the other hand, a trickier problem here is when a web crawler crawls without expressly accepting the social media's T/C or logging in such social media. Would the web T/C bind the web crawler since the T/C is shown at the bottom of the website, according to Article 548-2? To our knowledge there is no court precedent on this issue now. There are two possible interpretations as follows:

Firstly, the web T/C should be binding if the T/C is shown at the place which is easy to be discovered. Putting the link of the T/C at the bottom of the page already brings enough unity and should be considered that the T/C is properly "manifested" to users under Article 548-2 (1)(ii)[31].

Secondly, prohibiting web scraping here limits the fair use rights under Copyright Law of the users, so this provision must be expressly accepted by the users, otherwise it will not bind the users. Because of Article 548-2 (2), prohibiting web crawling for AI development purposes will be binding only when the web crawler expressly accepts the T/C. Since such prohibition limits the right to fair use of all users, and might unilaterally prejudice web users' interest, showing the T/C at the bottom of social network web pages should not suffice. There are articles which support this interpretation.[32]

The Ministry of Economy, Trade, and Industry (METI) has mentioned this issue in its Rule relating to Electronic Commercial Transaction and Information Goods Transactions[33] but its stance was

[31] See Hideki Muramatsu and Matsuo Hironori, *Teikei Yakkan no Jitsumu Q&A (定型約款の実務Q&A)*, Shojihomu, 2018, 72-73.

[32] See Sugiura Kenji, *Minpou Kaisei Kara 1 Nen, Web Saabisu no Riyoukiyakuj Jitsumu no Ima to Saikakunin no Point (Zenhan) (民法改正から1年、WEBサービスの利用規約実務のいまと再確認のポイント（前編）)*, BUSINESS LAWYERS , https://www.businesslawyers.jp/articles/940 (last visited Jan 8, 2022), and Sugiura Kenji, *Chosakuken Hou no Juunanna Kenri Seigen Kitei to Oobaaraido Mondai (著作権法の柔軟な権利制限規定とオーバーライド問題)*, STORIA法律事務所 (2021), https://storialaw.jp/blog/7658 (last visited Jan 8, 2022).

[33] Ministry of Economy, Trade, and Industry, *Denshi Shou Torihiki Oyobi Jouhouzai Torihiki Tou ni Kansuru Junsoku (電子商取引及び情報財取引等に関する準則)* (2019).

neutral. Specifically, it states that in principle, the provisions relating to fair use and copyright restriction in the Copyright Act (Article 30 to Article49) could be changed by an agreement of contracting parties, but there is also an interpretation asserting that the agreement which limits the scope of fair use is void.[34]

7. **What are some of the risks raised by collecting data from these social media directly?**

There might be a risk of collecting personal information, triggering an obligation to give notice under Article 18 of Privacy Act. Please refer to question 1 for further details.

8. **What changes if NLP researchers get direct consent from the users concerned?**

If NLP researchers collect personal information of the social media users through crawling and then get consent from the users concerned, they will have less chance of violating Article 18 of the Privacy Act. Please refer to question 1 for further details.

9. **Does the consent override the Terms of Use?**

Intentionally Omitted

10. **Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can't be scrapped?**

Intentionally Omitted

D. Privacy Questions

1. **Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?**

    i. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of common crawl, a published dataset based on a craw of the web in terms of PII?

Intentionally Omitted

2. **Are there concerns around distributing models that have PII stored within it? That can expose such PII?**

[34] Supra note, 252.

Under Act on the Protection of Personal Information（Privacy Act, 個人情報保護法), PII must be removed or replaced with other information from the database so that the individual will not be specified. (3-2-1, Guideline for Privacy Act) .

Individuals and organizations providing a personal information database are categorized as “personal information handling business operator” (個人情報取扱事業者）under Privacy Act (Article 2-5).

3. **What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exception for research purposes?**

Information is protected by both Privacy Act (see Question and Answer 2) and Privacy Act Guidelines concerning information in the specific fields, including medical record and genetic information. NLP researchers handling personal information must delete PII or replace particular PII data with other data. There is no particular exception for research purpose that is determined in the Privacy Act.

4. **How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?**

There is no regulation controlling NLP specifically. The information collected by NLP organizations would be protected by the laws applicable to general information. Generally, with informed consent of the individuals, the actor would be exempted from liability.

Furthermore, if the information fall under the category of “Special care-required personal Information” (hereinafter “sensitive personal information”要配慮個人情報) (note: the English translation is copied from government-operated website. This type of information is personal information which could be particularly cause discrimination, such as religion, race, birthplace (not just nationality but from certain discriminated tribe/area) , history of particular sickness, medical record, criminal record etc). The sensitive personal information is under stricter control than other personal information. Sensitive information must be obtained after the informed consent and cannot be the subject to opt-out exemption (distribution to third parties without consent of the provider of the information) .

5. **What are the privacy risks related to data collection directly from persons? For example when you interview people or they donate data etc.**

Intentionally Omitted

E. Prohibited content

1. **What types of data may be prohibited from being text data mined?**

Rule
There is no registration to prohibit text data mining per se. The Following are “caveats” concerning the relevant actions.

Liability of the individual who collected data used for other crime
Distribution could be in violation of the relevant laws such as Privacy Act and Copyright Act. Furthermore, if the data was used for the legally prohibited actions (e.g. data mining to collect information to build a prohibited weapon, and then the weapons are actually built, or the blueprint was distributed to third person) then the individual who conducted text data mining or the individual who distributed the blueprint could be liable for aiding and abetting the production of weapon without license (violation of Article 3 and 4 of Ordnance Manufacturing Act) or possession of weapon (violation of Article 3 of Act for Controlling the Possession of Firearms or Swords and Other Such Weapons)

Prohibition of possession of images
If the text data in case the “images” are collected in data mining, then possession of child prnography (only imagery of the actual body of the child, not computer graphics) with “for the purpose of satisfying one's sexual curiosity (limited to those who have come to possess it voluntarily, and are clearly deemed to as such.). (Article 7,Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children). Just accidental collection of prohibited imagery is not likely to be the violation of the Act. To establish “[T]he purpose of satisfying one's sexual curiosity intention to,“ the authority will review several elements such as amount the individual collects and the method the individual used.

2. What types of data may be prohibited from being 1) generated or 2) stored or 3) distributed?

1) There is no restriction on the generation of data per se, unless it is a) generated with an illegal means (see C-1) or b) child pornography (Violation of Article Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children (児童買春、児童ポルノに係る行為等の規制及び処罰並びに児童の保護等に関する法律)).
2) There is no restriction on storing data per se, unless the above-mentioned two situation (collected through illegal means, or child pornography).
3) Distribution of the information could be a violation of multiple laws including the following:

a) Privacy
   - Applicable Rules
     Act on the Protection of Personal Information(個人情報保護法）
b) Information used for WMDs, missiles and conventional weapon development and production
   - Applicable Rules
     Foreign Exchange and Foreign Trade Act（外国為替及び外国貿易法）
c) Trade Secret or licensed information
   - Applicable Rules

Copyright Act （著作権法）
Unfair Competition Prevention Act（不正競争防止法）

d) National Security Related Information
   - Applicable rules
     Act on the Protection of Specially Designated Secrets (特定秘密の保護に関する法律)
   - Protected by Act on Protection of Secrets Incidental to the "Mutual Defense Assistance Agreement Between Japan and the United States of America" (日米相互防衛援助協定等に伴う秘密保護法)
e) Damage individual's rights/business
   - Penal Code (刑法)
f) Porn
   - Applicable rules
     Article 175 (Distribution of Obscene Objects) of Penal Code（刑法）
     Violation of Article Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children（児童買春、児童ポルノに係る行為等の規制及び処罰並びに児童の保護等に関する法律)

3. What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?

Legal or cultural norms implicated by prohibited content

Japanese legal and social systems and culture are male-dominant. Japan's Global Gender Gap Index rank is 120th out of 156. *See* World Economic Forum, *Global Gender Gap Report 2021*(2021). Laws and policies related to women and children are slow to be treated. A well known story is that viagra was approved in Japan within half a year after being sold in the U.S., whereas low-dose pills were not approved until 1999, even after viagra. Japanese society and laws are torelant to sex industry in general (you will find red light district or "massage parlors" with lascivious signboard in the mainstreets where children can see), but the most relevant issue on this playbook is the torelance to pedphilia. In Japan, possession of child porn was not prohibited until 2015, pursuant to the revision of Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children（hereinafter "Act". 児童買春、児童ポルノに係る行為等の規制及び処罰並びに児童の保護等に関する法律）in 2014. Furthermore, there are two serious loopholes in this Act: 1. The Act only prohibits the image, videos and data of actual minors, not computer graphics or illustration. 2. The restriction of online access to child pornography is not legally binding, but controlled under "guidelines" by the internet providers.

With computer graphics and virtual reality devices, "almost real" lascivious content is created and distributed in Japan. These are legally allowed activities. With internet access, such content can

be easily accessed abroad, too. Such slow and passive attitudes to harmful contents could be explained based on this male-dominant decision making process.

Harms prevented and possible adverse impact of prohibition

This Act prevent child abuse and criminalized relevant activities. Openly discussed objection to this Act, especially to the 2014 revision to prohibit posession of child pornography is based on the concern of the freedom of expression. Japan has history of authority's suppression to the freedom of speech pursuant to Peace Preservation Law (治安維持法) (adopted in conjunction with the adoption of Universal Manhood Suffrage Law in 1925 and abolished in 1945). Article 3 of the Act prohibits arbitrary application of this Act to protect the rights and freedoms related to academic research, cultural and artistic activity and press. As protection of youth is an international norm, and the adverse impact can be smaller than the benefit of the prohibition of child porngraphy, this prohibition has little negative impact to the culture and legal system.

"Tatewari" (Sectionalism) in Japanese government

Aside from the legal and cultural norm implicated in relevant Japanese laws, sectionalism ("tatewari" (縦割り) in Japanese) is a political and cultural norm characterizes behavior of Japanese administrative organizations. "Demarcation" (in Japanese（デマケ）) is one of the most important jergon which all newly hired government offical learns at Day-1. This means each competent authority should stay in its own mandate, and should not catch the ball out of their side of the court. As data protection is an intergovernmental issue and a rapidly changing field, this organizational culture creates impedance for effective policy. Cabinet secretariat（内閣官房）is in charge of coordination of the ministrial works, but they have to assign the tasks to the ministries, and to do so the political objective are cut into the tasks which fit into the each ministry's area with unclaimed areas left untouched.

4. What types of licensing or other control mechanisms would be preferred under the applicable jurisdictions?

Intentionally Omitted

5. Is there a legal restriction on the distribution of data for national security reasons? Would a model trained on the data also fall under such restrictions?

Rules

Under the Foreign Exchange and Foreign Trade Act (FEFTA, 外国為替及び外国貿易法), goods and technology to be used for the development and production of 1) Weapons of Mass Destruction (WMDs) and missile and 2) conventional weapons.

List of items and technologies are listed on the cabinet order (Export Trade Control Order for goods, Foreign Exchange Act technology for technologies) . In addition, Japan has a "catch-all" framework under which export of any (non-listed) items and technologies which could be used for WMDs, missile and conventional weapon development and production can be controlled.

Application

Export is either 1) transfer of an item/technology to another country. 2) Within Japanese territory, transfer of technology to a “non-resident (a) foreigner who stayed in Japan less than 6 months, b) diplomats of other countries and international organization staff (non-Japanese), and c) Japanese residing abroad) Currently, Japanese government is planning to add the new technology transfer restriction to be applied to “resident under strong influence of foreign state” (either Japanese and foreigners). Further definition to be determined) within FY 2022 (April 2022-March 2023). A model trained on the data will be reviewed under the same criteria.

1) Transfer of items/technologies to another country

The exporter should review based on a) item-related risk b) end-user/user risk. The export license is required a) if the item/technology is listed in the above-stated Cabinet Orders b) if the country to be exported is restricted c) if the end-user is listed on Foreign End User List. d) if the exporter knew the possibility that the item/technology could be used for military purposes, or the Minister of Industry,Trade and Economy notified the exporter the requirement of the licensing ("catch-all" provision).

2) Transfer of technology in Japan

Transfer of technology to be used for WMDs, missiles and conventional weapons development and production is prohibited even within Japanese territory, if a) the recipient of the technology is above-stated non-resident b) the giver of the information knows that the recipient transfers the technology to be transferred to overseas (i.e. indirect transfer of technology to another country)
A foreign legal person is categorized as “resident” in Japanese law. Therefore, technically speaking if the information stays in Japan, giving information to the Japanese branch of foreign country is not a violation of FEFTA. However, if the giver knows or should have known the information is further transferred overseas, then the giver could be liable for the violation.

3) Liability of the exporter

If the exporter does a reasonable review of the item/technology export/transfer and follows the licensing procedure, the exporter is not likely to be liable for violation of FEFTA.

Conclusion

Transfer of Items and technology to be used for WMDs and missiles, as well as conventional weapons are prohibited under FEFTA. Transfer is either transfer of items/technologies to overseas and transfer of information to non-resident (foreigners and Japanese residing abroad)

6. Are there any AI technology, including NLP technology that should not be exported (or imported) by NLP researchers?

Intentionally Omitted

A. Useful Resource

General Issues on AI

- https://www.whitecase.com/publications/insight/regulation-artificial-intelligence-europe-and-japan giving a good outlook on AI regulation in Japan. Rather than establishing ruled-

based hard law, METI is prone to set up guidelines and goal-based regulation, since fast-pacing AI development causes difficulties in establishing the bright-lined rules.

- https://www.meti.go.jp/press/2019/12/20191209001/20191209001.html (Japanese)
  https://www.meti.go.jp/press/2019/04/20190404001/20190404001.html（English)
  Contract Guidelines on Utilization of AI and Data/ AI・データの利用に関する契約ガイドライン 1.1版（I find this document prepared by METI could really give a comprehensive perspective on datasets and machine learning development)
- https://www.meti.go.jp/policy/economy/chizai/chiteki/data.html
  不正競争防止法の観点から見たデータや限定提供データの利活用・制限
- In addition to METI, MIC seems to hold “AIネットワーク社会推進会議 AI経済検討会” regularly, and related documents are here. https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/index.html#keizai
  The annual report for 2021 seems to be here (Japanese)
  https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000100.html

Copyright

- https://www.jonesday.com/files/Publication/1f59ff24-9ab2-4f23-9c0e-5905cbe16cda/Presentation/PublicationAttachment/01dcbe95-e9d4-424f-873a-7379b01af09f/Japan%20Legal%20Update%20April%202018.pdf
  Training data and copyright: the amendment of Copyright Law in 2018 extends the definition of “fair use” to allow the use of copyrighted works w/o permission as training data in cases where the expression of the copyrighted work is not perceived by the user
- https://www.cric.or.jp/english/clj/doc/20210624_law.pdf
  Outlook and translation of Japan’s copyright laws from Copyright Research and Information Center

Trade secrets

- https://blogs.orrick.com/trade-secrets-watch/2016/04/18/were-not-gonna-take-it-significant-changes-to-japans-trade-secret-protection-law/
  Summary on the change in trade secret law (Unfair Competition Prevention Act)

Personal Information

- Recruit Career hit with correction advisory for selling data on student job-hunters https://mainichi.jp/english/articles/20190827/p2a/00m/0na/009000c
- Japan- Data Protection Overview
  https://www.dataguidance.com/notes/japan-data-protection-overview

General Legal Resource

E-gov 法令検索
https://elaws.e-gov.go.jp/
Japanese law database operated by the Japanese government.

Japanese Law Translation
http://www.japaneselawtranslation.go.jp/law/?re=01
Database of English translation of Japanese laws operated by the Japanese government. (but sometimes the translation was not the latest version. You must double check with the Japanese site above)

PII

Personal Information Protection Committee
https://www.ppc.go.jp/en/index.html
An independent organization mandated to address PII related issues and monitor the implementation of relevant laws.

Guidelines for Act on the Protection of Personal Information (個人情報の保護に関する法律についてのガイドライン) (in Japanese)
https://www.ppc.go.jp/personalinfo/legal/

Security Export Control

安全保障貿易管理ウェブサイト(Security Export Control Website)
https://www.meti.go.jp/policy/anpo/index.html (Japanese) : maintained by Ministry of Economy, Trade and Industry (METI)
Best source to follow the updates
https://www.meti.go.jp/policy/anpo/englishpage.html (English)
(English site contents are not fully updated. Always refer to a Japanese site!)

FRANCE

Madeleine Hahn de Bykhovetz; Michiel De Wolf; Alfredo Palasciano

Precedence of Laws in France

The French legal system is a civil law jurisdiction with a dual legal system organized based on the concept of the hierarchy of norms, as theorised by Hans Kelsen.

First, as a civil law jurisdiction, France emphasizes the importance of written sources of laws, found in legislative laws, executive instruments, and generally synthesized in written Codes. Unlike in common law jurisdictions, decisions taken by judges are not considered to be a source of law but a mere interpretation and application of the above-mentioned written sources.

Second, as a dual legal system, France distinguishes between public and private law. The former focuses on matters related to public officials and is heard before administrative courts. The latter focuses on relationships between private individuals and is heard before judicial courts.

Finally, according to the concept of the hierarchy of norms, sources of law in France are hierarchically distinguished and each source of law must be compliant with superior sources.

- The highest source of law is the “constitutionality bloc”, which includes the 1958 Constitution, the Preamble to the 1946 Constitution, the 1789 Declaration of the Rights of the Man and of the Citizen, the Environmental Charter, as well as a set of Fundamental Principles and Constitutional Objectives;
- The second highest source of law is the “conventionality bloc”, which includes international sources such as international treaties and conventions and European Law (as a matter of French law, European Law remains inferior to the National Constitution);
- The third highest source of law is the “legality bloc”, which includes laws by the legislature and executive “ordonnances” by the Government taken in lieu of Parliament;
- The fourth highest source of law includes general principles of the law, namely unwritten rules with a general scope;
- Finally, the lowest source of law is the “regulatory bloc”, which includes executive orders issued by the President or Prime Minister (“décrets”) and executive orders issued by administrative authorities (“arrêtés”).

Because France is part of the European Union, EU law is also a source of law within the country, as exemplified in the pyramid above. EU law is composed of various sources of law with different characteristics.

- EU law is composed of primary sources of law, that is EU Treaties (the Treaty on the EU and Treaty on the Functioning of the EU), as well as the Charter of Fundamental Rights and general principles of law delineated by the European Court of Justice. These principles define the institutions and competences of the EU.
- EU law is also composed of secondary sources of law, which includes Regulations and Directives adopted by EU institutions. Regulations are generally, mandatorily, and directly applicable in all Member states, while Directives adopt mandatory or optional rules that need to be transposed in and adapted by each Member state into national law to have legal effect.

Regarding the French jurisdiction's regulation of AI and data related issues, it is still a work in progress as legislative and executive rules are being adopted, updated, and implemented. Under the impulsion of the EU, France and other Member states have a set of regulations that allow them to distinguish themselves from many other jurisdictions, but these still remain insufficient to tackle all issues that arise in connection with the subject matter of this research.

The single most innovative of legislation has been GDPR, an EU regulation on personal data: due to the important size of the European market for most companies relying on user and customer data, the regulation has prompted a vague of compliance by major and minor companies on the use of personal data and is serving as a model for many foreign jurisdictions. And the EU wants to capitalise on this important market and the possibility to serve as a model for other jurisdictions through a new regulation focused on AI.

France also seems to be a jurisdiction ready to embrace technological innovations by providing them with a certain legal framework in order to gain a competitive advantage over other countries. It has for instance adopted provisions specifically regulating text and data mining regarding copyright protected information or has recognised a sui generis database right to protect intellectual property related to the creation of databases. Both innovations were the fruit of European directives, a unique judicial tool that allows the EU to harmonize legal rules among Member states while granting them enough autonomy to adapt rules that naturally blend within the country's legal and social culture.

Nonetheless, despite France's desire to regulate new technologies and the EU's impulsion to adopt innovative legislation, the French legal framework still contains judicial voids. For instance, copyrights are not entitled to a legal framework providing for their licensing, unlike brands and trademarks, which therefore needs to be done pursuant to general contract law principles.

A. IP Questions

1. **Are databases, data training sets and data models protected by IP rights and, if so, which ones?**

The EU Directive 96/9/CE introduced a dual protection of databases that has been implemented by France in its Code of Intellectual Property.

First, databases are protected through traditional copyrights law. Article L111-1 states that the author of an intellectual creation benefits from a protection of its creation through a copyright that is exclusive and enforceable against anyone. Article L112-1 defines the scope of the protection as being applicable to any intellectual creation, regardless of its genre, form of expression, merit, or purpose, while Article L112-3 expressly includes into the scope databases that constitute intellectual creations through the choice or arrangement of materials. Article L112-3 also defines databases as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means".

Second, databases are protected through a sui generis database right. Article L341-1 states that the producer of a database benefits from a protection of the content of the database: the provision defines producer as any person taking the initiative and risk of investing in the database while for the protection ; the provision subordinates it to a substantial financial, material, or human investment in the constitution, verification, and presentation of the content of the database. The provision specifies that this sui generis right is independent and cumulative to the above-described copyright applicable to databases.
The substantial investment requirement does not refer to the “resources used for the creation of materials which make up the contents of a database”, but to the “resources used to seek out existing independent materials and collect them in the database” (4 decisions by the CJEC issued on November 9, 2004: The British Horseracing Board Ltd e.a. ; Fixtures Marketing Ltd c/ Organismos Pronostikon ; Fixtures Marketing Ltd c/ Oy Veikkaus AB ; Fixtures Marketing Ltd c/ Svenska Spel AB).
Article L342-1 grants the producer two rights : the right to prohibit the extraction through a permanent or temporary transfer of all or of a substantial part of the content of the database ; the right to prohibit the reuse of all or of a substantial part of the content of the database by making it publicly available. Article L342-2 grants the producer an additional right to prohibit the repeated and systematic extraction or reuse of non substantial parts of the content of the database when such extractions and reuses exceed the normal use of the database. Article L342-3 introduces certain limitations on these rights when the database has been made publicly available by its rightsholder, including exceptions and limitations allowing for text and data mining analysed in Question 1 of Part C.

The copyright protection focuses on the database itself and is applicable when there is an element of originality to the database so as to make it an intellectual creation: the mere aggregation of existing information should therefore not be protected unless there is a certain originality to the architecture of the assembly. The database right focuses on the content of the database and is applicable where there has been a substantial investment in its constitution, verification, and presentation : regardless of originality, the aggregation of data should therefore be protected provided that it required a substantial investment.

The above-described protections should be transposable to data training sets and data models. One would need to consider whether the given set or model falls within the definition of a database and whether it complies with either the originality or the substantial investment requirement.

2. Are there legal IP concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web?

The issue is whether it is possible to publish a language model that has been trained on a published dataset, itself based on a crawl of the web.

Article L112-3 defines a database as “a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means”. As elaborated in Question 1 of Part A (above), IP rights applicable to databases can either be copyrights (Article L111-1 and L112-3) or sui generis database rights (Article L341-1).

When a database is protected by copyrights or a sui generis database right, there are limitations and exceptions to the rights of the rightsholder that allow the use of the database and its information by third-parties.

Article L122-5 of the French Code of Intellectual Property introduces exceptions and limitations to rightsholder protections regarding their works. If the rightsholder has divulged the work, then it may not prohibit:

- *2° the copy or reproduction of the work solely for private use purposes, unless such copy or reproduction concerns an electronic database (the provision is therefore never applicable to databases)*
- 3° a) analyses et short citations justified by the critical, polemic, pedagogical, scientific, or informational character of the work with which the analyses or short citations are incorporated into, provided that the name of the rightsholder and the source are mentioned
- 6° the temporary reproduction with transitory or accessory characters that is an integral and essential part of a technical proceedings and that has for sole purpose allowing the use of the work, provided that the temporary reproduction has no own economic value
- 8° the reproduction and presentation of a work for conservation purposes or for the purpose of preserving the conditions of its consultation in the framework of research or private studies by private individuals, in specific settings, and without economic or commercial advantages
- 10° the copy or reproduction in a digital form of the work for text and data mining purposes and pursuant to Article L122-5-3

Article L342-3 of the French Code of Intellectual Property introduces exceptions and limitations to rightsholder protections regarding their databases. If such a rightsholder has made the database publicly available, then it may not prohibit:

- 1° the extraction or reuse of a non substantial part of the content of the database by a person with lawful access to it
- 4° the extraction and reuse of a substantial part of the content of the database for illustrative purpose, in the framework of research, to a public primarily composed of researchers, provided that no commercial exploitation ensues and that compensation for the rightsholder is ensured
- 5° the extraction of reuse of a database pursuant to Article L122-5 8°, namely : the reproduction and presentation of a work for conservation purposes or for the purpose of the conditions of its consultation in the framework of research or private studies by private individuals, in specific settings, and without economic or commercial advantages.
- 6° the extraction, copy, or reproduction in a digital form of a database for text and data mining purposes and pursuant to Article L122-5-3

Provided that the initial data published on and crawled from the web is copyrighted and publicly available, the initial dataset on which the language model is trained could constitute a database within the meaning of Article L112-3 and protected under either a copyright or a sui generis database right. The fact that this dataset is made publicly available means that its rightsholder may not prohibit text and data mining and other extraction of information, provided that the miner or extractor falls within the scenarios enumerated above. The publication could potentially fall within various cited provisions, especially L342-3 4°, provided that its conditions are satisfied.

3. Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g. HTML tags or document structure)?

The question is whether there are any differences between the legal implications of publishing datasets that have plain text only or datasets that contain additional information (e.g. HTML tags or document structure).

Article L112-3 defines a database as “a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means”.
As elaborated more extensively in Question 1 of Part A (above), IP rights applicable to databases can either be copyrights (Article L111-1) or sui generis database rights (Article L341-1). Copyrights protections expressly cover databases (L112-3) and are only applicable where there is an element of originality in the created work through the choice or arrangement of materials (L112-3). Database rights protections are available to producers for the content of their database (L341-1) and are applicable only where there has been a substantial financial, material, or human investment in the constitution, verification, and presentation of the content of the database (L341-1). The copyright protection focuses on the database itself and is applicable when there is an element of originality to the structure of the database so as to make it an intellectual creation. The database right focuses on the content of the database and is applicable where there has been a substantial investment in its constitution, verification, and presentation. French law specifies that database rights are independent and cumulative to copyrights applicable to databases (L341-1).

If a dataset is considered to be a database, the originality of its structure or the investment undertaken in forming its content will determine whether the database is subject to copyrights and/or database base rights. While there is no legal distinction between databases containing plain text only or also additional information, the presence of such additional information might affect the originality of the structure or the investment needed to create the database, thus modifying the IP protections applicable to the database.

4. What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?

The issue is whether there are any legal IP concerns for publishing metadata that has been extracted from a dataset (e.g. URLs, publication dates, timestamps).

Based on the legal rules developed in detail for the questions above, such a practice would fall within Article L342-3, 1°, especially considering the extremely limited amount of information that is being published.

Article L342-3 of the French Code of Intellectual Property introduces exceptions and limitations to rightsholder protections regarding their databases. If such a rightsholder has made the database publicly available, then it may not prohibit: 1° the extraction or reuse of a non substantial part of the content of the database by a person with lawful access to it. The provision specifies that substantiality is determined either quantitatively or qualitatively.

Additionally, Article L122-5, 3°, 3) of the French Code of Intellectual Property states that where a rightsholder has divulged its work, it may not prohibit “analyses et short citations justified by the critical, polemic, pedagogical, scientific, or informational character of the work with with the analyses or short citations are incorporated into”, provided that the name of the rightsholder and the source are mentioned.

5. Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR/Pile) (for example, such a dataset may contain information like: in the nth entry of C4, there is a "<b>" html tag after the mth character)?

Based on the legal rules developed in detail for the questions above, such a practice would fall within Article L342-3, 1°, especially considering the extremely limited amount of information that is being published.

Article L342-3 of the French Code of Intellectual Property introduces exceptions and limitations to rightsholder protections regarding their databases. If such a rightsholder has made the database publicly available, then it may not prohibit: 1° the extraction or reuse of a non substantial part of the content of the database by a person with lawful access to it. The provision specifies that substantiality is determined either quantitatively or qualitatively.

6. What are the IP risks related to data collection directly from persons? For example when you interview people or can they donate data etc.

The issue is whether there are any legal IP concerns for data that is collected directly from individuals.
While such collection should not present any IP related risks, if it concerns personal data, a set of specific provisions based on or linked to GDPR (the EU General Data Protection Regulation) could be applicable. Please refer to Part D on privacy-related issues.

7. Can NLP researchers train a language model on borrowed library books - e.g., from the Internet Archive or other online book lender?

The issue is whether NLP researchers train a language model on library books that have been borrowed. The same legal rules and reasoning related to the use of works under copyrights protections are applicable to this issue. Please refer to Question 1 Part A (above).

8. How do rights on the source data (e.g. copyright) transfer to the trained model?
    a. Will this depend on where training occurs?
    b. Will this depend on where data is gathered?

B. Licensing Questions

1. Are data sets licensable? Are models licenseable? What restrictions are appropriate or preferable under the jurisdictions?

The issue is whether databases, data sets, and models are licensable and, if so, whether such licenses are subject to any restrictions.

Regarding private information, licensing agreements are regulated by contract law and licensing contracts are often assimilated to renting contracts of intangible objects.

First, provisions focused on copyrights and database rights provide for the right of the rightsholder of the database to license the database to another.
Regarding copyrights protections, Article L122-6 of the French Code of Intellectual Property provides that the right holder's right to exploit its software include the right effectuate are authorise another to effectuate (1°) a permanent or temporary reproduction, (2°) a translation, adaptation, arrangement, or modification of the software and the reproduction of such modified software, (3°) the presentation to the market, gratuitously or at cost, of the exemplaries of the software.
Regarding database rights protections, Article L342-1 of the French Code of Intellectual Property provides that extraction by transfer and reuse of protected databases can be the subject matter of license agreements.
More interestingly, Article L122-7-1 of the French Code of Intellectual Property provides that authors of copyrighted material are entitled to put their work at the gratuitous disposal of the public.

However, unlike for patents and brands, France does not expressly recognise licenses for copyrights. We could therefore only suppose that a contract could be drafted based on generally applicable contract law principles.
Article 1101 of the French Civil Code defines a contract as a voluntary agreement between parties to create, modify, transmit, or extinguish obligations , while Article 1128 of the French Civil Code requires there to be consent and capacity to contract for the parties and lawful and certain for the content.
Some doctrinal authors argue that such contracts should also comply with provisions usually applicable to renting contracts or lending contracts. Article 1709 of the French Civil Code defines a renting contract as a contract that obligates one party to allow the other party to use an object in exchange for consideration, while Article 1713 of the French Civil Code makes it applicable to intangible property too.

Regarding public information, France has adopted special provisions present in the French Code of Relations between the Public and the Administration.
Article L323-1 states that in general, when one reuses public information one can enter a license agreement, and if such reuse is subject to royalty payment then such license agreement is mandatory.
Article L323-2 specifies that the license agreement fixates the conditions of the reuse of information and may introduce restrictions on such reuse only on the basis of general interest motives and provided that they are proportionate and do not hinder competition.
Article D323-2-1 lists the possible licenses that can be entered into for when public data is reused gratuitously and provides a specific list for when such information takes the shape of a software. The latter licenses include ; permissive licenses (Berkeley Software Distribution License ; Apache ; CeCILL-B ; Massachusetts Institute of Technology License ") and licenses with an obligation of reciprocity (Mozilla Public License ; GNU General Public License ; CeCILL ; European Union

Public License ; Eclipse Public License). Article D323-2-2 adds that should the administration want to use a license that is not included in the above-mentioned list, it may ask the government to be allowed to use this different license.

In France, data sets published by the administration must be free to use. The data must be publicly available and free of charge (the only entities which can charge for their data are the IGN and SHOM, according to the Décret n°2016-1617 of Nov. 29, 2016). And the use of data must be free, meaning that anyone (public and private actors included), can use it.

2. What are the relationships between laws, licenses, and terms of use?

Unlike in the US, France has more mandatorily applicable laws that restrain the parties' freedom to contract. Thus, unless a French law is a default rule that can be contracted around, it has superior binding force to licenses and accepted terms of use, which are mere contracts. Indeed, the terms of a license and general terms of use must comply with mandatory applicable laws and any non-compliant rule will either be inapplicable or render the entire license or set of terms null. Regarding terms of use, it is important that they have been accepted to be considered a legally binding, unilateral contract to which the user has consented to.

3. What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?

Regarding the conditions of validity, please refer to Question 1 Part B (above). It is up to the NPL researchers using re-published or derived datasets to verify the existence and validity of licenses attached to the used datasets.

4. What about if the users download or copy their own data and then provide it to NLP researchers directly?

The issue and solution varies depending on what is meant by providing personal data researchers.

If providing is intended to mean donate, such donation of personal data is unlikely to be valid, despite the lack of express provisions on the matter. Personal data is intrinsically attached to individuals and is protected through various fundamental rights, most likely making it unalienable.

If providing is simply understood as putting at the disposal of the researchers, then the issue is most likely regulated by GDPR. Please refer to Part D on privacy-related issues.

5. Does the license that the dataset is shared under override the terms and conditions?

The issue is whether the license attached to a data set has overriding authority over terms and conditions.

As stated for Question 3 of Part B (above), both licenses that have been agreed upon and terms of use that have been accepted constitute legally binding contracts between the parties to it. Thus,

they have the same legal standing and the overriding nature of one over the other will be a case-by-case issue that will depend on various factors such as the date each license and term has been entered into, the compatibility of their respective clauses, etc.

6. Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?

As it has been explained for various other questions, the extraction, copy, and reuse of information that is protected by copyrights is permitted for a limited number of purposes, including scientific research and text and data mining, usually provided that it has been made publicly available. Please see the reasoning exposed under Question 2 Part A and Question 1 Part C.

7. Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?

The first issue is whether it is possible to release data gathered by text and data mining through a royalty free license. For a more expansive description of French licensing law, please refer to Question 1 Part B (above). Article L122-7-1 of the French Code of Intellectual Property provides that authors of copyrighted material are entitled to put their work at the gratuitous disposal of the public. However, this power is reserved to be the owner of the work that is placed at the public's disposal. It is therefore unlikely that a person mining data over which it has no legal right can just use to enter licenses.

The second issue is whether it is possible to release a model trained on the collected data through a royalty free license. If the data set or training set falls within the definition of a database (Article L112-3 of the French Code of Intellectual Property) and the database is recongised a copyright or database right protection, then the person owning the database benefits from the right to either put it at the disposal of the public (L122-7-1 of the French Code of Intellectual Property for databases subject to copyrights protections) or license it (L342-1 of the French Code of Intellectual Property for databases protected by a database right). Please see Question 1 Part A and Question 1 Part B for a more detailed explanation of copyrights, database rights, and licensing.

C. Text Data Mining and Fair Use Questions

1. What are the types of legally permitted text data mining?

IN ENGLISH

The issue is what types of text and data mining are permitted under French and EU law.

The French Code of Intellectual Property contains multiple exceptions and limitations to copyright and sui generis rights to allow for text and data mining.

Articles L. 122-5, 10° and L. 342-3, 6° introduce a Research Exception for, respectively, sources and databases. They prohibit a rightsholder that has made its work or database publicly available from preventing the extraction, copy, and reproduction in a digital form of the wotk or database for the purpose of text and data mining pursuant to Article L. 122-5-3.
Article L122-5-3, I defines text and data mining as the "implementation of an automated analysis technique of texts and data in a digital form in order to extract information, including patterns, trends and correlations". Article L122-5-3, II limits text and data mining under Article L. 122-5, 10°, without requiring the consent of the rightsholder, to mining executed for scientific research purposes by research organisms and other public actors. Article L122-5-3, II also allows rightsholder representatives and persons engaged in text and data mining to reach an agreement on the good practices related to the application of the provision.
While these provisions are applicable to text and data mining with absolute certainty, their application is limited as the text and data mining must be focused on scientific research. Articles L. 122-5 and L. 342-3 contain other provisions that do not specifically target text and data mining but that are considered by many as applicable to it. The uncertainty surrounding such applicability has encouraged the legislator to adopt the more specific, above-described provisions.
Prior to the adoption Article L. 342-3, 6°, Article L. 342-3, 5° already stated that when a database is made publicly available by its rightsholder, the latter may not prohibit the extraction and re-use of the database made pursuant to Article L. 122-5, 8°. Article L. 122-5, 8° allows the reproduction and presentation of the database for research or private-study purposes by private individuals, in specific settings, and for no economic or commercial purpose. The provision is still in place and suffers from limited scope that needs to be focused on scientific research.

The EU Directive 2019/790/EC for a Digital Single Market contains two mandatory exceptions and limitations for text and data mining. Article 2 defines text and data mining as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations".
Article 3 introduced an exception for text and data mining for scientific research purposes. The exception is limited to reproductions and extractions made by research organisations and cultural heritage institutions and for the purpose of scientific research. Article 2 defines research organisations as including any entity with a primary purpose of conducting scientific research or carrying out educational activities involving such research (condition 1), either on a not-for-profit basis, or based on a reinvestment of the profits in scientific research, or pursuant to a public interest mission (condition 2), and with an access to the results not preferentially limited to the undertaking with a decisive influence on the entity. This exemption has the advantage of being applicable to commercial and non-commercial text and data mining and cannot be prohibited by the rightsholder.
Article 4 introduced another exception for text and data mining for other purposes. Such mining remains possible provided that the rightsholder has not expressly reserved the subject-matter of the mining "in an appropriate manner, such as machine-readable means in the case of content made publicly available online". This exception is therefore universally applicable and grants the rightsholder an opt-out option, allowing it to prevent such mining, unless it is for scientific research purposes.

These mandatory provisions were supposed to be implemented by the EU Member States, including France, not later than 26 July 2021. However, most States including France have failed to do so and are now subject to legal action from the European Commission.

IN FRENCH

La question est de savoir quels types de fouilles de textes et de données sont autorisés par la législation française et européenne.

Le Code de la propriété intellectuelle contient de multiples exceptions et limites aux droits d'auteur et aux droits sui generis pour permettre l'exploration de textes et de données.
Les articles L. 122-5, 10° et L. 342-3, 6° introduisent une exception de recherche pour, respectivement, les oeuvres et les bases de données. Ils interdisent à un auteur qui a mis son oeuvre ou sa base de données à la disposition du public d'empêcher l'extraction, la copie et la reproduction numériques de l'oeuvre ou de la base de données réalisées conformément à l'article L. 122-5-3.
L'article L122-5-3, définit fouille de textes et de données comme "la mise en œuvre d'une technique d'analyse automatisée de textes et données sous forme numérique afin d'en dégager des informations, notamment des constantes, des tendances et des corrélations". Le II de l'article L122-5-3 limite les opérations de fouille de textes et de données visées au 10° de l'article L. 122-5 qui ne nécessitent pas le consentement de l'auteur aux fouilles réalisées à des fins de recherche scientifique par des organismes de recherche et d'autres acteurs publics. L'article L122-5-3, II permet également aux représentants des titulaires de droits d'auteur et aux personnes pratiquant la fouille de textes et de données de conclure un accord sur les bonnes pratiques liées à l'application de cette disposition.
Bien que ces dispositions soient applicables à la fouille de textes et de données avec une certitude absolue, leur application est limitée car la fouille de textes et de données doit être axée sur la recherche scientifique. Les articles L. 122-5 et L. 342-3 contiennent d'autres dispositions qui ne visent pas spécifiquement la fouille de textes et de données mais qui sont considérées comme applicables à cette pratique. L'incertitude entourant cette applicabilité a encouragé le législateur à adopter les dispositions plus spécifiques décrites ci-dessus.
Avant l'adoption de l'article L. 342-3, 5°, l'article L. 342-3, 5° précisait déjà que lorsqu'une base de données est mise à la disposition du public par son titulaire de droits, ce dernier ne peut interdire l'extraction et la réutilisation de la base de données effectuées en application de l'article L. 122-5, 8°. L'article L. 122-5, 8° permet la reproduction et la présentation de la base de données à des fins de recherche ou d'étude privée par des particuliers, dans un cadre spécifique, et sans finalité économique ou commerciale. Cette disposition est toujours en vigueur mais souffre d'un champ d'application limité qui doit être axé sur la recherche scientifique.

La directive européenne 2019/790/CE pour un marché numérique unique contient deux exceptions et limitations obligatoires pour l'exploration de texte et de données. L'article 2 définit la fouille de textes et de données comme "toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations".

L'article 3 introduit une exception pour la fouille de textes et de données à des fins de recherche scientifique. Cette exception est limitée aux reproductions et aux extractions effectuées par des organismes de recherche et des institutions chargées du patrimoine culturel et aux fins de la recherche scientifique. L'article 2 définit les organismes de recherche comme incluant toute entité dont l'objet principal est de mener des recherches scientifiques ou d'exercer des activités éducatives impliquant de telles recherches (condition 1), soit dans un but non lucratif, soit sur la base d'un réinvestissement des bénéfices dans la recherche scientifique, soit en vertu d'une mission d'intérêt public (condition 2), et dont l'accès aux résultats n'est pas limité de manière préférentielle à l'entreprise ayant une influence déterminante sur l'entité. Cette exception a l'avantage d'être applicable à l'extraction de textes et de données à des fins commerciales et non commerciales et ne peut être interdite par le titulaire des droits.
L'article 4 introduit une autre exception pour la fouille de textes et de données à d'autres fins. Cette exploration reste possible à condition que le titulaire des droits n'ait pas expressément réservé l'objet de l'exploration "d'une manière appropriée, par exemple par des moyens lisibles par machine dans le cas de contenus mis à la disposition du public en ligne". Cette exception est donc universellement applicable et offre au titulaire des droits une option de retrait, lui permettant d'empêcher une telle exploitation, sauf si elle est effectuée à des fins de recherche scientifique.

Ces dispositions obligatoires étaient censées être mises en œuvre par les États membres de l'UE, dont la France, au plus tard le 26 juillet 2021. Cependant, la plupart des États, dont la France, ne l'ont pas fait et font maintenant l'objet d'une action en justice de la part de la Commission européenne.

2. Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question?

After extensive research, there do not seem to be legal provisions regulating web crawling in France. The practice is therefore most likely regulated by intellectual property provisions described above as well as data regulation for personal information.

3. If NLP researchers can crawl data ourselves, what are the limits regarding using this data for training a language model? Can NLP researchers redistribute the data afterwards? If so, under which license and in which geographical regions?

After extensive research, there do not seem to be legal provisions regulating web crawling in France. The practice is therefore most likely regulated by intellectual property provisions described above as well as data regulation for personal information.

4. If NLP researchers use parts of the Pile, Common Crawl, OSCAR or C4, can NLP researchers redistribute the data later? If yes under which conditions and in which countries?
5. What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify? The edition conversations of Wikipedia?

6. Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can't be scrapped?

D. Privacy Questions

1. Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?

The issue is whether there are legal concerns around accessing web content that has PII, and legal concerns about datasets including PII. This answer will cover general provisions concerning accessing content and processing data with PII in France.

France is subject to the EU's General Data Protection Regulation (GDPR) law for questions surrounding private data, as France adopted the GDPR through amendments to its law n°78-17 (*Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*). The GDPR became applicable in France from May 25, 2018 onwards.

The concept of PII (Personally Identifiable Information) does not exist in Europe. Instead, there are two main types of personal data which are recognized in the GDPR rules. Firstly, there is personal data in the general sense, which can be processed under certain conditions in the EU. Then, there are "special categories of personal data", which are subject to article 9 of the GDPR and have stricter rules surrounding their processing.

Under the GDPR, personal data can be processed and accessed under certain conditions, including consent from the individual (GDPR article 6-1(a)). It must be collected fairly, for specific purposes, and limited to what is necessary for those purposes (GDPR article 5). (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434) Processing of personal data is lawful only under certain circumstances: if the data subject consented to the processing for specific purposes (GDPR Article 6-1(a)), if the processing is necessary for the performance of a contract (GDPR Article 6-1(b)); if it's necessary for compliance with a legal obligation (GDPR Article 6-1(c)); if it's necessary to protect the vital interests of the data subject (GDPR Article 6-1(d)); if it's necessary for the performance of a task carried out in the public interest (GDPR Article 6-1(e)); and if it's necessary for legitimate interests pursued by the controller, except when those interests are overridden by the data subject's rights (GDPR Article 6-1(f)).

However, some types of personal data cannot be processed; those are "special categories" of personal data. According to article 9 of the GDPR, among others, categories of data which cannot be processed are the following: personal data revealing ethnic origin, religious beliefs, genetic or biometric data, health data, and data about sexual orientation. However, article 9 also lays out several exceptions to that rule, and cases where special categories of data can be processed. Among those, there is if the data subject gave explicit consent to the processing of their data (GDPR Article 9-2(a)). Another case where the data could be processed is if it is done "in the course of its legitimate activities with appropriate safeguards" by a non-for-profit body; and on the

condition that the processing relates to members of the body or persons with regular contact with it, and that the data is not disclosed outside of that body without consent (GDPR Article 9-2(d)). Another exception is with data which is “manifestly made public” by the subject (GDPR Article 9-2(e)).

The question of what is “manifestly made public” is not defined by the GDPR itself. However, according to the *Guidelines 8/2020 on the targeting of social media users* by the European Data Protection Board (p.32-33), several elements can be used to determine whether data is made “manifestly” public (the word itself inferring a high threshold). (https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202008_onthetargetingofsocialmediausers_en.pdf)

Those elements include the default settings on social media platforms (whether the subject made their profile public), the nature of the platform, the accessibility of the page where sensitive data is published, the visibility of information, and whether the data was published by the subject themselves (or by a third party). According to the European Data Protection Board, a single element might not always be enough to prove it was “manifestly” made public.

One element specific to the GDPR is its extraterritoriality: according to Article 3 of the GDPR, processing of data belonging to individuals of the EU is subject to the GDPR, even if the processing and the processor do not belong to the EU. This applies when the processing activities are related to the offering of goods or services (even without payment), or the monitoring of their behaviour (GDPR Article 3-2).

2. Intentionally Omitted

3. What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exceptions for research purposes?

The issue is what mechanisms NLP researchers would have to put in place to ensure the takedown on personal data in accordance with French law. Since the issue in this case is “takedown”, the answer to the question will only include matters related to limitation or takedown of data, and will not cover consent (mentioned in the answer to question 4).

The first step that researchers would have to take is to be able to identify what data counts as personal data, and manage to filter and exclude it if it is not necessary for research. Personal data is defined by the GDPR in its article 4(1), as any information related to an “identifiable natural person”, who can be identified directly or indirectly, and in particular through an identifier such as: “a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;”. In principle, personal data can only be kept for “no longer than necessary” to complete the purpose it was collected for (GDPR article 5(e)), and there would therefore need to be filtering and takedown of data which has fulfilled that purpose.

However, the GDPR provides exceptions for (among others) scientific or historical research purposes in its Article 5(e), enabling the storage of personal data for longer periods of time. But

as is mentioned in Article 89(1) of the GDPR, even data gathered as part of research needs to respect safeguards. Those include putting in place mechanisms to respect the general principle of data minimization; pseudonymisation when possible; and to use processing which limits identification of data subjects if the research purpose can be achieved that way.

Also, the “right to be forgotten” could be another reason for the takedown of personal data: upon request and without undue delay, controllers are under the obligation to erase personal data (GDPR Article 17). However, there is also an exception in this case for data gathered for archiving, scientific or historical research, and statistical purposes, if it would impair the objectives of the processing. Unless researchers fit in those categories (scientific or historical research), they would have to prepare mechanisms ensuring a swift takedown of personal data upon request.

NLP researchers would therefore have to put in place mechanisms to either identify personal data and organize its takedown if it is unnecessary for research purposes (following the principle of data minimization of the GDPR) or upon request; or if the data is still needed, create mechanisms for pseudonymisation and limiting identification of personal data.

4. How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?

The issue is how consent of the individuals could influence what an organization could or could not do in France. Consent of individuals can enable organizations to gather data: as mentioned in question 1, consent is one of the conditions for processing to be lawful, under Article 6-1(a) of the GDPR. Even categories of personal information concerning sensitive data (race, political opinions, sexual orientation, etc.) which are usually forbidden, can be processed if the individual gives consent (GDPR Article 9-2(a)).

Article 7 of the GDPR defines conditions for consent. If consent is obtained through a written declaration which also includes other matters, the request for consent should be presented in an “intelligible and easily accessible form, using clear and plain language” (GDPR Article 7-2). Consent should be withdrawable at any time, as easily as it was to give consent (GDPR Article 7-3).

There are specific conditions for obtaining consent of children under 16, laid out in Article 8 of the GDPR. France, when adopting the GDPR into national law, changed the age of consent from 16 (GDPR Article 8), to 15 years old (*Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés,* Article 45). For children under 15, processing data is only lawful if the consent was also authorized by the holder of parental authority over the child (*Loi n° 78-17 du 6 janvier 1978,* Article 45); and the controller has to make “reasonable efforts” to verify that the consent was given by the holder of parental authority (GDPR Article 8). Finally, the request for consent has to be written with clear and simple words, which should be easily understandable by the child (*Loi n° 78-17 du 6 janvier 1978,* Article 45).

Consent of individuals can therefore be the key for an organization to gather data, and even sensitive data if necessary. However, the GDPR’s conditions for consent must be followed, and obtaining consent from children is subject to more conditions.

5. What are the privacy risks related to data collection directly from persons? For example when you interview people or they donate data etc.

The issue is what privacy risks are related to gathering data directly from individuals.

Privacy risks related to collecting data directly from persons could be related to free and informed consent. As mentioned in the answer to question 4, consent to data processing must follow the conditions mentioned in Article 7 of the GDPR, and the data must be processed in a fair and transparent manner (GDPR Article 5).

When data is collected directly from persons, there is an obligation to provide information, stated by Article 13-1 of the GDPR. The information to be provided includes (among others): identity and contact details of the controller; purposes for processing; legal basis for the processing; etc. Article 13-2 of the GDPR also requires controllers to inform the data subject of, among other things, the period for which the data will be stored; the existence of a right to erasure or rectification of personal data; the right to withdraw consent at any time; and the right to lodge a complaint with a supervisory authority.

Risks may therefore be of not informing data subjects well enough of the conditions surrounding the data processing. If the data is collected directly and immediately, through interviews, for example, another concern could be about the withdrawal of consent. According to Article 7 of the GDPR, consent should be withdrawable at any time – there should therefore be a mechanism put in place to ensure that data subjects are able to easily withdraw their consent, even after the interview is over.

E. Prohibited Content

1. What types of data may be prohibited from being text data mined?

The issue is whether some types of data are prohibited from being text data mined.

There does not seem to be a specific legal provision in French law which defines what content is allowed to be text data mined or not. However, it is probable that content which is forbidden from being created and published on the internet would also be prohibited from data mining.

The French law n°2004-575 (*loi n°2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique*) defines what content is prohibited. In its Article 6-I-7, the law defines content which is prohibited from the internet to be: negating crimes against humanity; inciting or committing acts of terrorism; inciting hate of persons because of their race, sex, sexual orientation, gender identity or disability; child pornography; encouraging violence, including inciting sexual and gender-based violence; degrading human dignity. (https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000043982464/)

2. What types of data may be prohibited from being stored or distributed?

The issue is what types of data could be prohibited from being stored or distributed. The term "distributed" has been interpreted here as meaning making data available for access (but without editing, creating, or publishing it).

The types of data which are prohibited in France are the same as in the question above. However, "storage" of prohibited data involves specific legal obligations.

France adopted the European Directive on electronic commerce (*Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000*), which defines in its article 14 obligations related to "hosting" information. "Hosting" is defined as a service consisting of storing information provided by a recipient of the service, at the latter's request.

Article 6-I-2 of French law n°2004-575 (*loi n°2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique*) defines hosting as providing (for public access in a communication service) storage of information including signals, text, images, sounds or messages of any nature. However, if the service provider is recognized as a "host" by the law, they would not be liable for the information they stored at the request of the service's recipient. A host is not liable for information they stored if they did not know that it was illicit (and did not see any facts or circumstances which showed it) (Article 6-I-2). Also, a host would not be held liable for the content they made available if, once they noticed it was prohibited, they immediately deleted the data or blocked access to it (Article 6-I-2).

Although the types of data which are prohibited are the same as in question 1, if the data was just unknowingly stored by a service provider, they wouldn't necessarily be held liable for it. However, they would have the obligation to promptly inform public authorities of any illegal activity which they noticed or was reported to them (Article 6-I-7). The service provider is also under the obligation of creating an accessible and visible reporting system which could allow any person to report the presence of illegal data on their platform (Article 6-I-7). However, they are not under the obligation to actively monitor or search for illegal content (EU Directive on electronic commerce, Article 15).

3. What types of data may be prohibited from being generated?

The issue here is what types of data are prohibited from being generated. The term "generated" was interpreted here to mean created (in whole or in part, including through editing).

The question of what types of data are prohibited in France was answered in question 1. However, as "generated" could be understood broadly, this answer will cover the consequences of editing such content, which is different from storing it.

Under French law, "editing" content is referred to in Article 6-II of law n°2004-575. According to an article published by the French Constitutional Council (Conseil constitutionnel, https://www.conseil-constitutionnel.fr/nouveaux-cahiers-du-conseil-constitutionnel/contenus-illicites-sur-internet-et-hebergeurs), editing content differs from hosting it as it involves "active" participation in the content.

Liability over online content is governed by the law on freedom of press of July 29, 1881 (*loi du 29 juillet 1881 sur la liberté de la presse*, Article 42), which makes editors liable for content they published.

Liability is therefore much broader for editors (who contributed to generate the content), compared to hosts who simply stored it or made it available.

4. What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?

Creating prohibitions for content generally calls for a balance between different interests. On the one hand, prohibiting content ultimately goes against the right to freedom of expression, freedom of information, and right to privacy. On the other hand, prohibiting harmful content is important to protect the rights of vulnerable persons. In the case of France, certain types of content are prohibited to prevent human rights violations, including discrimination and violence, and to protect children. Another main category of content which is prohibited in France is content which endangers national security.

An example of the balance of interests in prohibiting content can be found in the adoption of *Regulation 2021/784* by the European Parliament in April 2021 (referred to in answer 6). From 2015, France (as well as other European countries) was faced with several terrorist attacks, and governments found the internet to be one of the ways that terrorist propaganda was spread. (As an example, Europe's law enforcement agency (Europol) published a report concerning the role of internet in radicalization
https://www.europol.europa.eu/publications-events/publications/online-jihadist-propaganda-2020-in-review.)

The European Parliement therefore adopted its *Regulation 2021/784* in April 2021 with the goal to curb the spread of terrorist content online. Adopting the regulation was subject to strong debate however, with around 60 NGOs including Human Rights Watch and Amnesty International voicing concern over risks to freedom of speech, freedom of information, and privacy. Another concern was that the short timeframe given to platforms to delete terrorist content would require them to use automation tools, which might risk mistakingly deleting lawful content as well.
(https://www.lemonde.fr/pixels/article/2021/03/25/moderation-en-ligne-des-ong-appellent-le-parlement-europeen-a-rejeter-le-reglement-contre-les-contenus-terroristes_6074434_4408996.htm)
Although the threat to be prevented by that regulation was severe, the debate surrounding the adoption of the regulation could therefore be an example of the delicate balance to build between different rights.

Several harms are therefore prevented by prohibiting some types of content online: including harms related to discrimination; harms to children; harms related to illlegal activity and to national security threats.

5. What types of control mechanisms would be preferred under the applicable jurisdictions?

As mentioned in answers to questions 2 and 6, several control mechanisms are encouraged for online platforms to have. As defined in law n°2004-575, online service providers have to put in place a reporting mechanism to be made aware of illegal content, and are held liable if they were aware of the content and did not remove it.

Regulation 2021/784 requires for online service providers to publish a yearly transparency report, which involves reporting on mechanisms put in place to prevent the spread of terrorist content, including automation (which was specifically mentioned in Article 7). As referred to in answer 4, some NGOs have argued that the short timeframe given to platforms to delete terrorist content under Regulation 2021/784 encourages the use of AI to sort through content. AI may therefore be another (unofficial) control mechanism for prohibited content.

6. Is there a legal restriction on the distribution of data for national security reasons?

The issue is whether there is a legal restriction on the distribution of data for national security reasons in France. There are specific provisions in France which restrict content for national security reasons, and specifically to fight against the threat of terrorism. What to do with data related to that content is also defined.

The European Parliament adopted its *Regulation 2021/784 on the dissemination of terrorist content online* on April 29, 2021.
(https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R0784)
The Regulation gives members states the power to issue removal orders for terrorist content on the internet (Article 3). This sets an obligation for hosting service providers to remove terrorist content or disable access to it, within one hour following the order's issuance (Article 3). The scope of removal orders is extensive, and there even exists a procedure for cross-border removal orders within the EU member states (Article 4).

There does not seem to be a condition concerning the size of the providers subject to the obligation: it could therefore include large social media platforms such as Facebook and Twitter, search engines such as Google, as well as much smaller websites. A systematic failure to comply with the obligations set by article 3 could be punished by a financial penalty, up to 4% of the provider's global turnover for the preceding year (article 18). The amount of the financial penalty shows the importance given to this issue: for a company such as Google, 4% of turnover would easily be over a billion dollars.

This Regulation also includes measures designed to prevent circulation of terrorist content, such as a yearly transparency report to be published by providers, explaining the measures they take to address the dissemination of such content (Article 7).

Data of removed terrorist content has to be preserved by providers (with safeguards surrounding its access), if it is necessary for judicial proceedings and complaints, or for the prevention and investigation of terrorism. It should be preserved for six months after the removal order, or longer upon request by a relevant authority (Article 6).

However, there is an exception concerning educational and research purposes. This Regulation does not consider as terrorist content, material which was disseminated for educational, journalistic, artistic or research purposes, as well as material preventing terrorism — including material representing controversial views (Article 1-3).

BELGIUM

Michiel De Wolf

In Belgium, the hierarchy of laws is as follows: at the very top of the pyramid are international law and EU law. These will always take precedence over national Belgian law. So, in case of a conflict, it is the EU law that will prevail and the national provision or contrary national interpretation will have to be set aside.
Moving on to the internal Belgian order, the Belgian Constitution is the most important legal norm. It notably enumerates several fundamental rights. The Constitutional Court reviews federal and federate laws for compliance with those fundamental rights.
Then, the hierarchy is split between the federal and the federate level.

- On the federal level, the order of precedence is as follows: special laws and ordinary laws (this is the "residual" category in which the large majority of legislative measures fall), both of which are adopted by the legislator, and royal and ministerial decrees adopted by the executive branch to give effect to the laws adopted by the legislator.
- On the federate level, both the communities and the regions share competences. At this level, the legislator adopts decrees and ordinances, while the executive branch adopts governmental decrees.

At the very bottom of this scheme are the rules adopted by provinces and municipalities.
Case law does not follow the rule of precedent, and decisions by courts and tribunals are only binding on the parties to the dispute. The sole exception to this general rule are the judgments by the Constitutional Court. Judgments by the Supreme Court nevertheless enjoy strong persuasive force.

Introduction to the Belgian legal framework on AI / NLP / data. It is important to note that most of the Belgian rules in the domain are made at the level of the European Union, either in the form of directly applicable Regulations or Directives to be transposed in the Belgian legal order by the competent government body. There is a notable absence of purely "Belgian-made" regulation in this domain. Finally, Belgium is one of 14 European Union member states that have published a White Paper in favor of soft-law regulation of AI in response to the Commission's Proposal for a European AI Regulation. These countries fear that too stringent regulation may hinder the development of the European AI space.

A. IP Questions

In the Belgian legal system, intellectual property rights are dealt with under Title XI of the Economic Law Code (ELC).

1. Are the data training sets and models protected by IP rights, and if so, which IP rights?

Article XI.186 ELC states that datasets that by reason of the choice of ranking of their subject matter form an intellectual creation of their author, are covered by **copyright**. Such copyright is independent of the copyright that applies to the data, components or underlying works in the dataset. Note however the requirement of intervention by the author, demonstrating a sufficient degree of originality. At the current state of the law, datasets produced through machine learning cannot be protected by copyright.

Datasets are also protected by a ***sui generis* database right**. *See*: Article XI.307 ELC. The producer of a database shall be entitled to prohibit the use of the database or prohibit the copying of the contents of the database. Thus, without the permission of the producer of the database, it is not allowed to extract a substantial part of the contents of a website with the intention of placing those contents on one's own website. Without such permission, it is also prohibited to distribute copies of the database by renting it out or making it available on the Internet. The extraction and re-utilization of insubstantial parts (i.e. negligible parts which in themselves do not require significant investment by the producer) do not fall within the scope of protection of the sui generis right.

Attention: the extraction and re-utilization of insubstantial parts does constitute an infringement (i) if it is done in a repeated and systematic manner and (ii) if it conflicts with a normal exploitation of that database or (iii) if it unreasonably prejudices the legitimate interests of the maker of the database. This is the case when the repeated retrievals make it possible to compile a similar database.

2. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?

Article XI.191/2 ELC: the author of a lawfully published dataset cannot oppose the reproduction thereof for scientific purposes insofar such purposes are not for monetary gain and do not impair the normal exploitation of such dataset. Reproduction for these purposes requires reference to the source and name of the dataset's author.

Other exceptions to dataset's authors' copyrights are citation in the context of education or scientific research; or in the context of pedagogical activities (Article XI.191/2, 1°, 2° and 6° ELC).

Outside of the limited exceptions, the author of the published dataset on which the pre-trained language model is trained will be able to assert their copyright. Licensing will be needed to legally use such datasets.

As far as the *sui generis* data right is concerned, once the dataset has been lawfully made public, other users can without the consent of the producer use such dataset for private use (which is of limited use in the context of large language models), for educational or scientific purposes, or to guarantee public security or in the context of administrative or judicial proceedings.

3. Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g. HTML tags or document structure)?

Based on the definition of "database" provided in Article I.13 ELC, there is no distinction made between datasets based on formatting, structure or HTML tags. Dataset means: a collection of works, data, or other independent elements, arranged systematically or methodically, and accessible individually by electronic means or otherwise.

4. What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?

Intentionally Omitted.

5. Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR/Pile) (for example, such a dataset may contain information like: in the nth entry of C4, there is a "<b>" html tag after the mth character)?

Intentionally Omitted.

6. What are the IP risks related to data collection directly from persons? For example when you interview people or they donate data etc.

In this case, there would not be any IP-related risks. The issue faced here is related to data protection of the persons. In the EU, the General Data Protection Regulation (GDPR) regulates the collection of personal data per this question. In short,such persons should give their consent for the collection and use of the data given, and this consent will only be deemed valid if the purpose of the use is clear from the beginning on.

7. Can NLP researchers train a language model on borrowed library books - e.g., from the Internet Archive or other online book lender?

The fact that a work is borrowed through a library service does not affect the copyright protections attached thereto. NLP researchers wanting to use a copyrighted book would either need to obtain approval from the rightsholder, or find themselves in one of the copyright exceptions. Think about scientific research purposes, or private use. (Subsection 1 of Section 6 of Chapter 2 of Title 5 ELC of Book XI ELC).

8. How do rights on the source data (e.g. copyright) transfer to the trained model?
a. Will this depend on where training occurs?
b. Will this depend on where data is gathered?

Intentionally Omitted.

B. Licensing questions

1. Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?

- The rights to a database (*sui generis*) are licensable. There are no special requirements for such licensing other than the typical validity conditions for contracts under general civil law. Article XI.308 ELC.

- To the extent that copyright applies, licensing is subject to more stringent conditions, such as exploitation of the work in accordance with professional practices. Article XI.167 ELC.

2. What are the relationships between laws, licenses, and terms of use? Which is more binding?

   This question falls back on general contracts law. As licenses are contracts between private parties, they cannot go against the law. Nevertheless, within the bounds of the law parties are free to contract as they bargain. Terms of use of a license can only bind a party when such party effectively knew of the terms of use, and accepted those .

3. What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?

   Please refer to the answer in question #1. In general, licenses are contracts. They are validly construed if: (i) each party gives free and informed consent; (i) each party has the capacity to contract (*e.g.*, no minors); (iii) the object of the contract is determined or determinable and lawful; and (iv) a legitimate cause. These conditions are evaluated for accurateness/respect at the moment of conclusion of the license.

   In order to determine whether such a license actually applies, researchers should look at the scope of the license and possible exclusions contained therein.

4. What about if the users download or copy their own data and then provide it to NLP researchers directly?

   In such a case, no license question arises. Rather, the question is situated in the domain of privacy and data protection. The users' personal data are theirs. NLP researchers have to make sure that the users consent to the use and processing of their personal data.

5. Does the license that the dataset is shared under override the terms and conditions?

   Intentionally Omitted.

6. Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?

   The key element of the answer to the first prong of this question is legal access: NLP researchers that have lawful access to licensed/copyrighted data (or otherwise protected as intellectual property) can use such data to train their language model.

7. Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?

   Intentionally Omitted.

C. Text data mining and fair use questions

1. What are types of legally permitted text data mining?

Following Article 3 of European Union Directive 2019/790, the Belgian legislator has to provide an exception to copyright for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access. Lawful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or access to content that is freely available online. This is a strict limit.
Apart from this exception for text and data mining for scientific research, a more limited exception allowing text data mining exists in all cases where the works are lawfully accessible and the holder of the copyright has not made an express reservation against data mining (Article 4 of the Directive).

All text and data mining of copyright protected works outside the scope of these two exceptions is unlawful.

Note that this is the future law in Belgium. The Directive had to be transposed into national law already at the time of writing of this playbook, but Belgium failed to do so timely. The preparatory works for the transposition are being undertaken. These exceptions – created by the EU legislator – are meant to fill the current vacuum in the law.

Note finally that in contrast to the US, there is no “fair use” principle in the EU.

2. Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question?

See: the answer in question #1 here above. This would precisely be the scientific research copyright exception. Note that even for this exception lawful access is required, meaning either through an open access policy (freely available content) or appropriate contractual arrangements.

3. If NLP researchers can crawl data ourselves, what are the limits regarding using this data for training a language model? Can NLP researchers redistribute the data afterwards? If so, under which license and in which geographical regions?

Intentionally Omitted.

4. If NLP researchers use parts of the Pile, Common Crawl, OSCAR or C4, can NLP researchers redistribute the data later? If yes under which conditions and in which countries?

Intentionally Omitted.

5. What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify? The edition conversations of Wikipedia?

Intentionally Omitted.

6. For example, do the terms and conditions of Twitter, Facebook, Youtube, etc., tell us whether NLP researchers can collect data from them for a project such as BigScience?

Intentionally Omitted.

7. What are some of the risks raised by collecting data from these social media directly?

Information crawled from the social media themselves by researchers can (or most likely will) be personal data in the sense of GDPR.

8. What changes if NLP researchers get direct consent from the users concerned?

Again, this is situated more in the field of data protection and privacy rather than data mining itself. If the social media user whose data will be used in the research undertaking gives consent to use them, that is allowed. Note however that the various conditions described in Question 1 of the Privacy questions here below need to be respected.

9. Does the consent override the Terms of Use?
10. Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can't be scrapped?

D. Privacy questions

1. Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?

i. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of common crawl, a published dataset based on a crawl of the web in terms of PII?

In Belgium, data protection is governed by GDPR and the Implementing Law (*See* the following useful overview: Belgium - National GDPR Implementation Overview | Guidance Note | DataGuidance).

GDPR does not use the concept of PII, but uses the much broader concept of Personal data. The Regulation applies to the data of all EU citizens, independent of the jurisdiction from which the data are being collected/processed in the sense of the Regulation.

In short, data subjects need to consent to the use of their Personal data. Such consent is not a blank check. It needs to be specific, in the sense that the Personal data can only be used for certain purposes that the data subject agreed with (the purpose limitation).

Nonetheless, the Regulation contains an exemption for scientific research purposes. Scientific research on the data is considered not to be incompatible with the initial purposes that the data subject consented to. Researchers wishing to rely on this exemption should include the following information in its record of processing:

- a justification for the non-use of pseudonymised data;
- the reasons why the exercise of data subject rights is likely to seriously impair or render impossible the pursued purposes; and
- the DPIA (data protection impact assessment).

Furthermore, in addition to what is required under Article 13 of the GDPR (*i.e.*, information to be provided where personal data are collected from the data subject), inform the data subject as to whether the personal data are anonymised or not, and the reasons why the exercise of the data subject rights is likely to seriously impair or render impossible the achieved purposes.

Where a controller processes personal data for scientific or historical research purposes which were not obtained directly from the data subjects, the controller must enter into an agreement with the original controller, unless an exception applies. This agreement must contain the details of both controllers and the reasons why the exercise of the data subject rights is likely to seriously impair or render impossible the pursued purposes. The agreement must be added to the record of processing.

Scientific or historical research must be performed on the basis of anonymised data. If it is not possible to achieve the research purpose with anonymised data, then the controller must use pseudonymised data. If it is not possible to achieve the research purpose with pseudonymised data, then the controller may use non-pseudonymised data.

Personal data obtained directly from the data subject must be pseudonymised/anonymised after collection.

In case of further processing for scientific or historical research purposes, the personal data must be pseudonymised/anonymised before initiating further processing or before disclosure to another controller for further processing.

Pseudonymised data may only be de-pseudonymised if necessary for the research and after advice from the DPO.

In case of further processing by another controller, the other controller may not have access to the pseudonymisation keys.

2. Are there concerns around distributing models that have PII stored within it? That can expose such PII?

These cannot be distributed insofar the data contained therein allow for the identification of the individual data subjects. Processing of data needs to happen on an anonymised or pseudonymised basis.

3. What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exceptions for research purposes?

Anonymisation or pseudonymisation is a requirement for valid takedown of personal data, within the context of scientific research (see the answer to question 1).

4. How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?

Consent is the cornerstone of GDPR. Without consent of the data subject, no action can be undertaken. Take also purpose limitations into consideration.

5. What are the privacy risks related to data collection directly from persons? For example when you interview people or they donate data etc.

In this case the same GDPR rules apply as formulated in the answers here above. There needs to be consent and purpose limitation, taking into account the scientific research exemption.

Note furthermore that donation of data is a concept that does not exist under EU (GDPR) law, and such "donation" doesn't change anything to the applicability of the GDPR requirements. Under GDPR data cannot be given "free to use" in a blank check manner. *See* in this regard the following post: Personal data donation: a legal oxymoron beneficial to science? | Timelex.

E. Prohibited content

1. What types of data may be prohibited from being text data mined?

There are no legal rules in Belgium that deal with text data mining (with the exception of the soon-to-be transposed TDM copyright exceptions), consequently no provision in Belgian law specifically deals with text data mining and related prohibitions.

Under Article 22(4) GDPR, in automated decision-making no sensitive data may be used, unless the data subject has given explicit consent to do so or the processing is necessary for reason of public interest. Sensitive data is any data that reveals a person's information.

2. What types of data may be prohibited from being stored or distributed?

Intentionally Omitted.

3. What types of data may be prohibited from being generated?

Intentionally Omitted.

4. What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?

Intentionally Omitted.

5. What types of licensing or other control mechanisms would be preferred under the applicable jurisdictions?

Intentionally Omitted.

6. Is there a legal restriction on the distribution of data for national security reasons? Would a model trained on the data also fall under such restrictions?

Intentionally Omitted.

7. Are there any AI technology, including NLP technology that should not be exported (or imported) by NLP researchers?

Regulation (EU) 2021/821 of the European Parliament and of the Council of 20 May 2021 setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items (recast). Under this Regulation, dual use items – civilian goods and technologies with possible military or security use – can be subject to export control, which entails the need to obtain an authorization from a competent authority. Nonetheless, controls on "technology" transfer do not apply to information "in the public domain" or to "basic scientific research".
See also Recital (13): Various categories of persons can be involved in the export of dual-use items, including natural persons such as service providers, researchers, consultants and persons transmitting dual-use items electronically. It is essential that all such persons are aware of the risks associated with the export and the provision of technical assistance regarding sensitive items. In particular, academic and research institutions face distinct challenges in export control due to, inter alia, their general commitment to the free exchange of ideas, the fact that their research work often involves cutting edge technologies, their organisational structures and the international nature of their scientific exchanges. Member States and the Commission should, where necessary, raise awareness among the academic and research community and provide them with tailored guidance to address those distinct challenges. In alignment with multilateral export control regimes, the implementation of controls should provide, to the extent possible, for a common approach with respect to certain provisions, in particular regarding the academia related de-control notes ‘basic scientific research’ and ‘public domain’.

SWITZERLAND

Florian Fuhrimann

INTRODUCTION

Switzerland has a civil law system. Therefore, the main sources of Swiss law are comprehensive, frequently updated legal codes. The federal constitution is the highest law, followed by federal statutes. While a lot of important aspects of the administration of justice in Switzerland are unified on the federal level, there are still various areas of law which are governed by cantonal or even communal statutes and regulations. Further, in many areas of law, codes on the cantonal level supplement the federal codes applicable on the national level. In the present context, it should be noted that in particular in the area of data privacy law, each of the 26 cantons has its own laws and regulations. However, the deviations between the cantons are usually not very significant and mainly concern procedural issues.

Case law is a secondary source for Swiss law; the stare decisis principle in adjudication is not adapted in Switzerland. In deciding any given legal issue, precedents serve merely a persuasive role. Still, courts are expected to take past decisions into account when there is a sufficient level of consistency in case law. This holds particularly true for judgments of Switzerland's highest court, the Federal Supreme Court, which inter alia hears direct appeals from the Federal Criminal Court, the Federal Patent Court, the Federal Criminal Court (the court of first instance for certain federal offenses), and the Federal Administrative Court. It also hears appeals from the Cantonal Courts of Appeal, and some special cantonal courts.Despite its geographical location in the heart of Europe, Switzerland is not a member state of the European Union ("EU"). Hence, EU law is not directly applicable in Switzerland. Yet still, the indirect influences (e.g., via bilateral treaties) of EU law even on purely domestic issues are manifold. Further, many Swiss statutes are voluntarily adapted to, or at least inspired by, the respective regulations applicable in the EU.

A. IP Questions

1. Are the data training sets and models protected by IP rights and if so which IP rights?

The issue is what types of text and data mining are permitted under French and EU law. The answer mainly depends on the particular nature of the training sets and models used.

In general, Swiss law foresees a *numerus clausus* of IP rights (*i.e.*, there cannot be any other IP rights than the ones enshrined in the relevant Swiss statutes, which stipulate the protection for trademarks, patents, designs, and the general copyright). As regards data training sets and models, only the general copyright and the protection of patents are relevant.

Copyright mainly protects the authors of literary and artistic works (Art. 1 of the Federal Act on Copyright and Related Rights [CopA]). It is the way in which an idea is expressed that is protected, not the idea or concept itself. Copyright protection therefore applies to the form of the work and not its content. Pursuant to Art. 2(3) CopA, computer programs also constitute "works" in the sense of Art. 1 CopA. In particular, the source code of computer programs is protected by copyright. However, algorithms which form the basis of software, for example, are excluded from copyright protection.[1] Moreover, unlike the legal framework applicable in the EU[2], Swiss law

does not foresee specific or *sui generis* copyright protection of databases.[3] While Art. 4 CopA does stipulate a specific protection for collected works (*i.e.,* intellectual creations with individual character with regard to their selection and arrangement), this notion hardly fits to data training sets.[4] Copyright pursuant to Swiss law therefore provides only limited protection for training data sets, since only those data which, taken in isolation, constitute works within the meaning of Art. 2 of the CopA, such as creative texts, images, videos, or music, will be protected under this title. In Switzerland, copyright protection expires after 50 years for computer programs and after 70 years for most other works . Copyright protection begins automatically from when a work is created (see Art. 29 CopA). There are no formalities or registration required. There is no copyright register in Switzerland.

For the limited data that is protected, it might be possible that users wishing to use such data to train their model can avail themselves of an exception provided for by the law if they do not already own rights to these data, e.g., (i) temporary reproduction (Art. 24a CopA), or (ii) use of works for scientific research purposes (Art. 24d CopA).[5]

Patents, just as in most countries of the world, are IP rights for technical inventions. They allow the inventors to prevent others from using their invention for commercial purposes for up to 20 years. Inventors of the patent can decide who is allowed to produce, sell or import the invention in those countries in which they own a valid patent. They can also trade their patent, *e.g.,* sell it or licence the use of the invention.

Swiss legal scholars have discussed the question of whether algorithms can be considered inventions in the sense of Art. 1 of the Federal Act on Patents for Inventions (PatA). Pursuant to the PatA, an invention uses technology to solve a specific problem. The technical features of an invention have a function through which the problem – the purpose of the invention – is solved. The technical character necessary for patenting requires that the laws of nature are used to achieve the objective. An invention is also known as "a technical teaching". In light of this definition, and in particular because of the alleged lack of technical character in the sense of the PatA, legal scholars generally deny that algorithms can be protected by patents.

Lastly, it has been considered by legal scholars that the training data sets and algorithms might also be protected as trade secrets.[6] Indeed, there are rulings of the Supreme Court of Germany and the European Court of Justice in which algorithms have been classified as trade secrets.[7] However, there seem to be no similar judgments issued by Swiss courts. In any event, the protection of trade secrets, at least from a purely legal point of view, would not be an intellectual property right.

2. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?

The issue is whether the published dataset on which the pre-trained language model is trained on is protected by copyright or other IP rights. As outlined in question 1. above, the answer to this question will mainly depend on whether the dataset at hand constitutes as a “work” in the sense

of the CopA. While under Swiss law, there is no *sui generis* protection of datasets, it might well be that the individualized data constitutes as a "work" in the sense of the CopA.

Even if a copyright protection existed, the protection of the data would not be absolute. Indeed, it might be possible that users wishing to use such protected data to train their model can avail themselves of an exception provided for by the law if they do not already own rights to these data, e.g., (i) temporary reproduction (Art. 24a CopA), or (ii) use of works for scientific research purposes (Art. 24d CopA).[8]

3. Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g. HTML tags or document structure)?

From a Swiss legal perspective, there is no relevant difference between publishing datasets that contain only plain text to publishing datasets with additional information, like HTML tags or document structure. Again, the main issue is whether the dataset or the individualized data standing on its own is protected by copyright. To profit of such protection, the data(set) would need to qualify as "work" in the sense of Art. 2 CopA.

4. What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?

If the metadata extracted from a dataset corresponds to a "work" in the sense of Art. 2 CopA, publishing it could be a violation of Art. 39c CopA. This provision stipulates:

"1 Rights management information on copyright and related rights may not be removed or altered.

2 Electronic information that identifies works and other subject-matter or information about the terms and conditions of use as well as any numbers or codes that represent such information are protected when such information:

a.is affixed to a phonogram, audio-visual fixation or data carrier; or

b. appears in conjunction with the communication of a work or other subject-matter without tangible medium.

3 Works or other subject-matter from which the rights management information concerning copyright and related rights has been removed or altered may not be copied, imported, offered, transferred or otherwise distributed or broadcast, made perceptible or made available in this form.

Hence, if the dataset or the individualized data is a "work" in the sense of Art. 2 CopA (e.g., a picture or a photo), then the extraction and use of the corresponding metadata might be a violation of Art. 39c CopA.[9]

5. Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR/Pile) (for example, such a

dataset may contain information like: in the nth entry of C4, there is a "<b>" html tag after the mth character)?

There are no specific IP related concerns to publishing a dataset that only refers to locations in another dataset with restricted availability. Even if such other dataset were protected by copyright, the scope of the copyright under Swiss law does not reach as far as prohibiting pure references (see Art. 25(1) CopA, explicitly stating that published works may be quoted if the quotation serves as an explanation, a reference or an illustration, and the extent of the quotation is justified for such purpose).

6. What are the IP risks related to data collection directly from persons? For example when you interview people or they donate data etc.

If the data is collected with the explicit consent of the persons, and the persons also consented to the specific use of the data they provided, there are no IP related risks with regard to the consenting person.

However, if no such clear-cut consent is given, collection of data from persons is indeed risky; not mainly from an IP rights perspective, but from the point of view of the applicable general data protection rules in Switzerland.[10]

7. Can NLP researchers train a language model on borrowed library books - e.g., from the Internet Archive or other online book lender?

To answer the question properly, the specific conditions of the rental agreements would need to be reviewed.

In general, “library books” and other books are “works” in the sense of Art. 2 CopA, as elaborated more extensively in question 1 above. Irrespective of whether such works are bought or borrowed, the copyright protecting the work generally remains with the author. Usually, the author will have assigned his copyright to the publisher.

Art. 10(1) CopA stipulates that the author has the exclusive right to decide whether, when and how his work is used. Pursuant to Art. 11(1) CopA, the author has the exclusive right to decide: (i) whether, when and how the work may be altered and (ii) whether, when and how the work may be used to create a derivative work or may be included in a collected work.

8. How do rights on the source data (e.g. copyright) transfer to the trained model?

According to Art. 16(1) CopA, copyright is assignable. There are no *lex specialis* rules on the assignment of copyright. Hence, the general rules of Art. 164 *et seqq.* Swiss Code of Obligations (“CO”) are applicable. Notably, Art. 165 CO stipulates that the assignment is valid only if done in writing.

The assignment of a copyright is to be distinguished from the situation in which a derivative work in the sense of Art. 3(1) CopA is created. Derivative works are intellectual creations with individual character that are based upon pre-existing works, whereby the individual character of the latter remains identifiable. Derivative works are protected as works in their own right, as stated in Art.

3(3) CopA. Hence, the copyright of the source work is not transferred, but a new copyright is created.

As regards other rights on the source data, any transfer or assignment generally has to follow the rules set out in Art. 164 *et seqq.* CO.

a) ***Will this depend on where training occurs?***

There is no apparent connection to the place where the training occurs.

b) ***Will this depend on where data is gathered?***

There is no apparent connection to the place where the data is gathered.

B. Licensing Questions

1. Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?

If the data sets or models are protected by copyright, they are generally licensable (see Art. 16 CopA). The Swiss Code of Obligations does not contain any specific rules on license agreements, as this is a so-called *sui generis* innominate contract under Swiss law. Often, analogous rules of the tenancy law are applied by courts when interpreting license agreements.

2. What are the relationships between laws, licenses, and terms of use? Which is more binding?

Generally speaking, (mandatory) laws precede the license agreements. Indeed, the terms of a license and general terms of use must comply with mandatory applicable laws and any non-compliant rule will either be inapplicable or render the entire license or set of terms null. Regarding terms of use, it is important that they have been accepted to be considered a legally binding, unilateral contract to which the user has consented to.

3. What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?

The validity of the license agreement follows the general rules of contract conclusion pursuant to the General Part of the Swiss Code of Obligations. There is no form requirement. It is impossible to provide general guidelines on whether a license or specific term of a license applies under Swiss law. Rather, such assessment must be made in each individual case.

4. What about if the users download or copy their own data and then provide it to NLP researchers directly?

The answer to this question mainly depends on what is meant by providing personal data to NLP researchers. If providing is intended to mean donate, such donation of personal data is unlikely to be valid, despite the lack of express provisions on the matter. Personal data is intrinsically

attached to individuals and is protected through various fundamental rights, most likely making it unalienable. If providing is simply understood as putting at the disposal of the researchers, then this would more likely constitute a privacy-related issue. For an overview on these questions, please refer to Sect. D. below.

5. Does the license that the dataset is shared under override the terms and conditions?

Both licenses that have been agreed upon and terms of use that have been accepted constitute legally binding contracts between the parties to it. Thus, they have the same legal standing and the overriding nature of one over the other will be a case-by-case issue that will depend on various factors such as the date each license and term has been entered into, the compatibility of their respective clauses, etc.

6. Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?

This depends on the specific rights that have been assigned to the NLP researchers. If these rights include the use of the licensed/copyrighted data to train a language model, the NLP researchers may do so. To determine whether such rights are included, the specific contractual provisions need to be analyzed.

7. Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?

If the data gathered is copyright protected, it can only be released by means of a royalty free license, if such right was first assigned to the NLP researchers. If no such assignment had taken place, it is the exclusive right of the original author of the copyright protected work to release it by means of a royalty free license.

C. Text Data Mining and Fair Use Questions

1. What are types of legally permitted text data mining?

The Text Data Mining (“TDM”) process normally starts with the creation of a data set on which the TDM software is then utilized for the purpose of analyzing the data set. In this process, data are regularly copied and stored. If the data in question involve works as defined by copyright law (e.g. scientific or technical texts, images, audio-visual media or collections of data; see Sect. A. above), the TDM process involves acts of reproduction of copyrighted works, e.g. through permanent storage of the works in the data set, through reformatting of the works contained in the data set or through their temporary storage during the analysis. Without either the author's consent or statutory exceptions, such reproduction acts constitute copyright infringements.

In 2020, the CopA was revised. Prior to the revision, the CopA did not contain an exception to the author's exclusive right in favour of scientific research or specific provisions on TDM. Granted,

acts of reproduction performed in the context of TDM and relevant to copyright law were sometimes privileged by the existing exceptions for personal use (for private and internal company purposes), as well as for temporary reproductions, and were thus permissible without the rightsholders' consent. However, the consent of all affected rightsholders was required for all other uses of works or acts of reproduction for other purposes, such as the permanent storage of the data set. This always entailed a great deal of effort, high costs, difficulties in assessing the steps of the technical TDM process from a copyright standpoint and other legal uncertainties.

The principal goal of the comprehensive, newly introduced exception to copyright protection in the context of scientific research was to remove the aforementioned legal uncertainties existing in terms of the permissibility of TDM (and similar technologies) for research purposes under copyright law. The newly incorporated Art. 24d CopA now expressly allows technical acts of reproductions for the purpose of scientific research, provided that the reproduced works have been lawfully accessed, as well as the storage of such reproductions for archiving and backup purposes once the research is complete. Consequently, such technical acts of reproduction are now lawful and royalty free even without the author's consent. However, the reproduction of computer programs is excluded from this exception.[11]

The new provisions cover both basic research and applied research. To avoid classification problems, the new statutory exception applies to both non-commercial and commercial research. It remains uncertain, however, if TDM processes for purposes other than scientific research are permissible under Swiss copyright law.

These statutory exceptions are similar to those set out on the EU level through the Copyright Directive 2019/790.

2. Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question?

As outlined more extensively in the question just above, Swiss copyright law allows researchers to crawl data from the internet themselves. Of course, this exception for the use of copyright protected works for research purposes is only relevant under Swiss law. It has no extraterritorial effects.

3. Intentionally Omitted

4. Intentionally Omitted

5. What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify? The edition conversations of Wikipedia?

Scraping social media content will most likely lead to data privacy issues. For an overview on the applicable data privacy rules under Swiss law, see Sect. D. below.

6. For example, do the terms and conditions of Twitter, Facebook, Youtube, etc., tell us whether NLP researchers can collect data from them for a project such as BigScience?

The TAC of the big social media companies cannot override the mandatory Swiss law rules on data privacy. However, they can be an additional hurdle if data from users of these social media platforms is scraped.

7. What are some of the risks raised by collecting data from these social media directly?

Again, there are mainly data privacy risks associated with collecting data from social media. There might also be contractual and further risks under the applicable TAC.

8. What changes if NLP researchers get direct consent from the users concerned?

See Sect. D, question 4 for an overview on the legal implications of consent to the use of private data.

9. Does the consent override the Terms of Use?

The terms of use a *per se* not relevant with regard to the direct relationship between the concerned user and the NLP researchers.

10. Are there any meaningful legal differences between different kinds of social media (Facebook/Twitter, content-based social media such as Youtube, forums such as Reddit/StackExchange) that can or can't be scrapped?

Answering the question would require to analyze the contractual terms and conditions of the social media platforms concerned. It does not seem that this question is particularly related to Swiss law.

D. Privacy Questions

Overview

Swiss data protection law is rooted in the civil law protection of personality rights. The Federal Constitution of the Swiss Confederation ("SFC") provides a constitutional right to privacy. Article 13 SFC protects the right to privacy in personal or family life and in a person's home. Article 28 of the Swiss Civil Code ("Civil Code") and the Federal Act on Data Protection 1992 ("FADP") put this fundamental right to privacy into concrete terms at a statutory level.

In essence, the data processing principles set out in the FADP provide for protection against infringements of personality rights (data privacy) through excessive use of personal data. Article 28 of the Civil Code remains relevant, from a privacy law perspective, where libel, slander, or defamation is the concern. Furthermore, Article 28 of the Civil Code is relevant for the protection of personality rights of legal entities.

In addition to criminal liability governed by the FADP, a number of provisions of the Swiss Criminal Code ('the Criminal Code') are relevant in a data protection and privacy context. These include criminal law protection of a person's reputation against defamation (including libel and slander) and criminal law provisions prohibiting unauthorised recording of private conversations or wiretapping.

Sector-specific data protection and security requirements set out in laws regulating businesses and organisations in certain sectors (including the healthcare, pharmaceutical, energy, telecommunications, and finance), provide more specific requirements applying to the processing of e.g. patient personal data, bank customer data, or smart metre data. Sector-specific provisions typically supersede the provisions of the FADP.

As outlined in the introductory remarks above, the 26 Cantons, the federal states of the Swiss Confederation, have enacted their own data protection acts. These govern the processing of personal data by Cantonal authorities.

On September 25, 2020, the Federal Parliament enacted a revised FADP. The Revised FADP will enter into force in the course of 2022 or at the beginning of 2023. It implements the requirements of the Council of Europe's modernised Convention 108 on the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), and it aligns the FADP with the European Union's General Data Protection Regulation (Regulation (EU) 2016/679) ('GDPR') with the aim of retaining the European Commission's adequacy finding.

1. Are there legal concerns around accessing web content that have PII? What about datasets? What about distributing such datasets, including across borders?

The concept of PII (Personally Identifiable Information) is not generally used in Switzerland. Swiss law mainly differentiates between Personal data and sensitive data:

Personal data:

The FADP defines “personal data” as any information relating to an identified or identifiable person. This includes information that directly identifies a (natural) person (e.g. a full name or picture showing a person's face) and information that allows identification indirectly by reference to additional information (e.g. email address, telephone number, social security number, or customer number). A 'relative' approach to identification applies. Information may qualify as personal data in the hands of one party and as anonymous data in the hands of another party. Identifiability means that the party holding or receiving the information has (or will reasonably likely gain) access to means it will reasonably likely use to identify the (natural) person directly or indirectly. To ascertain whether such identification is reasonably likely, account is taken of the costs of and the amount of time the holder or receiver of the information requires for identification, taking into consideration the technology available to such business, organisation, or natural person. Note that the current FADP also governs the processing of information relating to an identified or identifiable legal entity. The Revised FADP will not apply to processing of information relating to an identified or identifiable legal entity.[12]

Sensitive data:

Under the FADP, the following categories of personal data qualify as 'sensitive:'

· personal data concerning religious, ideological, political, or trade union-related views or activities;

· personal data concerning health, the intimate sphere, or the racial origin of an individual;

· personal data concerning social security measures; and

· personal data concerning administrative or criminal proceedings and sanctions.

These categories of personal data will continue to be considered sensitive under the Revised FADP. The Revised FADP will add two new categories:

· genetic data; and

· biometric data that uniquely identifies an individual.

What kind of processing is allowed under the FADP?

In contrast to the principle of lawfulness of processing on which the GDPR is based, the processing of personal data by businesses, organizations, or natural persons is generally allowed under the FADP. Only public authorities require a legal basis for processing. Private controllers do not need a legal basis for lawful processing of private data.

Legal bases – or rather "justifications" are relevant only as a basis to justify an otherwise unlawful personality rights infringement (see Art. 12 FADP). Normally there is no breach of privacy if the data subject has made the data generally accessible and has not expressly prohibited its processing (Art. 12(3) FADP). A breach of privacy is unlawful unless it is justified by the consent of the injured party, by an overriding private or public interest or by law (Art. 13 FADP). Pursuant to Art. 13(2)(e) FADP, an overriding interest of the person processing the data shall in particular be considered if that person processes personal data for purposes not relating to a specific person, in particular for the **purposes of research**, planning and statistics and publishes the results in such a manner that the data subjects may not be identified.

This concept will remain the same under the Revised FADP (Article 30(2) of the Revised FADP). Personality rights infringements may be justified on grounds of overriding private or public interests, necessity to comply with a legal obligation laid down in Swiss law, or the consent of the data subject.

The following processing principles are key principles and responsibilities of controllers (i.e. persons or companies processing personal data) under the FADP:[13]

· Lawfulness: Businesses or organisations (controllers) may only process personal data that has been collected in accordance with other applicable laws. For example, processing personal data that has been collected through unlawful trespassing or wiretapping would infringe the 'lawfulness' principle. Note that, in contrast to the principle of 'lawfulness of processing' on which

the GDPR is based, the processing of personal data by businesses, organisations, or natural persons is generally allowed under the FADP (see also above Section 5). Only public authorities require a legal basis for processing.

· Fairness (good faith): Controllers may only perform such processing activities as data subjects may reasonably expect. Furthermore, fairness (good faith) means that processing must be performed as described in privacy notices.

· Transparency: Controllers have to convey to data subjects all information necessary in order to ensure transparent data processing. The information needs to enable data subjects to exercise their rights under the FADP. The Revised FADP will set out in more detail the types of information that controllers need to convey to data subjects. At a minimum, controllers will need to inform data subjects about:

· the identity and contact details of the controller;

· the contact details of the DPO (if any);

· the contact details of the Swiss representative (if any);

· the purposes of the processing;

· (if any) the recipients or categories of recipients of the personal data;

· (if the controller intends to transfer personal data internationally) the countries the controller intends to transfer personal data to and (in the absence of an adequacy decision taken by the Federal Council) based on which safeguards (e.g. Standard Contractual Clauses ('SCCs'));

· (if the controller has not obtained the personal data directly from the data subject) the categories of personal data collected and processed; and

· (if any) the existence of automated individual decision-making.

· Purpose limitation: Controllers may only process personal data for the specified purposes that have been notified to or are obvious to data subjects; and may only process personal data in a manner compatible with those purposes. The information about the purposes of the processing needs to be specific. Controllers also need to ensure that further processing of personal data received from other controllers is compatible with the purposes determined and communicated to the data subjects at the time of collection.

· Proportionality: The processing of personal data needs to be proportionate; that is, limited to what is necessary to achieve the specified purposes, considering the type of personal data concerned and the scope and duration of the processing. The data minimisation and storage limitation principles are key aspects of the proportionality principle. This means that controllers need to limit the scope of personal data collected and processed to what is necessary for the intended purposes, and to delete personal data once it is no longer needed for the specified purposes.

· Accuracy: Controllers need to ensure they only process personal data that is accurate and kept up to date. They must take all reasonable steps to ensure that personal data that is inaccurate or incomplete, having regard to the purposes for which it is processed, is deleted or rectified.

· Data security (integrity and confidentiality): Both controllers and (under the Revised FADP) processors are under an obligation to ensure an adequate level of data security. They are required to protect the integrity, confidentiality, and availability of personal data by means of adequate technical and organisational security measures. In assessing the appropriate level of security, controllers and processors have to account for the purpose, type, and scope of the data processing, the assessment of potential risks for data subjects, and the state-of-the-art security solutions.

If businesses and organisations process personal data in accordance with the processing principles set out above, the processing will generally be considered lawful as long as the data subject has not expressly objected to the processing. Infringements of these processing principles (e.g. processing for further purposes than those initially specified, or processing for longer than necessary for the specified purposes), or continued processing despite the data subject's objection, are breaches of personality rights of the affected data subject. In addition, disclosure of sensitive personal data to third parties without a valid ground for justification is deemed a breach of personality rights.

Breaches of personality rights are deemed unlawful unless the controller can demonstrate that the relevant data processing is justified on grounds of overriding private or public interests or the necessity for its compliance with legal obligations laid down in Swiss law. See section 5 above.

These general principles of processing personal data also apply to accessing and distributing datasets that contain personal data.

2. Are there concerns around distributing models that have PII stored within it? That can expose such PII?

Yes, distributing models that have personal data stored within them, in particular when there is a risk that these personal data could be exposed, must follow the principles set out in question 1 above.

3. What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exception for research purposes?

NLP researchers need to implement mechanisms that ensure the compliance with the general principles of personal data processing as set out above in question D.1.. BReaches of privacy can be justified by overriding public or private interests (see question 1 above). Pursuant to Art. 13(2)(e) FADP, an overriding interest of the person processing the data shall in particular be considered if that person processes personal data for purposes not relating to a specific

person, in particular for the **purposes of research**, planning and statistics and publishes the results in such a manner that the data subjects may not be identified.

4. How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?

Consent is valid as a ground for justification of personality rights infringements only if it is informed and freely given. If the controller seeks to justify the disclosure of sensitive personal data or so-called 'personality profiles' to third parties (other controllers) or if it otherwise seeks to justify a personality rights infringement (e.g. processing for further purposes or for longer than necessary) concerning sensitive personal data, consent needs to be given expressly (clear affirmative action).

Under the Revised FADP, consent will be valid only if it is informed, freely given and specific to one or several processing activities (Article 6(6) of the Revised FADP). Further, if a controller seeks to justify personality rights infringements involving sensitive personal data or high-risk profiling, consent needs to be expressly given (Article 6(7) of the Revised FADP).

Despite the lengthy Parliamentary debate and the misconceptions surrounding it, there will be no general requirements under the Revised FADP to obtain consent for so-called high-risk profiling (a concept introduced late in the Parliamentary debate; meaning profiling that poses a high risk to the privacy of individuals by pairing between data that enables an assessment of essential aspects of the personally of a natural person). Rather, consent or another valid ground for justification (such as overriding private or public interests or a legal obligation) will only be required to justify high-risk profiling that does not comply with the fair processing principles. See section 6 for details.

E. Prohibited content

1. What types of data may be prohibited from being text data mined?

The issue is whether certain types of data are prohibited from being text data mined under Swiss law.

Generally speaking, there are no specific legal provisions on which data is prohibited from being text data mined. However, general Swiss criminal law applies also to text data mining.

Preliminarily, it should be noted that felonies or misdemeanours under Swiss law can also be “committed” by AI. In fact, offences committed using robots can be prosecuted like any other crime committed by a person using an object. Swiss criminal law requires the personal culpability of the offender. If an AI robot or system commits a criminal act, it cannot be criminally liable under the current and traditional Swiss criminal law doctrine. The same is true if AI causes someone to commit a crime Therefore, attribution of the criminal act to the creator/programmer or the user of the AI robot or system should be considered. If an AI robot or system was intentionally

programmed to commit a criminal act, the creator or programmer is criminally liable. If it was programmed correctly but intentionally used in a way that resulted in the committing of a criminal act, the user is criminally liable. The creator/programmer as well as the user can only be punished for the negligent commission of a criminal offence if negligence is also explicitly punishable for such criminal offence. Under Art. 102 of the Swiss Criminal Code ("SCC"), it is even possible to assign criminal liability to a corporation if the activity cannot be attributed to a natural person, and if the criminal offence was committed in the exercise of commercial activities in accordance with the object of the undertaking. The undertaking can be fined up to CHF 5 million for such liability. If AI commits a felony or misdemeanour and the requirements mentioned above are met, the corporation using the AI can be held liable.

In the context of the present question, the following provisions of the SCC are relevant, as they effectively also prohibit certain data from being text data mined:

Obtaining personal data without authorisation (Art. 179^{novies} SCC): Any person who without authorization obtains from a data collection personal data or personality profiles that are particularly sensitive and that are not freely accessible shall be liable on complaint to a custodial sentence not exceeding three years or to a monetary penalty.

Pornography (Art. 197 SCC):

(1) Any person who offers, shows, passes on or makes accessible to a person under the age of 16 pornographic documents, sound or visual recordings, depictions or other items of a similar nature or pornographic performances, or broadcasts any of the same on radio or television is liable to a custodial sentence not exceeding three years or to a monetary penalty.

(4) Any person who produces, imports, stores, markets, advertises, exhibits, offers, shows, passes on or makes accessible to others, acquires, or procures or possesses via electronic media or otherwise items or performances as described in paragraph 1 above that contain sexual acts involving animals, acts of violence involving adults or non-genuine sexual acts with minors is liable to a custodial sentence not exceeding three years or to a monetary penalty. If the items or performances contain genuine sexual acts with minors, the penalty is a custodial sentence not exceeding five years or a monetary penalty.

Representations of acts of violence (Art. 135 SCC):

(1) Any person who produces, imports, stores, markets, promotes, exhibits, offers, shows, makes accessible or makes available sound, film or video recordings or other products in which acts of extreme violence against persons or animals are portrayed, without reasonable cultural or scientific grounds therefor, and in doing so seriously offends basic human dignity is liable to a custodial sentence not exceeding three years or to a monetary penalty.

(1bis) Any person who acquires, procures by electronic or any other means, or possesses the recordings or other products mentioned in paragraph 1 above, provided these portray acts of

violence against persons or animals is liable to a custodial sentence not exceeding one year or to a monetary penalty.

2. What types of data may be prohibited from being stored or distributed?

The limitations as outlined above under question E.1. just above apply also to storage and distribution of the respective data.

3. What types of data may be prohibited from being generated?

The limitations as outlined above under question E.1. just above apply also to the creation of the respective data.

4. What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?

Art. 179^{novies} SCC is an offence in breach of privacy. Hence, its purpose is to protect the privacy of people and their freedom to choose with whom they want to share their personal data with.

The provision of Art. 197 SCC (pornography) is listed as an offence against sexual liberty and honor. The paragraphs declaring child pornography to be illegal obviously particularly protect the sexual and physical integrity of children.

Lastly, Art. 135 SCC (representations of acts of violence) is listed in the chapter on offences against life and limb. However, the prohibition of representations of acts of violence is rather an indirect protection of offences against life and limb. The represented acts of violence themselves constitute the direct offences against life and limb.

* * *

[1] See Widmer, *Ada, die Algorithmen und das Immaterialgüterrecht*, AJP 2016, 1273-1274 with further references.

[2] See EU Directive 96/9/CE.

[3] See Gilliéron, *Intelligence artificielle: la titularité de données*, RSDA 2021, 435-449, 440 *et seq.*

[4] Gilliéron, *Intelligence artificielle: la titularité de données*, RSDA 2021, 435-449, 439; see also Gordon/Gurovits, AI, Machine Learning & Big Data Laws and Regulations 2021 | Switzerland, available online: https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/switzerland.

[5] Gilliéron, *Intelligence artificielle: la titularité de données*, RSDA 2021, 435-449, 440.

[6] See Widmer, *Ada, die Algorithmen und das Immaterialgüterrecht*, AJP 2016, 1273-1274, 1274.

[7] BGH, VI ZR 156/13, Jan. 28, 2014, para. 27; ECLI:EU:T:2007:289, para. 151–289.

[8] Gilliéron, *Intelligence artificielle: la titularité de données*, RSDA 2021, 435-449, 440.

[9] See Auf der Maur, SHK Urheberrechtsgesetz, Art. 39c para. 5.

[10] For a concise overview of the revised Data Protection Laws of Switzerland see Rosenthal, *Data Protected - Switzerland*, Oct. 2020 (available online: https://www.linklaters.com/en/insights/data-protected/data-protected---switzerland).

[11] EY Global, *Data Mining under the revised Swiss Copyright Act*, lexology, May 2020; https://www.lexology.com/library/detail.aspx?g=fc874515-d63a-4d05-aa60-fd06c9e19fa6.

[12] Steiner, *Switzerland - Data Protection Overview*, 2021, available online: https://www.dataguidance.com/notes/switzerland-data-protection-overview.

[13] See Steiner, Switzerland - Data Protection Overview, 2021, available online: https://www.dataguidance.com/notes/switzerland-data-protection-overview

SOUTH AFRICA

Giulia Musmeci
Tien Hsin Sheu
Jingyuan Dai
Eugenia Scipioni

A. IP Questions

1. **Are the data training sets and models protected by IP rights and if so which IP rights?**

Issue: What kinds of IP rights would apply to data training sets and models.

Rules: As a member of WTO and Berne Convention, South Africa has amended its IP legislation to conform with the TRIPS agreement. Main IP statutes are (i) Copyright Act 98 of 1978, which covers copyrights on creative works of authorship (latest amendment in 2002), (ii) Patents Act, 1978, which covers patent related regulation, (iii) Designs Act, 1993, which covers the protection of artworks with aesthetic design or functional design. (iv) Protection of Personal Information Act (POPIA), which covers the information privacy protections for the data subjects.

Analysis:

Copyrights: Copyright Act 98 of 1978 states that the "computer programs" are eligible for copyrights (Section 2). If a person writes any source code or compiles any object code from your source code, then that code is automatically afforded copyright protection merely by virtue of writing or compiling your code. Unlike patents or designs, there is no need for you to register for copyright protection. Thus, the data training sets and models are protected by copyright. The copyright protection prevents others from using the same data training sets and models without the explicit or implied authorization from the owner.

However, data training sets and models may be legally used without the authorization in fair dealing situations. Under Copyright Act 1978, fair dealing allows for copying and modification of copyright works in specified situations provided that certain conditions are met. The definition of fair dealing, however, is vague. For example, the 1978 Act states that the extent of copying must be "compatible with fair practice" and must not "exceed the extent justified by the purpose." Thus, a computer program owner's copyright protection of data training sets may yield to fair dealing.

Patents: Pursuant to Patents Act section 25(2), a computer program shall not be an invention for the purposes of patent act; therefore, the computer programs are not proper subject matter for registration as a patent. The data training sets and models are not protected by patent rights.

Design: Designs Act section 14(1) states that the designs eligible for registration are (a) new and original aesthetic design or (b) new and not commonplace in the art in question of functional design. Because programming is not art-oriented design, the Design Act does not apply to data training sets and models.

Conclusion: Data training sets and models are protected by copyrights; however, are excluded from patent and design protection.

2. **Are there legal concerns around publishing a pre-trained language model that is trained on a subset of a published dataset based on a crawl of the web in terms of IP?**

Issue: Is the training of pre-trained language models on the published dataset based on a crawl of the web in terms of IP without individual authorization permitted under Copyright Act? Does the training fall into the category of fair dealing?

Analysis:

Pursuant to Copyright Act 1978, the inclusion of copyrighted content in datasets and its storage require authorization from the copyright holder. This makes web crawled datasets vulnerable to copyright infringement claims since they copy protected content from various websites without individual authorization. The right of the copyright owner is infringed whenever a protected work is the object of unauthorized reproduction, inclusion, and adaptation.

Section 19 of Copyright Amendment Bill (2017) also states that even if a person has a right to access data, the use for the development, production, or marketing of a computer program substantially similar in its expression to the program contemplated in subsection requires another authorization. The data training process may fall into the category of the use of data for "development" and require an additional consent from the data owners for the development purpose under this provision. It is also not compatible with fair practice because it is beyond the original use permitted. Thus, if a person intends to exercise data training with a pre-trained language model, it requires specific authorization from data owners under the Copyright Act. It is worth noting that if a person exercises a pre-trained language model to crawl the data from open source, the person may have to follow the licensing agreement in each data pool for the use of data.

In conclusion, when a person intends to exercise data training with a pre-trained language model trained on a subset of a published dataset based on a crawl of the web in terms of IP, it requires specific authorization from data owners under Copyright Act. Failing to obtain the authorization may be vulnerable to copyright infringement claims.

3. Are there any legal differences between publishing datasets that contain plain text only vs publishing datasets with additional information (e.g., HTML tags or document structure)?

Issue: Whether publishing datasets that contain plain text only is protected by the Copyright Act? Whether adding additional information to the dataset creates a new copyright different from the original one protected under Copyright Act?

Analysis:

Copyright requires the original works of authorship. Pursuant to section 2 (2) of Copyright Act, work shall not be eligible for copyright unless (a) sufficient effort or skill has been expended on making the work to give it a new and original character; and (b) the work has been written down, record or otherwise reduced to material form. Therefore, publishing datasets that contain plain text only is vulnerable to copyright infringement claims due to the lack of original character and sufficient effort within the authorship.

However, once the additional information is added into the datasets (e.g., HTML tags or document structure), the datasets include new content which would be eligible for copyright.

Conclusion: The legal differences between publishing datasets that contain plain text only and publishing datasets with additional information: the former without original character and sufficient effort is more vulnerable to copyright infringement claims while the latter would be eligible for copyright.

4. **What are the legal concerns for publishing metadata extracted from a dataset (e.g. URLs of the documents, publication date, timestamps)?**

Analysis:

(1) One of the concerns may be whether metadata extracted from a dataset has copyright. It relates to the characteristics of work protected by the Copyright Act. It depends on whether the metadata extracted from a dataset has sufficient effort or skill has been expended on making the work to give it a new and original character. If the magnitude of the dataset somehow encompasses the entirety of the original dataset with, it would be vulnerable to copyright infringement claims.

(2) Another legal concern may be whether publishing metadata extracted from a dataset infringes the data owner's copyright. Since the metadata extracted from datasets is trained through crawling online data. Under section 19 of Copyright Amendment Bill (2017), this data development process requires additional authorizations from the data owners.

5. **Are there any legal concerns to publishing a dataset that only refers to locations in another dataset with restricted availability (e.g., C4/OSCAR/Pile) (for example, such a dataset may contain information like: in the nth entry of C4, there is a "<b>" html tag after the mth character)?**

Issue: Whether publishing a dataset that only refers to locations in another dataset with restricted availability violates the copyright of the original dataset owner?

Analysis:

(1) The dataset that only refers to locations in another dataset is similar to an index. Pursuant to section 2 (2) of Copyright Act, work shall not be eligible for copyright unless (a) sufficient effort or skill has been expended on making the work to give it a new and original character; and (b) the work has been written down, record or otherwise reduced to material form. If the new dataset only refers to location with restricted availability to another dataset without further information, it probably lacks sufficient effort or skill in its work and would not be protected by copyright. On the contrary, if the new dataset combines with different locations into its work, it would be protected by copyright.

(2) Referring to locations in another dataset with restricted availability is more like quoting someone else without citation. Whether it violates the copyright of the original dataset depends on the original dataset's licensing agreement and the fair practice within the purpose of publishing a new dataset.

6. **What are the IP risks related to data collection directly from persons? For example, when you interview people, or they donate data etc.**

Issue: Does collecting special personal information involve any legal risk under Protection of Personal Information Act when collection is directly from persons?

Analysis:

(1) Collecting data directly from persons shall abide by the Protection of Personal Information Act (POPIA), especially when it comes to sensitive personal data. Section 26 of POPIA prohibits the processing of special personal information, subject to exceptions provided for in section 27(1). In terms of section 26, a responsible party may not process any of the following special personal information of a data subject: religious beliefs; philosophical

beliefs; race; ethnic origin; trade union membership; political persuasion; health; sex life; biometric information; other criminal behavior of a data subject to the extent that such information relates to the alleged commission by a data subject of any offence; or any proceedings in respect of any offence allegedly committed by a data subject or the disposal of such proceedings.

(2) There are some exceptions to use the special personal data provided for in section 27(1). A data trainer may obtain the consent from a data subject to process the data. Processing for historical, statistical or research purposes to the extent that the purpose serves a public interest, and the processing necessary for the purpose concerned is also permitted. Therefore, when interviewing people or they donate data, a person should not involve these categories of data into training unless obtaining consent with specific purpose for the use of data from the data subject.

(3) Once a person acquires personal information directly from interviewing, it is required, in terms of section 19(1) of POPIA, to secure the integrity and confidentiality of personal information in its possession or under its control by taking appropriate, reasonable technical and organizational measures to prevent loss of damage to or unauthorized destruction of personal information. Therefore, the data training system should establish appropriate safeguards to protect personal information from unlawful or unauthorized access to the data.

7. Can NLP researchers train a language model on borrowed library books - e.g., from the Internet Archive or other online book lender?

Issue: The legal concern from training a language model on borrowed library books. What is the copyright duration for library books?

Analysis:

(1) Borrowed library books are protected by their own copyright. Therefore, training a language model on borrowed library books is vulnerable from copyright infringement claims. However, copyright has a certain duration. NLP researchers may use the materials in borrowed library books after its expiration. Under the 1978 Copyright Act, the lifespan of copyright of computer programs lasts for 50 years after the first copies were made available to the public. The duration of copyright in ‘anonymous or pseudonymous works’ is calculated as 50 years from the date on which the work was made available to the public with the consent of the owner or the end of the year in which it is ‘reasonable to presume’ that the author died. Hence, NLP researchers should check the duration of copyright attached to borrowed library books. The online e-book and the materials in the internet archive are also subject to copyright protection. However, the materials in the open-source archive may be available for data training if the user follows the license agreement of the data pool.

(2) In conclusion, NLP researchers may use the borrowed library books after 50-year of copyright duration expires or use the materials in the open-source archive following the license agreement of the data pool.

B. Licensing Questions

1. **Are data sets licensable? Are models licensable? What restrictions are appropriate or preferable under the jurisdictions?**

There is no fixed list on the types of IP that can be licensed in South Africa, Models, databases and their contents do not have their own specific protection. However, they may qualify for copyright protection (if the requirements for the subsistence of copyright, particularly relating to originality, are met). As such, they can be licensed through either an exclusive or a non-exclusive license agreement.

An exclusive license permits only the licensee and persons authorized by the licensee to exploit the invention and therefore exclude all other persons from doing so, including the patentee.

A non-exclusive license allows the patentee to retain the right to exploit the licensed property as well as the right to grant additional licenses to third parties.

2. **What are the relationships between laws, licenses, and terms of use? Which is more binding?**

Laws are enacted and enforced by the state and everyone must abide by them, while licenses and terms of use are contracts agreed upon by civil subjects such as persons and corporations per autonomy of will. Laws are more binding because violation of law will absolutely result in liability and punishment. However, licenses and terms of use may be deemed as void or voidable if certain provisions violate the law. Detailed definition and illustration are as follows:

- Laws are legal documents that are enacted, revised and promulgated in accordance with legal procedures by the legislative organs with legislative powers, and enforced by the coercive force of the state. In China, laws can be classified into: (1) Constitution (2) laws (3) administrative regulations (4) local regulations (5) autonomous regulations and separate regulations;
- Licenses are agreements where the owner authorizes others to use something within a certain period and scope without changing the ownership. According to the scope of licensing, they can be classified into: (1) proprietary license (2) exclusive license (3) ordinary license;
- Terms of use are agreements between service providers and users regarding the service provided and rights and obligations of both parties. For example, users need to read and consent to the terms of use before they use certain online services such as apps, websites and other e-platforms;

In South Africa, the principle of freedom of contract applies. Therefore, parties are free to negotiate the terms of the license for the whole or part of the bundle of the relevant IP rights, with jurisdictional restrictions and time limits.

A license can include a waiver of moral rights or consent in favor of the licensee, to the reproduction of the copyright works and their communication to the public and the making of adaptations of the copyright works by the licensee, in such a way that would, but for the consent, infringe the author's moral rights.

A license can also place an obligation on the licensor to obtain such a waiver/consent from the author of the works to which the copyright relates. Some limits set out by the law will apply. For example, a licensor cannot license more extensive rights than those they hold.

3. **What makes a valid license? How can NLP researchers determine whether the license attached to a re-published or derived dataset actually applies?**

An exclusive copyright license only has effect if it is in writing and signed by or on behalf of the licensor or an exclusive sub-licensor. Instead, there are no formal requirements for the creation of a non-exclusive license or a sole license. Such a license can be granted orally, in writing, tacitly or can even be inferred from the parties' conduct. However, it is good practice and advantageous to record a license in writing for evidentiary purposes.

When in writing, an IP license should:

- Identify the parties.
- Include a clear definition of the IP which is the subject of the license.
- State whether the license is exclusive, non-exclusive or a sole licensee.
- State whether sub-licensing is permissible.
- State the field of use, especially where the license only relates to specific goods and/or services.
- Determine the range of protected acts that the licensee is authorized to perform in relation to the IP.
- Provide for appropriate quality controls.
- Deal with confidentiality, if applicable.
- Provide for a duration of the license.
- Define the territory of the license.
- Set out the rights and obligations of the licensor and licensee.
- Clearly define what acts will be considered a breach of the licence.
- Provide for the manner in which the licence can be terminated.

4. **What about if the users download or copy their own data and then provide it to NLP researchers directly?**

According to Section 11 of POPIA, personal information may only be processed if there is a lawful justification for the processing, which includes consent from the data subject. In the case of data that are not personal data, the owner of the information is the one who collected it and therefore, it is necessary to refer to the agreement entered into with the person who provided the information. If the destination of the information was exclusive, then the owner will not be able to download the data and deliver it to anyone since there would be a copyright violation, if the information was not collected with exclusive use or through confidentiality agreements, they could be downloaded and shared but only those in which that person is the resource.

5. **Does the license that the dataset is shared under override the terms and conditions?**

As licenses are different from terms and conditions, they do not necessarily override the terms and conditions applicable to a data set. Notably, terms of use are agreements that must be adopted and structured to legally collect, process, use, and share personal data, and they cannot be completely overridden by licenses, although such licenses may include covenants obligating the parties to modify or amend certain provisions of such terms. While licenses are binding agreements to transfer the use of a set of intangible assets over certain work, terms of use and conditions are agreements in which the owner of a digital portal establishes the way in which they will manage the portal and the information collected from the public through it. However, to the extent that licenses can be freely negotiated between the parties, it would

be possible to include a clause on the obligation to apply certain terms of use or to adopt a new set of terms to any person accessing the portal.

Additionally, licenses normally regulate the relationship between the owner of the intangible asset whose usage is being transferred and the person receiving such asset, while the terms of use govern the relationship between those accessing the digital platform and the one providing such platform, regardless of whether the latter is the owner or the licensee. Therefore, the license under which the dataset is shared will coexist with the terms of use applicable to those entering the portal or digital platform.

6. **Can NLP researchers use licensed/copyrighted data to which they legally have access to train a language model? If so, how does this affect the publication and distribution of the model?**

Most likely yes, but not necessarily. Access to licensed or copyrighted material is often limited. The right to use copyrighted or licensed material can be transferred through agreements between the owner of the work and third parties. As mentioned above, these agreements include (1) contracts of assignment of copyrighted material; or (2) license agreements. Those contracts will be the ones determining whether it is possible for NLP researchers to use the licensed or copyrighted material to train a language model.

Through a contract of assignment of copyrighted material the owner of a copyrighted work transfers the economic rights of its creation to a third party Copyright assignments must be in writing and signed by the assignor. To the extent that the norm does not impose any further restrictions on the contract, the parties are free to set the terms of the usage. Moreover, if no limitations are included in the agreement, the person that receives the rights over the work can use it without any limitations. Therefore, if there are no explicit limitations included in the contract, the NLP researchers can legally train a language model with the data.

Furthermore, in this agreement the owner of the copyrighted work transfers the rights over the work and is no longer entitled to exercise such rights unless a limitation is set forth in the contract. An assignment should clearly identify:

- the parties.
- the intention and consensus of the assignor to transfer ownership and the assignee to receive ownership.
- the subject of the assignment.
- the applicable territory for the assignment.
- any consideration payable by the assignee for the assignment.

In addition, it should include warranties and undertakings by the parties and provide for an effective date of the assignment, if different to the date of signature by the parties.

It is also worth mentioning that under South Africa law, the assignment of IPRs by South African citizens and companies to foreign citizens or companies requires approval from the South African Exchange Control Authorities. Provision should be made in an assignment agreement, as a condition precedent, for obtaining the necessary approvals.

License agreements, also called ‘authorizations to use’, allow third parties to access copyrighted material with the authorization of the owner. Contrary to the contract of assignment, where the rights are ceded to the acquiring party, the license agreement only allows third parties to use the copyrighted work for free or in exchange for a fee. Nevertheless, if this latter agreement does not include any particular limitations to the usage of the work,

the NLP researcher could give any reasonable use to such material, such as using the content of it to train new models.

In those cases where the terms of the agreement do not restrict the ways in which the copyrighted work can be used, and assuming that such work meets the requirements of originality and existence in material form and creation by a qualified person, the language model trained by the NLP researcher would be protected by copyright law as a separate work. In consequence, the derivative work created by NLP researchers can be published and distributed without limitation. They will be the owners of all the legal rights attached to such derivative work.

7. **Can data gathered by text data mining be released by means of a royalty free license? What about a model trained on the collected data and the training dataset (i.e. the compilation of the data structured in a specific way)?**

   There can be problems about text data mining when the information used after the mining is protected by copyrights such as books or articles. Since these pieces are protected, they cannot be reproduced or used without authorization and data is not licensable without royalties.

   If the model or data sets are original enough, they can be licensed and no work from another person is being used or if it was authorized to be transferred, therefore it could be licensed under a royalty free license.

C. Text Data Mining and Fair Use Questions

1. **What are types of legally permitted text data mining?**

Copying for text data mining (TDM) is an infringing activity which requires a license or an exception. In South Africa there are no specific exceptions governing the use of text data mining. In fact, in spite of being a member of WTO and Berne Convention, South Africa has amended its IP legislation to conform with the TRIPS agreement, it did not adopt any TDM exceptions and the three-step test contained in Article 13 TRIPS.

It means that TDM software used to process corpora of big data might infringe rights in databases that are protected either by copyright right, thus creating a barrier to TDM. The rule that copyright works reproduced in a big data corpus retain independent copyright protection has not been altered. This means that images, texts, musical works, and other copyright subject-matter contained in a big data corpus are still subject to copyright protection until the expiry of the term of protection.

Perhaps only TDM tools involving minimal copying of a few words or crawling through data and processing each item separately could be operated without running into a potential liability for copyright infringement.

2. Can NLP researchers crawl data from the internet ourselves? If so, are there regional restrictions, for instance, can the crawl be done in the specific jurisdiction in question? If NLP researchers can crawl data ourselves, what are the limits regarding using this data for training a language model?

In South Africa there are no laws or regulations governing web crawling. Therefore, there are no explicit prohibitions specifically applicable to such activity. However, since data

crawling from the internet typically involves downloading contents from websites, if such a process requires collecting, processing, using or sharing personal data, the researcher must comply with the applicable privacy laws.

To the extent that there are no specific regulations explicitly prohibiting web crawling, as long as such activity does not violate privacy, IP, or criminal laws, NLP researchers could legally crawl the web themselves.

Assuming that the applicable laws and regulations to data crawling are privacy and IP laws, there are no regional restrictions applicable. Laws enacted by the President and published in the Government Gazette and the regulations issued by the government regulating such laws are applicable on the entire national territory.

3. What are the risks of data scraping social media content? What counts as social media content? Does the comment section of a news website qualify? The edition conversations of Wikipedia?

When data appear in public sources (i.e. social media posts), collecting them using scraping methods is considered legal in South Africa. However, South Africa does not have a special regime applicable to social media platforms. Therefore, there is no definition of social media content or any provision that allows individuals to exactly determine what can be considered social media content.

Notwithstanding the above, on 1 March 2022, the Films and Publications Amendment Act 11 of 2019 (FPAA), which was signed into law by the President and published in the Government Gazette on 3 October 2019, came into operation. The FPAA amends the Films and Publications Act 65 of 1996 (FPA) and provides more clarity on the regulation of online commercial distributors and the processes that they are required to follow to distribute online content in South Africa. In terms of the FPAA and the regulations, all online distributors will be required to register with, and submit all content to the newly established Film and Publications Board (FPB) for classification.

The FPA also prohibits distribution through any medium including the internet and social media, of any film, game or publication, which amounts to propaganda for war, incites imminent violence or advocates hate speech. The definition of a publication has been widened in the FPAA to include any content made available using the internet, excluding a film or game.

The FPAA provides that any person who knowingly distributes in any medium, including the internet and social media, any such film, game or publication will be guilty of an offence. This includes a possible fine not exceeding ZAR 150 000 and/or imprisonment for a period not exceeding two years.

Finally, South African common law, as codified in the Constitution of the Republic of South Africa, 1996, also protects individual’s right to privacy on a professional and personal basis. Thus, posting on social media what may be considered as personal information, although falling out of the ambit of POPIA, might be still protected by the common law in the event that the post breaches the privacy of the subject.

D. Privacy Questions

1. Are there legal concerns around accessing web content that has personal identifiable information ("PII")? What about datasets?

Rules and definitions from the Protection of Personal Information Act 4 of 2013 ("POPIA"):

Section 1 - Definitions:

"**personal information**" means information relating to an identifiable, living, natural person, and where it is applicable, an identifiable, existing juristic person, including, but not limited to:

(a) information relating to the race, gender, sex, pregnancy, marital status, national, ethnic or social origin, colour, sexual orientation, age, physical or mental health, well-being, disability, religion, conscience, belief, culture, language and birth of the person;

(b) information relating to the education or the medical, financial, criminal or employment history of the person;

(c) any identifying number, symbol, e-mail address, physical address, telephone number, location information, online identifier or other particular assignment to the person;

(d) the biometric information of the person;

(e) the personal opinions, views or preferences of the person;

(f) correspondence sent by the person that is implicitly or explicitly of a private or confidential nature or further correspondence that would reveal the contents of the original correspondence;

(g) the views or opinions of another individual about the person; and

(h) the name of the person if it appears with other personal information relating to the person or if the disclosure of the name itself would reveal information about the person;

"**processing**" means any operation or activity or any set of operations, whether or not by automatic means, concerning personal information, including—

(a) the collection, receipt, recording, organization, collation, storage, updating or modification, retrieval, alteration, consultation or use;

(b) dissemination by means of transmission, distribution or making available in any other form; or

(c) merging, linking, as well as restriction, degradation, erasure or destruction of information.

Section 11 - Consent:

(1) Personal information may only be processed if:

(a) the data subject or a competent person where the data subject is a child consents to the processing;

(b) processing is necessary to carry out actions for the conclusion or performance of a contract to which the data subject is party;

(c) processing complies with an obligation imposed by law on the responsible party;

(d) processing protects a legitimate interest of the data subject;

(e) processing is necessary for the proper performance of a public law duty by a public body; or

(f) processing is necessary for pursuing the legitimate interests of the responsible party or of a third party to whom the information is supplied.

Section 12 - Collection directly from data subject:

(1) Personal information must be collected directly from the data subject, except as otherwise provided for in subsection (2).

(2) It is not necessary to comply with subsection (1) if:

(a) the information is contained in or derived from a public record or has deliberately been made public by the data subject;

(b) the data subject or a competent person where the data subject is a child has consented to the collection of the information from another source;

(c) collection of the information from another source would not prejudice a legitimate interest of the data subject.

Section 13 - Collection for specific purpose:

(1) Personal information must be collected for a specific, explicitly defined and lawful purpose related to a function or activity of the responsible party.

(2) Steps must be taken in accordance with section 18(1) to ensure that the data subject is aware of the purpose of the collection of the information unless the provisions of section 18(4) are applicable.

Section 18 - Notification to data subject when collecting personal information:

(1) If personal information is collected, the responsible party must take reasonably practicable steps to ensure that the data subject is aware of:

A. the information being collected and where the information is not collected from the data subject, the source from which it is collected;
B. the name and address of the responsible party;
C. the purpose for which the information is being collected;
D. whether or not the supply of the information by that data subject is voluntary or mandatory;
E. the consequences of failure to provide the information;
F. any particular law authorising or requiring the collection of the information;
G. the fact that, where applicable, the responsible party intends to transfer the information to a third country or international organisation and the level of protection afforded to the information by that third country or international organisation;
H. any further information such as the:
    1. recipient or category of recipients of the information;
    2. nature or category of the information;
    3. existence of the right of access to and the right to rectify the information collected;
    4. existence of the right to object to the processing of personal information as referred to in section 11(3); and right to lodge a complaint to the Information Regulator and the contact details of the Information Regulator, which is necessary, having regard to the specific circumstances in which the information is or is not to be processed, to enable processing in respect of the data subject to be reasonable.

(2) The steps referred to in subsection (1) must be taken:

A. if the personal information is collected directly from the data subject, before the information is collected, unless the data subject is already aware of the information referred to in that subsection; or
B. in any other case, before the information is collected or as soon as reasonably practicable after it has been collected.

(3) A responsible party that has previously taken the steps referred to in subsection (1) complies with subsection (1) in relation to the subsequent collection from the data subject of the same information or information of the same kind if the purpose of collection of the information remains the same.

(4) It is not necessary for a responsible party to comply with subsection (1) if:

(a) the data subject or a competent person where the data subject is a child has provided consent for the non-compliance;

(b) non-compliance would not prejudice the legitimate interests of the data subject as set out in terms of this Act;

[…]

(f) the information will:

not be used in a form in which the data subject may be identified; or be used for historical, statistical or research purposes.

Section 26 - Prohibition on processing of special personal information:

(1) A responsible party may, subject to section 27, not process personal information concerning:

(a) the religious or philosophical beliefs, race or ethnic origin, trade union membership, political persuasion, health or sex life or biometric information of a data subject; or

(b) the criminal behavior of a data subject to the extent that such information relates to: (i) the alleged commission by a data subject of any offence; or (ii) any proceedings in respect of any offence allegedly committed by a data subject or the disposal of such proceedings.

Section 27 - General authorisation concerning special personal information:

(1) The prohibition on processing personal information, as referred to in section 26, does not apply if the:

(a) processing is carried out with the consent of a data subject referred to in section 26;

[…]

(e) information has deliberately been made public by the data subject.

Importantly, sections 28 – 33 of POPIA include additional authorisations relating to specified categories of information. These sections may provide additional justifications in the event that consent cannot be obtained.

Analysis:

Generally, in order to access web content and datasets that include PII, a lawful justification must exist. POPIA specifies a list of lawful justifications, one of which is consent. Information should be collected directly from the data subject, and they should be notified of a list of things upon collection. There are exceptions to these rules.

POPIA provides a broad definition of PII, which is included in section 1. PII must relate to a living person and accordingly doesn't apply to deceased persons. The definition includes a non-exhaustive list of the types of information that constitute PII, which include things such as contact information, any identifying number, symbol and online identifiers. What should be noticed is that POPIA includes personal opinions, views or preferences as long as they can be related to the data subject. This clause may be broader than regulations in other countries. Moreover, POPIA generally prohibits processing special personal information which is prescribed in section 26, such as religious or philosophical beliefs, race or ethnic origin. However, if data subjects have consented or this information has deliberately been made public by the data subject, the prohibition will be removed.

Second, accessing web contents and datasets constitutes processing under POPIA. For NLP activities, researchers will need large amounts of data to train models, usually including collecting, using, and merging data which falls within the scope of POPIA.

Therefore, accessing web contents and datasets containing PII is governed by POPIA. To process the information lawfully, NLP researchers should:

1. gather consent to process from data subjects;
2. collect directly from data subjects, or from public records such as public webs o datasets, or from another source agreed to by data subjects; and
3. notify data subjects of the source of information, the purpose of the collection and other information specified in section 18.1, unless the information will not be used in a form in which the data subject may be identified or it is used for research purposes.

2. What about distributing such datasets, including across borders?

Rules:

Section 15 - Further processing to be compatible with purpose of collection

Further processing of personal information must be in accordance or compatible with the purpose for which it was collected in terms of section 13.

(1) To assess whether further processing is compatible with the purpose of collection, the responsible party must take account of:

A. the relationship between the purpose of the intended further processing and the purpose for which the information has been collected;
B. the nature of the information concerned;
C. the consequences of the intended further processing for the data subject;
D. the manner in which the information has been collected; and
E. any contractual rights and obligations between the parties.

(2) The further processing of personal information is not incompatible with the purpose of collection if:

A. the data subject or a competent person where the data subject is a child has consented to the further processing of the information;
B. the information is available in or derived from a public record or has deliberately been made public by the data subject;

[…]

E. the information is used for historical, statistical or research purposes and the responsible party ensures that the further processing is carried out solely for such purposes and will not be published in an identifiable form.

Section 72 - Transfers of personal information outside Republic

(1) A responsible party in the Republic may not transfer personal information about a data subject to a third party who is in a foreign country unless:

(a) the third party who is the recipient of the information is subject to a law, binding corporate rules or binding agreement which provide an adequate level of protection that:

a) effectively upholds principles for reasonable processing of the information that are substantially similar to the conditions for the lawful processing of personal information relating to a data subject who is a natural person and, where applicable, a juristic person; and

b) includes provisions, that are substantially similar to this section, relating to the further transfer of personal information from the recipient to third parties who are in a foreign country;

(b) the data subject consents to the transfer;

(c) the transfer is necessary for the performance of a contract between the data subject and the responsible party, or for the implementation of pre-contractual measures taken in response to the data subject's request;

(d) the transfer is necessary for the conclusion or performance of a contract concluded in the interest of the data subject between the responsible party and a third party; or

(e) the transfer is for the benefit of the data subject, and:

- it is not reasonably practicable to obtain the consent of the data subject to that transfer; and
- if it were reasonably practicable to obtain such consent, the data subject would be likely to give it.

(2) For the purpose of this section:

(a) ''binding corporate rules'' means personal information processing policies, within a group of undertakings, which are adhered to by a responsible party or operator within that group of undertakings when transferring personal information to a responsible party or operator within that same group of undertakings in a foreign country; and

(b) ''group of undertakings'' means a controlling undertaking and its controlled undertakings.

Analysis:

POPIA prohibits the further processing of PII, if the further processing is not compatible with the original purpose of the collection. Distributing datasets that contain PII may constitute "further processing" and should be compatible with the initial purpose of collection under section 15.

Distributing datasets is likely to be considered as incompatible with the initial purpose of collection, and requires consent from data subjects as a justification. If the purpose for which the information has been collected is to establish datasets, distributing datasets to a third party is not a common or necessary next step. The relationship between the initial purpose and the intended furthering processing is not close. Also, when people agree to submit their PII for information gathering, they usually do not expect that their PII will be distributed to a third party. The distribution will increase the risk of data leakage. Therefore, NLP researchers should gather consent from data subjects to distribute datasets that contain PII.

Under limited circumstances, consent from data subjects is not required for distributing datasets. First, if the responsible party of the initial collection has notified data subjects that the collection is intended to form a dataset for NLP, and the distribution of datasets is common among NLP researchers, distributing may not be considered as further processing. Therefore, additional consent to distribution is not required. Second, if PII of datasets are collected from public record or solely for research purposes without an identifiable form of publication, there will be no need to assess whether further processing is compatible with the purpose of collection.

Regarding distributing datasets across borders, under section 72, NLP researchers should gather consent from data subjects or transfer to a third country that provides the same degree of protection for PII. The Information Regulator, which is the independent body mandated to oversee and enforce compliance with POPIA, has not yet released a list of countries that it considers as providing an adequate level of protection.

If it involves transferring special personal information, as referred to in section 26, or the personal information of children as referred to in section 34, to a third party in a foreign country that does not provide an adequate level of protection for the processing of personal information as referred to in section 72, the responsible party must obtain prior authorization from the Information Regulator, in terms of section 58, prior to any processing.

3. Are there legal concerns around publishing a pre-trained language model that is trained on a subset of common crawl, a published dataset based on a crawl of the web in terms of PII?

The publishing will raise potential issues under POPIA.

First, the training of the model using public datasets containing PII constitutes “processing.” It should follow the same procedures under the analysis of Question 1, which includes notifying data subjects and responsible parties of the public datasets.

Second, if a party only publishes a pre-trained model without being responsible for its training, what matters is whether the model contained PII. If it contains PII, publishing the model constitutes “distribution” of PII under the definition of “processing” and the party should gather consent from data subjects to publish the model under section 11.

4. Are there concerns around distributing models that have PII stored within it? That can expose such PII?

Distributing models that contain PII constitutes “further processing” under section 15 and should be compatible with the initial purpose of collection.

If data subjects only consent to collect their information to train a model, it is highly likely that distributing the model in a manner that may expose such PII is incompatible with the initial purpose of collection because data subjects may not anticipate the consequences of the intended further processing, such as the potential exposure of their PII to a third party. Therefore, NLP researchers will need consent from data subjects for further distribution.

It will not be necessary to examine the compatibility with the purpose of collection if the PII in the model is derived from a public dataset or other public source under section 15. Distributing models under this circumstance does not need consent from data subjects. However, the training of the model still needs to meet the procedures in the analysis of Question One.

5. What are the mechanisms NLP researchers would have to put in place in each case to ensure the takedown of personal data that is protected by local privacy laws? Any exceptions for research purposes?

NLP researchers are required to take appropriate measures to secure PII. Please refer to the following rules for protective mechanisms:

Section 19 Security measures on integrity and confidentiality of personal information

(1) A responsible party must secure the integrity and confidentiality of personal information in its possession or under its control by taking appropriate, reasonable technical and organisational measures to prevent—

(a) loss of, damage to or unauthorised destruction of personal information; and

(b) unlawful access to or processing of personal information.

(2) In order to give effect to subsection (1), the responsible party must take reasonable measures to:

(a) identify all reasonably foreseeable internal and external risks to personal information in its possession or under its control;

(b) establish and maintain appropriate safeguards against the risks identified;

(c) regularly verify that the safeguards are effectively implemented; and

(d) ensure that the safeguards are continually updated in response to new risks or deficiencies in previously implemented safeguards.

(3) The responsible party must have due regard to generally accepted information security practices and procedures which may apply to it generally or be required in terms of specific industry or professional rules and regulations.

Section 21 - Security measures regarding information processed by operator

(1) A responsible party must, in terms of a written contract between the responsible party and the operator, ensure that the operator which processes personal information for the responsible party establishes and maintains the security measures referred to in section 19.

(2) The operator must notify the responsible party immediately where there are reasonable grounds to believe that the personal information of a data subject has been accessed or acquired by any unauthorized person.

Section 22 - Notification of security compromises

(1) Where there are reasonable grounds to believe that the personal information of a data subject has been accessed or acquired by any unauthorized person, the responsible party must notify:

(a) the Regulator; and

(b) subject to subsection (3), the data subject, unless the identity of such data subject cannot be established.

(2) The notification referred to in subsection (1) must be made as soon as reasonably possible after the discovery of the compromise, taking into account the legitimate needs of law enforcement or any measures reasonably necessary to determine the scope of the compromise and to restore the integrity of the responsible party's information system.

(3) The responsible party may only delay notification of the data subject if a public body responsible for the prevention, detection or investigation of offences or the Regulator determines that notification will impede a criminal investigation by the public body concerned.

(4) The notification to a data subject referred to in subsection (1) must be in writing and communicated to the data subject in at least one of the following ways:

(a) Mailed to the data subject's last known physical or postal address;

(b) sent by e-mail to the data subject's last known e-mail address;

(c) placed in a prominent position on the website of the responsible party;

(d) published in the news media; or

(e) as may be directed by the Regulator.

(5) The notification referred to in subsection (1) must provide sufficient information to allow the data subject to take protective measures against the potential consequences of the compromise, including—

(a) a description of the possible consequences of the security compromise;

(b) a description of the measures that the responsible party intends to take or has taken to address the security compromise;

(c) a recommendation with regard to the measures to be taken by the data subject to mitigate the possible adverse effects of the security compromise; and

(d) if known to the responsible party, the identity of the unauthorized person who may have accessed or acquired the personal information.

(6) The Regulator may direct a responsible party to publicize, in any manner specified, the fact of any compromise to the integrity or confidentiality of personal information, if the Regulator has reasonable grounds to believe that such publicity would protect a data subject who may be affected by the compromise.

There is no exemption regarding the protective mechanisms for research purposes. However, under section 37, NLP researchers may apply to the Information Regulator for an exemption to process information, even if it is not compliant with the requirements under POPIA. The exemption is discretionary, but the Information Regulator may grant it if the public interest in the processing outweighs any interference with the privacy of the data subject. The NLP activities solely for research purposes may satisfy this public interest requirement and receive an exemption from the Information Regulator concerning retention periods (section 14), further processing (section 15), notification for data subjects (section 18), prohibitions on processing special PII (section 27) and other requirements for processing of PII under POPIA.

6. How does consent of the individuals affect what an NLP organization can and cannot do under each of these regulations?

Generally, without consent from data subjects, an NLP organization cannot process PII. The NLP organization must have a lawful justification to process PII. Consent is one such lawful justification. Given the broad definition of “processing”, there are very few things an NLP organization can do without receiving individuals’ consent or an exemption from the Information Regulator.

Under section 11, some alternatives for individuals’ consents are as follows: (a) processing is necessary to carry out actions for the conclusion or performance of a contract to which the data subject is party; (b) processing complies with an obligation imposed by law on the NLP organization; (c) processing protects a legitimate interest of the data subject; (d) processing is necessary for the proper performance of a public law duty by a public body; or (e) processing is necessary for pursuing the legitimate interests of the NLP organization or of a third party to whom the information is supplied.

An NLP organization may refer to its purposes and activities to evaluate whether it meet the requirements of exceptions under section 11 or the discretionary exemption that may be granted by the Information Regulator under section 37.

7. What are the privacy risks related to data collection directly from persons? For example, when you interview people or they donate data etc.

First, interviewees or donators should consent to the collection. Second, notification must be given to data subjects before NLP researchers directly collect information from data subjects. Section 18 lists all necessary information a data subject should be aware of, such as the data being collected, the purposes of the collection, the name and address of the responsible party and the existence of the right of access to and the right to rectify the information collected, etc. NLP researchers should provide protective mechanisms as well. Please refer to Rules of Question 1&5 for detailed information.

8. Does the applicable law change depending on whether the content of the collected data is relevant to the individual?

POPIA only regulates information relating to an identifiable, living, natural person, and where it is applicable, an identifiable, existing juristic person. Thus, if the collected data is not relevant to the individual, it will not be governed by POPIA. However, if such information is derived from the creation of human intelligence, it may fall within the scope of copyright law in South Africa.

E. Prohibited content

1. What types of data may be prohibited from being (i) text data mined, (ii) stored or distributed, or (iii) generated?

This section analyzes the presence of legal constraints other than those related to intellectual property and information privacy laws in the South African jurisdiction. The issue is whether the jurisdiction provides for content-based prohibitions or restrictions on data that can be (i) text mined, (ii) lawfully stored or distributed, and (iii) generated by AI.

Rules:

The Constitution of the Republic of South Africa, 1996, under Section 16 gives everybody the right to freedom of expression, including freedom of the press and other media, freedom to receive or impart information or ideas, freedom of artistic creativity, and academic freedom and freedom of scientific research, but the right to freedom of expression does not extend to propaganda for war, incitement of imminent violence or advocacy of hatred that is based on race, ethnicity, gender or religion, and that constitutes incitement to cause harm. The limitations to the exercise of freedom of expression can be further determined through the lens of Section 10 of the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000 (hereafter the "Equality Act"). Section 10(1) of the Equality Act provides for the prohibition against hate speech based on one or more of the prohibited grounds where one advocates, propagates or communicates words against any person that can reasonably be construed to show a clear intention to be hurtful, harmful or to incite harm.

Moreover, Sections 9(3) and (4) of the Constitution provide that neither the State nor any person may, directly or indirectly, discriminate unfairly against anyone on one or more grounds, including race, gender, sex, pregnancy, marital status, ethnic or social origin, color, sexual orientation, age, disability, religion, conscience, belief, culture, language and birth, and that national legislation must be enacted to prevent or prohibit unfair discrimination.
Section 20 of the Equality Act allows individuals that have been subjected to hate speech or unfair discrimination to institute civil litigation against the perpetrator. The remedies may include the payment of damages, an order restraining specified practices or an apology, amongst others. In addition, the Cybercrimes Act 19 of 2020 regulates cybercrime relating to electronic communication such as social media. Chapter 4 of the Act gives both police officials and investigators the authority to search, seize or access any resource suspected to have been used for the commission of a cyber-crime.
In relation to personal information of children, Section 35 of POPIA prohibits the processing of personal information concerning a child. However, it may be processed if one or more of the following exceptions apply and the processing is:

a) carried out with the prior consent of a competent person;
b) necessary for the establishment, exercise or defense of a right or obligation in law;
c) necessary to comply with an obligation of international public law;
d) for historical, statistical or research purposes to the extent that:

- the purpose serves a public interest and the processing is necessary for the purpose concerned; or
- it appears to be impossible or would involve a disproportionate effort to ask for consent,
- and sufficient guarantees are provided for to ensure that the processing does not adversely affect the individual privacy of the child to a disproportionate extent; or

e) of personal information which has deliberately been made public by the child with the consent of a competent person.

Analysis:
In light of the applicable legal framework, activities of data processing should be aware of prohibitions on (i) disclosure of national security information (as better outlined under the following section 2), (ii) child pornography, (iii) conveyance of hate speech, promotion of terrorism, unfair discrimination, war, and incitement of imminent violence.

A) Text mining of confidential, secret and top secret information in relation to national security reasons (as better defined under section 2 below) may be held a crime against national sovereignty. Although the law does not explicitly prohibit the storage of this kind of data, its distribution is a crime.

B) Section 18 of the Children's Act, 2005 provides that a child is anyone under the age of 18 years old. The Films and Publications Act, 1996 defines "child pornography" as any image, real or simulated, however created, depicting a person who is or who is shown as being under the age of 18 years, engaged in sexual conduct or a display of genitals which amounts to sexual exploitation, or participating in, or assisting another person to engage in sexual conduct which amounts to sexual exploitation or degradation of children.
There are different types of classification of publications. Focusing exclusively on "refused classification" which is the publication that contains child pornography, the law prohibits any

person who knowingly broadcasts, distributes, exhibits in public, offers for sale hire or advertises for exhibition, sale or hire any film, game or a publication which has been classified as a refused classification and impose upon conviction a fine or imprisonment for a period not exceeding five years or both a fine and such imprisonment.
Under Section 19 of the Criminal Law (Sexual Offences and Related Matters) Amendment Act, 2007, child pornography legislation is aimed at the prohibition of any image, publication, depiction, description or sequence of child pornography. Therefore, it extend to texts to the extent that their content involves textual description of child pornography, including fictional pornography depicting minors.
Therefore, text data mining, storage, distribution and generation of data "describing" child pornography is punished under the applicable criminal legislation on child pornography.

C) Hate speech, promotion of terrorism, unfair discrimination, war, incitement of imminent violence are often conveyed in the form of texts, therefore requiring special attention when it comes to text data mining, distribution and generation. Although the text mining of data containing any of these contents, or their storage, would not be unlawful per se, the use of such data in algorithm training might lead to further generation and distribution of data.

2. What legal or cultural norms are implicated by prohibited content? What are the harms being prevented?

The legal norms that are implicated by prohibited content are (1) constitutional norms protecting democratic values of the Republic of South Africa; (2) laws and regulations governing privacy and personal data; (3) criminal laws protecting fraudulent or unauthorized access to data.

Each of these norms is intended to protect different interests, but they all seek to protect the constitutional values of social justice, human dignity, equality and the advancement of human rights and freedoms, non-racialism and non-sexism and to prevent the usage of private information for illegitimate or criminal purposes.

3. What types of licensing or other control mechanisms would be preferred under the applicable jurisdictions?

According to many scholars, the "fair use provision" – which is less restrictive than the fair dealing provision currently provided under Section 12A of the Copyright Amendment Bill – combined with the specific list of exceptions would provide South Africa with the "best of both worlds" combining openness and predictability. On the one hand, the open "fair use exception" makes the exceptions future-proof. It permits the law to adapt to new uses, technologies, and purposes which may not be anticipated in the specific exceptions. On the other hand, the list of specific exceptions provides a higher degree of predictability for the set of uses long authorized in South Africa copyright law. This approach is also consistent with the international law (e.g., the Berne Convention and the TRIPS Agreement to which South Africa is a party) "three-step" test which requires that countries confine copyright limitations or exceptions "to certain special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder".
For these reasons, proposals to amend the Section 12 of the Copyright Amendment Bill have been recently submitted to the South African Parliament's Portfolio Committee on Trade and Industry.

4. Is there a legal restriction on the distribution of data for national security reasons?

Information that is deemed to be confidential, secret and top secret, and state information determined as valuable should be protected and controlled by the state.
“National security” is not specifically defined in any law in South Africa. However, section 198 of the Constitution of the Republic of South Africa prescribed principles which govern national security in the Republic, including: (a) National security must reflect the resolve of South Africans, as individuals and as a nation, to live as equals, to live in peace and harmony, to be free from fear and want and to seek a better life; (b) The resolve to live in peace and harmony precludes any South African citizen from participating in armed conflict, nationally or internationally, except as provided for in terms of the Constitution or national legislation; (c) National security must be pursued in compliance with the law, including international law; (d) National security is subject to the authority of Parliament and the national executive.

According to section 7 and 11 of the Protection of Information Bill, 1) ‘‘Valuable information’’ means information contemplated in this Act whose unlawful alteration, destruction or loss is likely to infringe on the constitutional rights of the public or individuals or deny them of a service or benefit to which they are entitled; 2) State information may be classified as confidential if the information is sensitive information, the disclosure of which is likely or could reasonably be expected to cause demonstrable harm to the national security of the Republic; 3) State information may be classified as secret if the information is sensitive information, the disclosure of which is likely or could reasonably be expected to cause serious demonstrable harm to the national security of the Republic; 4) State information may be classified as top secret if the information is sensitive information, the disclosure of which is likely or could reasonably be expected to demonstrably cause irreparable or exceptionally grave harm to the national security of the Republic. Therefore, NLP researchers should not attempt to access or distribute abovementioned data.