TOWARDS THE STANDARDIZATION OF DATA LICENSES

Misha Benjamin[1], **Paul Gagnon**[1], **Negar Rostamzadeh**[1], **Chris Pal**[1,2,3], **Yoshua Bengio**[3,4,5], **Alex Shee**[1]
[1] Element AI
[2] Polytechnique Montréal
[3] MILA
[4] Canada CIFAR AI Chair
[5] Senior CIFAR Fellow

ABSTRACT

This paper provides a taxonomy for the licensing of data in the fields of artificial intelligence and machine learning. The paper's goal is to build towards a common framework for data licensing akin to the licensing of open source software. Increased transparency and resolving conceptual ambiguities in existing licensing language are two noted benefits of the approach proposed in the paper. In parallel, such benefits may help foster fairer and more efficient markets for data through bringing about clearer tools and concepts that better define how data can be used in the fields of AI and ML. The paper's approach is summarized in a new family of data license language - the Montreal Data License (MDL). Alongside this new license, the authors and their collaborators have developed a web-based tool to generate license language espousing the taxonomies articulated in this paper (which will be made available after peer review).

1 WHY A NEW FRAMEWORK FOR DATA LICENSING IS NECESSARY?

This paper provides a taxonomy for the licensing of data in the fields of artificial intelligence and machine learning. The paper's goal is to build towards a common framework for data licensing akin to the licensing of open source software. Increased transparency and resolving conceptual ambiguities in existing licensing language are two noted benefits of the approach proposed in the paper. In parallel, such benefits may help foster fairer and more efficient use of data through bringing about clearer tools and concepts that better define how data can be used in the fields of AI and ML. The paper's approach is summarized in a new family of data license language - the Montreal Data License (MDL).

Understanding how data may be used in the field of AI and ML is a challenge that imposes costs on individuals and institutions. These transaction costs consist of resources allocated to harmonizing the data itself as well as assessing whether the data itself is usable (both technically and from a legal standpoint). Technical metadata does not clarify how data may legally be used. Further, those who are making data available do not benefit from a modular framework to better reflect their intent. If the data that is released is sensitive or contains personal information, it is necessary to have better and clear ways to convey what actions can and cannot be performed with data. In turn, not being able to modulate what rights may be granted makes it harder to enforce licensing terms.

The biggest companies in the world suffer least from such limitations – their own scale and data-generating capabilities through massive use of their platforms mean they collect and use data on a scale that most other market participants cannot match. For the benefits of ML and AI to be accessible and benefit a wider realm of humanity, all participants need to be on a level playing field. One way to bridge this gap is through a more transparent, predictable ways to license data with clear legal language and modularity that better reflects intent.

Recent efforts in standardizing the presentation of metadata for ML models are highly relevant, and contribute to fostering transparency in the fields of ML and AI (Mitchell et al., 2019; Gebru et al., 2018). For ML and AI to continue their growth, and for that growth to beneficial for all, standardized terminology and increased predictability is necessary. This article aims to provide a first step towards such standards with respect to data licensing.

2 “Use” of Data in ML and AI

2.1 Issues with Existing Database License

Commonly used language accompanying data is vague and unclear as to the permissions that are granted. Analyzing these permissions requires the time, resources and skill to appropriately ascertain whether a certain database can be used. The following are illustrative examples of the conceptual ambiguities present in commonly used license terms, which the taxonomy and licensing language presented in this paper aims to resolve.

Lack of Nuance on “Use”: Licenses typically grant the right to “use” data - without regard to what “use” actually means. Indeed, licenses may define use-cases or restrictions for “research use” or a “commercial use”. However, devoid of further context, the licenses all posit one homogenous notion of “use”. This forgoes the intricacies of ML and AI as to how data is actually used.

Commercial vs Non-Commercial Use: When assessing “commercial use” as used in data licenses, whose intent or purpose is considered? If an employee participates in fundamental research, but is employed by a for-profit entity, should it be considered commercial? Conversely, consider a researcher within a university’s whose technology transfer office may patent or commercialize uses of a dataset – would this otherwise academic context be considered commercial? What about the timing of such determination – could an initially non-commercial use be later requalified? If so, by whom? Another example of how the restriction against commercial use is problematic is the use-case of a for-profit company working for a non-profit on a specific project. Could it use the database to train a model for use by a non-profit? If so, could it reuse that model for other purposes after? What if it generated significant positive publicity or goodwill for the for-profit company?

Commercial vs Non-Commercial Use: When assessing “commercial use” as used in data licenses, whose intent or purpose is considered? If an employee participates in fundamental research, but is employed by a for-profit entity, should it be considered commercial? Conversely, consider a researcher within a university’s whose technology transfer office may patent or commercialize uses of a dataset – would this otherwise academic context be considered commercial? What about the timing of such determination – could an initially non-commercial use be later requalified? If so, by whom? Another example of how the restriction against commercial use is problematic is the use-case of a for-profit company working for a non-profit on a specific project. Could it use the database to train a model for use by a non-profit? If so, could it reuse that model for other purposes after? What if it generated significant positive publicity or goodwill for the for-profit company?

Research The development of ML and AI faces tremendous constraints on talent to drive the changes, and, more importantly, to ensure that such talent meaningfully participates in transmitting such knowledge through universities and other academic fora. Hence, “commercial” and “non-commercial” may also be a false dichotomy, or at least not reflect the realities of research. Moreover, researchers from academia may themselves partake in research in collaboration with peers working within for-profit companies, thus blurring the lines further. In this context, some database licenses limit the purpose for which they are licensed to “research” use. By restricting use of data to “research”, is the intent then to posit “research” in opposition to “commercialization of research products”? There is conceptual ambiguity that needs to be be resolved.

Lack of Uniformity Another apparent issue with the language accompanying datasets is that there is relatively no uniformity or standard terminology used across these different licenses. Free and Open Source Software (FOSS) differs in that, while there exist a great number of FOSS licenses that can be used, commonly used licenses have centered around a relatively limited number of FOSS licenses. Further splintering or creation of new licenses is generally discouraged. Consider, however, that the term “use” for software is conceptually sound when considered in contrast to “use” of data in ML and AI. This means that the difficulties encountered in standardizing FOSS license terms may be heightened for datasets used in ML and AI.

Share-Alike Requirements There are datasets that are licensed under licensing terms that bring functional ambiguities that are difficult to reconcile with uses cases in ML and AI. One such license requirement is the “share alike” requirement that is found under the Creative Commons Share Alike license (CC-SA). The CC-SA, as with most licenses of the Creative Commons family, is particularly well-suited to be used in conjunction with copyrighted works in creative fields. The share alike requirement prevents downstream users from re-appropriating licensed content under other license terms. For example, one cannot repackage a collection of photographs licensed under CC-SA and resell it online, or make it available under restrictive terms.

2.2 Consequences of uncertainty

The consequences of the uncertainty related to usage rights are manifold. Many actors in the AI community are forced to spend time and effort attempting to interpret vague language and make risk assessments based on that uncertainty. Depending on how the terms are interpreted, the user may (i) refrain from using a database that it would have the right to use, resulting in less productivity, less advancement of research, lower quality of data (which can often cause more bias or unfairness) and/or more data acquisition costs; or (ii) use a data in violation of the terms of the dataset - this can result in privacy-related harms and copyright violations, as well as lower the incentives to advance research while respecting data and privacy rights in the underlying data points. This creates perverse incentives, where the more scrupulous actors actually have less data to work with and are less successful, whereas the less scrupulous actors benefit from the uncertainty - effectively making a disregard for copyright and privacy a competitive advantage. In an era when there is a lot effort from regulators, government, tech companies and NGOs devoted to ensuring respect for data, and particularly personally identifiable information (PII), creating a market distortion where violations of those rights are rewarded seems particularly regrettable. This also may disproportionately affect small and medium enterprises (SMEs) and research undertaken by institutions with little or no means to obtain legal services.

3 Proposed Taxonomy for Standardized AI Rights to Data

How is data “used” in ML and AI? What does it mean when it is stated that algorithms need be exposed to data in order to fundamentally achieve their purpose and harness the potential of the advanced techniques being discussed? In AI and ML, different actions may be taken with data: it is “used” in different ways, for different purposes[1]. Our Data License articulates and clarifies the notion of **use** in the ways presented in the table.

4 Conclusion

The emergence of ML and AI is already shaping society, political systems and our economies. The underlying assets driving such changes are largely informational. Access and licensing of data can thus be understood as one of the cornerstone of the development of ML and AI. This is true in an abstract sense, but when combined to the fact that there exists a widening data gap between multinational firms with platform-based business models on one hand, and governments, citizens and other businesses on the other, the need for clarity in data licensing becomes imperative.

This paper aimed to bring a step forward to bring about this clarity from a legal standpoint by providing a licensing framework anchored in practical realities of ML and AI. The goal is ambitious: providing a common frame of reference to create standards for data licensing to compare with those found in open source software. Doing so will foster transparency, algorithmic fairness and fairer market dynamics.

[1] complimentary information is provided in supplementary materials: As outlined in Appendix 1. These use cases are the foundation of the Montreal Data License, the proposed text of which is found at Appendix 2. An online tool for everyone to generate the relevant licensing language they wish to use is made available through a website that will be available after peer review

The Data itself

Access	To access, view and/or download the Data to view it and evaluate it (evaluation algorithms may be exposed to it, but no Untrained Models).
Labelling	To build upon Data by adding tags, labels or other metadata to the dataset or subsets of the Data.
Distribute	Make all or part of the Data available to third parties.
Represent	Transform the data into a new representation, thereby re-representing each data element in a way that mimics the effects of the initial data itself (i.e. the purpose or end-result consists of a suitable alternative to such Data).

Use of the Data in Conjunction with Models

Benchmark (case 1: without training a model, Case 2: where a model is trained on the data so as to evaluate it)	To access the Data, use the Data as training data to evaluate the efficiency of different Untrained Models, algorithms and structures, but excludes reuse of the Trained Model, except to show the results of the Training. This includes the right to use the dataset to measure the performance of a Trained or Untrained Model, without however having the right to carry-over weights, code or architecture or implement any modifications resulting from such evaluation.
Research	To access the Data, use the Data to create or improve Models, but without the right to use the Output or resulting Trained Model for any purpose other than evaluating the Model Research under the same terms.
Publish	To make available to third parties the Models resulting from Research, provided however that third parties accessing such Trained Models have the right to use them for Research or Publication only.
Internal Use	To access the Data, use the Data to create or improve Models and resulting Output, but without the right to Output Commercialization or Model Commercialization. The Output can be used internally for any purpose, but not made available to Third Parties or for their benefit.
Output Commercialization	To access the Data, use the Data to create or improve Models and resulting Output, with the right to make the Output available to Third Parties or to use it for their benefit. The Trained Model itself however cannot be not made available to Third Parties. This would allow SaaS commercialization.
Model Commercialization	Make a Trained Model itself available to a Third Party, or embodying the Trained Model in a product or service, with or without direct access to the Output for such Third Party.

A small step towards this is the creation of a Standard Data Licence – a modular approach to data licensing in AI and ML that the authors hope will break ground and be adopted by the AI and ML communities.

REFERENCES

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAT**, 2019.