Reclaiming the Digital Commons: A Public Data Trust for Training Data

Alan Chan [1 2] **Herbie Bradley** [3 4] **Nitarshan Rajkumar** [3]

Abstract

Democratization of AI means not only that people can freely use AI, but also that people can collectively decide how AI is to be used. The rapid pace of AI development and deployment currently leaves little room for collective control. Monopolized in the hands of private corporations, the development of the most capable foundation models has proceeded largely without public input. There is currently no implemented mechanism to account for their negative externalities like unemployment and the decay of the digital commons. In this work, we propose that a public data trust assert control over training data for foundation models. First, we argue in detail for the existence of such a trust. We also discuss feasibility and potential risks. Second, we detail a number of ways for a data trust to incentivize model developers to use training data only from the trust. We propose a mix of verification mechanisms, potential regulatory action, and positive incentives. We conclude by highlighting other potential benefits of our proposed data trust and connecting our work to ongoing efforts in data and compute governance.

1. Introduction

Private companies dominate the development of the most capable AI systems (Giattino et al., 2022). The staggering amounts of compute involved (Sevilla et al., 2022; Giattino et al., 2022) mean that large tech companies or those backed by massive amounts of venture capital have disproportionate power in guiding the direction of technological progress. From a resource perspective, it remains difficult for academic or non-profit collaborations to match the financial weight of the private sector. From a philosophical perspective, democratization of AI is not solely about the free deployment of AI without regard for social consequence. Rather, we hold as Shevlane (2022) does that democratization also means collective decision-making power over how AI is to be developed and deployed. Narrow democratization could frustrate the broad democratic ideal; unstructured access to AI systems could hinder societies from restricting certain uses they deem undesirable.

Collective decision-making power over AI is deficient in two key respects. First, data creators cannot prevent AI developers from using their data. Opt-out mechanisms are lacking and the training datasets of many of the largest models are private. Second, there is no implemented mechanism to ensure that the profits of AI development and deployment are distributed widely, particularly as a way to redress negative externalities. Even if an individual were to threaten to withhold their data from a model developer, they would have effectively no bargaining power since a few data points likely make no significant difference in the final performance of a model.

We focus on the large training datasets scraped from the **digital commons**—the collective intellectual and cultural contributions of humanity that are in digital form—and also on bespoke crowdworker data as a point of intervention for redressing the power imbalance between model developers and human data creators. The digital commons is the product of humanity's cumulative efforts, yet in AI development the fruits of the commons are captured by the few.

To address the imbalance of power, we propose the creation of a public data trust. We intend this data trust to be national and located in a jurisdiction with a high concentration of AI development, such as the US or the UK. Our data trust would gate access to the most important data for model training: pre-training data from the internet and human feedback data from annotators. Our gating is meant to apply primarily to commercial AI developers. We focus our attention on general-purpose AI systems such as foundation models, given their likely role as important components of future AI systems and their increasingly wide adoption. Our contributions are as follows. **1)** We argue for the creation of a public data trust to hold training data, so as to address

[1]Mila - Quebec AI Institute, Canada [2]Université de Montréal, Canada [3]CAML Lab, University of Cambridge, UK [4]EleutherAI. Correspondence to: Alan Chan <alan.chan@mila.quebec>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023.

the private concentration of power in AI development and safeguard the digital commons. **2)** We sketch a plan for how the data trust could assert control over training data.

2. The Case for a Data Trust

An outline of our case for a data trust is as follows.

1) AI development heavily depends upon the **digital commons**: the collective intellectual and cultural contributions of humanity that are in digital form.

2) AI development is extremely concentrated in the private sector. Those who contribute to the digital commons, including the general public and sector-specific individual such as artists, have little decision-making power over the development of deployment of AI compared to the AI developers.

3) AI deployment results in negative externalities to the public; there are currently no effective mechanisms to address these negative externalities.

4) A data trust that gated training data access to the digital commons would help to correct the power imbalance so as to redress negative externalities, such as by setting up a digital commons fund financed by model royalties.

3. Building a Data Trust

We briefly sketch a plan for building a public data trust for training data.

3.1. Obtaining Data for the Trust

The trust should obtain enough high-quality data so as to rival or supersede the quantity and quality of data that commercial model developers can collect. Data would include both pre-training data and human feedback data generated by human annotators. To obtain data, the trust should scrape the internet, partner with data communities, encourage the entrustment of personal data from online platforms, and work with annotators to include human feedback data.

The data trust should curate and document the collected data in detail, following best practices (Gebru et al., 2021; Hutchinson et al., 2021; Mitchell et al., 2023). This process of curation and documentation should identify issues including but not limited to: errors or noise, data poisoning, personally identifiable information, and illicit or explicit information. The choice of data to exclude from a pre-training set can be difficult. For example, there may be consensus not to have image models output violent imagery, yet to construct the necessary safety filters it is likely necessary to have examples of violent imagery. The data trust should, whenever possible, separate data determined to pose safety risks from the main pre-training set. Since the act of doing so is inherently value-laden, the trust should carry out this process through or under the supervision of a diverse panel of experts across disciplines, with explicit representation of voices from marginalized communities. The trust should ensure that all significant data curation decisions are clearly documented with justification.

Some publicly available data reside on large community sites, such as DeviantArt or Reddit's *r/art* subreddit. Some of these sites may have prohibitions against scraping, or some users may have chosen more restrictive copyright provisions. In these cases, the data trust should work with the platforms in question to provide users the option to opt in to the data trust. Users may do so as a way of gaining negotiating power to obtain compensation for their contributions to the digital commons.

To obtain high-quality human feedback data, which is increasingly responsible for the strong performance of widely deployed models foundation models (Bai et al., 2022), data trusts could work with both crowdworker collectives (Irani & Silberman, 2013) and crowdsourcing platforms like Surge and Upwork to include human feedback data from crowdworkers in the trust. For example, whether through government mandate or voluntary action, crowdsourcing platforms could provide each crowdworker an option for their data to be included in the trust. Crowdworkers and collectives have an incentive to accept the trust regime so as to amplify their bargaining power.

3.2. Verifying Compliance

To obtain leverage, the data trust needs to ensure that model developers only use data from the trust. We consider it infeasible to ban scraping outright. Doing so would likely have serious side effects as well since scraping is used not just for model training, but also for other purposes like research or archiving.

We propose technical verification of a model developer's claim that they are only using the trust's data, under the assumption that a model developer has committed, for example through contract, only to use data to which the trust grants them access. Possible options for this verification include techniques based on data poisoning (Li et al., 2020; Carlini & Terzis, 2022; Carlini et al., 2023) and proof-of-learning (Jia et al., 2021). The goal is to verify the following: **1)** Anybody who obtains data from the data trust actually trains the model with the trust's data. **2)** The data trust's dataset is the only dataset used to train the model. **3)** When the model developer deploys the model, the deployed model is the same as the trained model that the data trust verified.

3.3. Incentives to Submit to the Data Trust Regime

To incentivize model developers to submit to the data trust's regime, there are a few options. First, governments could

stipulate that commercial model developers use only data from the trust. Second, the data trust could provide certifications for companies that voluntarily agree to its regime. Third, model developers may have positive reasons to use the trust's data given that the trust would take on the burden of collecting and curating data.

Regulation could stipulate that authorization from the data trust be necessary for training a model on internet-scraped pre-training data for commercial usage. Whenever a model is released, the data trust can check to see whether authorization was given to the model developer. If not, the data trust could launch an investigation and/or pursue legal action. If yes, the data trust could proceed with the verification mechanisms in Section 3.2.

As an alternative to regulation, the data trust could provide certifications for companies that voluntarily agree only to use the trust's data and submit to the verification regime in Section 3.2. Such a certification would work similarly to Fair Trade labels (Dragusanu et al., 2014). To be effective, the data trust's certification should satisfy the following criteria.

1. Consumers can easily distinguish between model developers who have certification and those who do not.

2. There are consumers that care about model developers having certification.

3. The buying power of consumers who care about certification is enough to offset the increased cost of a model developer's complying with certification requirements.

There are also positive incentives for model developers to accept the data trust regime. Data collection tends to be an arduous, costly process. Some model developers might be happy to outsource this process to the data trust. Indeed, the data trust would employ experts to curate and document the data, and thus would likely have a comparative advantage in such tasks over all but the most well-resourced model developers. Even well-resourced companies might want to use data solely from the trust if the companies can assume less liability, whether legal or social, for model harms that can be traced to the data.

4. Conclusion

Through data, the construction of today's most advanced AI systems depends upon the digital commons. However, the public holds relatively little power over the conditions of AI deployment. We propose a data trust to hold key sources of training data to begin to rectify this power imbalance. Our data trust would collect training data, create a verification regime, and support a variety of methods to incentivize developers to submit to the regime. Our proposal is a high-level overview of how such a trust might function; further work needs to be done to sketch out a legal framework for how governments might implement a data trust.

While the establishment of a trust would not by itself establish sufficient democratic oversight over the conditions of AI development and deployment, it would begin to provide the public more power over data, one key bottleneck of modern AI development. To ensure broad distribution of the fruits of AI progress, future work should aim to improve democratic control over both data and other bottlenecks such as compute.

Acknowledgements

We benefited greatly from insightful comments from the following individuals: Lauro Langosco, Usman Anwar, Shahar Avin, Stella Biderman, Henry Ashton, Micah Carroll, Yawen Duan, David Krueger, Robert Harling. Herbie Bradley's contributions were supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (reference EP/S022961/1).

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

Carlini, N. and Terzis, A. Poisoning and Backdooring Contrastive Learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iC4UHbQ01Mp.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning Web-Scale Training Datasets is Practical, February 2023. URL http://arxiv.org/abs/2302.10149. arXiv:2302.10149 [cs].

Dragusanu, R., Giovannucci, D., and Nunn, N. The Economics of Fair Trade. *Journal of Economic Perspectives*, 28(3):217–236, September 2014. ISSN 0895-3309. doi: 10.1257/jep.28.3.217. URL https://www.aeaweb.org/articles?id=10.1257/jep.28.3.217.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, December 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3458723. URL https://dl.acm.org/doi/10.1145/3458723.

Giattino, C., Mathieu, E., Broden, J., and Roser, M. Artificial Intelligence. *Our World in Data*, 2022.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure, January 2021. URL http://arxiv.org/abs/2010.13561. arXiv:2010.13561 [cs].

Irani, L. C. and Silberman, M. S. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 611–620, New York, NY, USA, April 2013. Association for Computing Machinery. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2470742. URL https://doi.org/10.1145/2470654.2470742.

Jia, H., Yaghini, M., Choquette-Choo, C. A., Dullerud, N., Thudi, A., Chandrasekaran, V., and Papernot, N. Proof-of-Learning: Definitions and Practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056, May 2021. doi: 10.1109/SP40001.2021.00106. ISSN: 2375-1207.

Li, Y., Zhang, Z., Bai, J., Wu, B., Jiang, Y., and Xia, S.-T. Open-sourced Dataset Protection via Backdoor Watermarking, November 2020. URL http://arxiv.org/abs/2010.05821. arXiv:2010.05821 [cs].

Mitchell, M., Luccioni, A. S., Lambert, N., Gerchick, M., McMillan-Major, A., Ozoani, E., Rajani, N., Thrush, T., Jernite, Y., and Kiela, D. Measuring Data, February 2023. URL http://arxiv.org/abs/2212.05129. arXiv:2212.05129 [cs].

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute Trends Across Three Eras of Machine Learning, February 2022. URL https://arxiv.org/abs/2202.05924v2.

Shevlane, T. Structured access: an emerging paradigm for safe AI deployment, April 2022. URL http://arxiv.org/abs/2201.05159. arXiv:2201.05159 [cs].