On Durable Data Usage Policies across AI Model Supply Chain

RUI ZHAO, University of Oxford, UK
TERESA HUANG, University of Oxford, UK
MOMING DUAN, East China Normal University, China
JUN ZHAO, University of Oxford, UK

The issue of ethical data curation in AI (foundation) models is not alone to the dataset itself, but also to the whole lifecycle of AI supply chain, from all aspects about training data, model training, and model usage practices. We highlight the entanglement of these aspects, and draw an attention to the necessity of whole-chain policy management. We also discuss some promising directions and solutions for achieving that goal, particularly through practical formal policy encoding and reasoning.

ACM Reference Format:
Rui Zhao, Teresa Huang, Moming Duan, and Jun Zhao. 2025. On Durable Data Usage Policies across AI Model Supply Chain. 1, 1 (September 2025), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Background: AI and Data

Modern artificial intelligence (AI) relies on machine learning (ML) that identifies and generalizes patterns from training data, leading to significant hunger of data curation [4, 7]. This has become more prevailing with the wide utilization and competition of large language models (LLM), where people estimated that all data from the Internet will soon be exhausted [12, 15]. On the other hand, profound debates and actions have surrounded using artwork (including other creational work such as music) in ML model training and output generation [11, 13], with relevant research in "poisoning" (ways to circumvent using the data in model training) [10]. Essentially, wider sources of data and the corresponding ethical issues have thus become an active area of discussion.

Many existing discussions (e.g. [3]) concerned the ethical practices for (manually) curating data and then forming a dataset. However, that often falls short when data collection accompanies AI tool *usage*, such as LLM applications/services like ChatGPT or copilots, where user data may later be used for training improved models. This issue is further amplified in other AI domains and applications where active data collection is mandated, such as smart speakers or autonomous vehicles [2, 5, 6]. Guidance or technologies to assist those parts of the practices are necessary to make the whole AI field more ethical.

2 Position

We take the position that we need to design and develop (semi-)automated and future-proof technologies to govern artifact usages across the whole lifecycle of AI supply chain, supporting the requirements of all stakeholders, especially the data providers. Fig. 1 depicts the relevant steps in the supply chain, and the relevant *cycle branches*, where an artifact (data or model) may flow to a different step for a different purpose.

Authors' Contact Information: Rui Zhao, rui.zhao@cs.ox.ac.uk, University of Oxford, UK; Teresa Huang, University of Oxford, UK; Moming Duan, duanmoming@gmail.com, East China Normal University, Shanghai, China; Jun Zhao, jun.zhao@cs.ox.ac.uk, University of Oxford, UK.

ACM XXXX-XXXX/2025/9-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

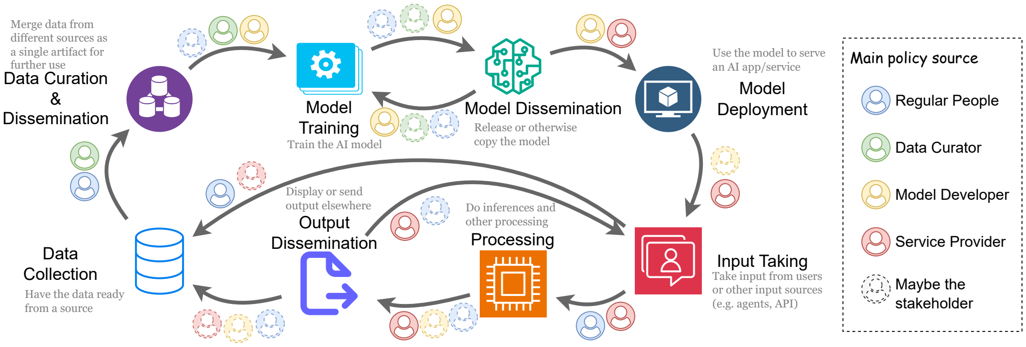
Merge data from different sources as a single artifact for further use
Data Curation & Dissemination
Model Training
Train the AI model
Model Dissemination
Release or otherwise copy the model
Use the model to serve an AI app/service
Model Deployment
Main policy source
Regular People
Data Curator
Model Developer
Service Provider
Maybe the stakeholder
Display or send output elsewhere
Output Dissemination
Do inferences and other processing
Processing
Input Taking
Take input from users or other input sources (e.g. agents, API)
Data Collection
Have the data ready from a source

Fig. 1. Artifact (data and model) lifecycle(s) in AI model supply chain, and main policy sources governing step transitions

Because of the multi-cyclic nature, we believe *correct* "policy" should be associated with each artifact, and being (semi-)automatically verified whenever the artifact is used, including downstreams. We identify five requirements:

(1) Each policy is precisely and unambiguously specified, by the relevant stakeholders themselves;
(2) The policy language should be extensible to allow additional concepts to be added;
(3) The policy should permit future users by default, given their compliance of the requirements;
(4) Multiple policies from multiple artifacts may be checked at the same time;
(5) Derived artifact should have derived policy based on those from inputs, and the step it has gone through.

We believe our prior work on usage control and licence compatibility, together with others' work, showed evidence of the necessity, and formed innovative foundations for achieving the goal.

More specifically, in [8], through formal licence modelling and reasoning, Duan et al revealed the overlook of sensible licences for ML models and datasets and the urgent need of appropriate versatile licences; it also showed a promising direction of composable licences, further discussed in [9], with an extensible formal foundation. The "perennial" policy language [19] forms an appropriate foundation in fulfilling the requirements above, as a significant improvement compared to existing policy languages that fall short in different aspects in meeting the needs. This model forms a stark contrast to many existing models by allowing both data providers and data consumers to encode their own policies independently, and use the generic reasoning to verify compliance across data usages cycles, across downstreams. Its key designs have also been used in contexts of provenance data generated from scientific workflows [18], demonstrating the ability in handling complex data-flows that can be unfolded as directed acyclic graphs.

We believe, as a next step, a unification of different formal languages should be performed, including standards such as Open Digital Rights Language (ODRL) [1]. That would form a flexible, comprehensive and interoperable foundation to construct policies by different vendors, independently. Formal reasoning is performed to a) verify if artifact providers' policies are respected by the consumer (e.g. a data curation activity, a computational task such as model training, or a deployed app/service), and b) derive policies for any output artifact by the consumer, so further consumers of these output artifacts will receive appropriate policies to check against.

Appropriate supporting mechanisms may also be developed, such as undeniable and indelible recording of activities and policies, for future verification of compliance; corresponding protocols and standards for handling policy update should also be developed. The *recording* aspect may involve techniques like digital signature, verifiable credentials [17] and blockchain [14]; the data schema can be built on extensible and interoperable standards like W3C PROV [16]. Appropriate user-facing tools are also desirable for facilitating lay-person's adoption, such as policy advisors.

References

[1] 2018. ODRL Information Model 2.2. https://www.w3.org/TR/odrl-model/

[2] Wael Albayaydh and Ivan Flechais. 2024. "Innovative Technologies or Invasive Technologies?": Exploring Design Challenges of Privacy Protection With Smart Home in Jordan. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 76:1–76:54. doi:10.1145/3637353

[3] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical Considerations for Responsible Data Curation. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 55320–55360. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ad3ebc951f43d1e9ed20187a7b5bc4ee-Abstract-Datasets_and_Benchmarks.html

[4] Ms. Aayushi Bansal, Dr. Rewa Sharma, and Dr. Mamta Kathuria. 2022. A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. *ACM Comput. Surv.* 54, 10s (Sept. 2022), 208:1–208:29. doi:10.1145/3502287

[5] Johana Bhuiyan. 2023. TechScape: Self-driving Cars Are Here and They're Watching You. *The Guardian* (July 2023). https://www.theguardian.com/technology/2023/jul/04/smile-youre-on-camera-self-driving-cars-are-here-and-theyre-watching-you

[6] Cara Bloom, Joshua Tan, Javed Ramjohn, and Lujo Bauer. 2017. Self-Driving Cars and Data Collection: Privacy Perceptions of Networked Autonomous Vehicles. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. 357–375. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/bloom

[7] Zefeng Chen, Wensheng Gan, Jiayang Wu, Kaixia Hu, and Hong Lin. 2025. Data Scarcity in Recommendation Systems: A Survey. *ACM Trans. Recomm. Syst.* 3, 3 (March 2025), 27:1–27:31. doi:10.1145/3639063

[8] Moming Duan, Mingzhe Du, Rui Zhao, Mengying Wang, Yinghui Wu, Nigel Shadbolt, and Bingsheng He. 2025. Position: Current Model Licensing Practices Are Dragging Us into a Quagmire of Legal Noncompliance. In *Forty-Second International Conference on Machine Learning Position Paper Track*. https://openreview.net/forum?id=1rh8iTehBc

[9] Moming Duan, Rui Zhao, Linshan Jiang, Nigel Shadbolt, and Bingsheng He. 2024. "They've Stolen My GPL-Licensed Model!": Toward Standardized and Transparent Model Licensing. arXiv:2412.11483 [cs] doi:10.48550/arXiv.2412.11483

[10] Hanna Foerster, Sasha Behrouzi, Phillip Rieger, Murtuza Jadliwala, and Ahmad-Reza Sadeghi. 2025. LightShed: Defeating Perturbation-based Image Copyright Protections. In *34th USENIX Security Symposium (USENIX Security 25)*. 7271–7290. https://www.usenix.org/conference/usenixsecurity25/presentation/foerster

[11] Paul Glynn. 2025. Artists Release Silent Album in Protest against AI Using Their Work. *BBC News* (Feb. 2025). https://www.bbc.com/news/articles/cwyd3r62kp5o

[12] Dan Milmo. 2025. Elon Musk Says All Human Data for AI Training 'Exhausted'. *The Guardian* (Jan. 2025). https://www.theguardian.com/technology/2025/jan/09/elon-musk-data-ai-training-artificial-intelligence

[13] Dan Milmo. 2025. 'Mass Theft': Thousands of Artists Call for AI Art Auction to Be Cancelled. *The Guardian* (Feb. 2025). https://www.theguardian.com/technology/2025/feb/10/mass-theft-thousands-of-artists-call-for-ai-art-auction-to-be-cancelled

[14] Ricardo Neisse, Gary Steri, and Igor Nai-Fovino. 2017. A Blockchain-based Approach for Data Accountability and Provenance Tracking. In *Proceedings of the 12th International Conference on Availability, Reliability and Security (ARES '17)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3098954.3098958

[15] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will We Run out of Data? Limits of LLM Scaling Based on Human-Generated Data. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24, Vol. 235)*. JMLR.org, Vienna, Austria, 49523–49544.

[16] W3C. 2013. PROV-overview : An Overview of the PROV Family of Documents. https://www.w3.org/TR/prov-overview/

[17] W3C. 2022. Verifiable Credentials Data Model v1.1. https://www.w3.org/TR/vc-data-model/

[18] Rui Zhao, Malcolm Atkinson, Petros Papapanagiotou, Federica Magnoni, and Jacques Fleuriot. 2021. Dr.Aid: Supporting Data-governance Rule Compliance for Decentralized Collaboration in an Automated Way. In *The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*. doi:10.1145/3479604

[19] Rui Zhao and Jun Zhao. 2024. Perennial Semantic Data Terms of Use for Decentralized Web. In *Proceedings of The ACM Web Conference 2024*. ACM, Singapore, 2238–2249. doi:10.1145/3589334.3645631