

Vision-Language Assisted Data Curation for Next-Generation Trajectory Prediction Systems

Mugdha Chhabra

School of Computing, Clemson University, Clemson, SC, USA.
chhabra1989@clemson.edu

Keelan Mulvey

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
keelanmail@ucf.edu

Dario Antonioni

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
antonionidario@ku.edu

Jeffery C. Zamora

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
jefferyczamora697@uc.edu

Abstract

Trajectory prediction systems are becoming foundational components of autonomous mobility, robotic navigation, and urban analytics, yet their performance and safety depend less on model architecture alone than on the quality, balance, and interpretive richness of the data used to train and evaluate them. This paper examines vision-language assisted data curation as a socio-technical infrastructure for next-generation trajectory prediction. It argues that vision-language models can support semantic labeling, rare-scenario discovery, cross-modal consistency checking, and human-in-the-loop review, but that these benefits introduce new dependencies, opacity, and governance obligations. The discussion is organized around system-level trade-offs involving annotation scalability, spatial and temporal coverage, fairness across road users, robustness under distribution shift, deployment sustainability, and policy accountability. The paper connects advances in sequence modeling and multimodal representation learning with established concerns in data documentation, model reporting, infrastructure studies, and algorithmic auditing. It also considers why curation pipelines must be designed as durable institutions rather than transient preprocessing steps. Particular attention is given to the risks of automated curation reproducing historical biases, marginalizing underrepresented environments, and concealing labor through interface abstraction. The analysis concludes that vision-language assistance is most valuable when it augments rather than replaces human judgment, when its outputs are auditable, and when curation decisions are traceable across the full lifecycle of trajectory prediction systems. The paper proposes that next-generation trajectory prediction requires curation architectures that treat data quality, governance, and evaluation as continuous operational commitments rather than one-time engineering tasks.

Keywords

trajectory prediction, vision-language models, data curation, socio-technical infrastructure, fairness, governance, autonomous systems, robustness.

1. Introduction

Trajectory prediction has moved from a specialized motion forecasting problem to a core capability within autonomous driving, mobile robotics, air traffic management, and smart-city analytics. Early sequence models established the importance of temporal representation and attention mechanisms for learning long-range dependencies in movement data [1], while later multimodal systems demonstrated that paired visual and textual supervision can create transferable representations across tasks and domains [2]. In parallel, trajectory prediction research developed specialized architectures for social interaction, generative multimodality, dynamic feasibility, and map-aware reasoning [3], [4], [5], [6]. These technical advances have been accompanied by large-scale datasets that provide multimodal sensor streams, high-definition maps, and increasingly diverse interaction scenarios [7], [8], [9], [10]. The result is a field with powerful model families but fragile data foundations, because the cost of producing, validating, and maintaining high-quality trajectory data remains high.

The central claim of this paper is that vision-language assistance should be understood as a curation infrastructure rather than as a model improvement technique. Curation infrastructure includes the people, routines, tools, standards, contracts, and review processes that transform raw sensor observations into trustworthy training and evaluation assets. This perspective draws on foundational work on foundation models and their systemic risks [11], critiques of large-scale language technologies and their hidden social costs [12], and structured documentation practices such as datasheets and model cards [13], [14]. It also aligns with infrastructure studies, which emphasize that durable systems are shaped by invisible maintenance, standards, and organizational arrangements [18]. When vision-language models are inserted into curation, they do not simply reduce labeling cost; they redistribute interpretive authority, create new failure modes, and alter the accountability relationships among data collectors, annotators, model developers, auditors, and affected publics.

A system-level view is necessary because trajectory prediction failures rarely originate in a single algorithmic component. They emerge from mismatches between training distributions and deployment environments, incomplete semantic labels, inconsistent map associations, scarce representation of rare but safety-critical maneuvers, and evaluation protocols that reward average accuracy while ignoring tail behavior. Vision-language models can help surface such problems by connecting visual scenes to natural-language descriptions, retrieving similar scenarios, identifying anomalous interactions, and generating structured metadata. However, the same models can also amplify biases, hallucinate plausible but incorrect scene interpretations, and obscure the provenance of curation decisions. Therefore, the design question is not whether vision-language assistance is useful, but under what architectural, institutional, and policy conditions it improves the reliability of trajectory prediction without creating new forms of fragility.

2. Foundations and System Context

Trajectory prediction systems operate within layered technical and social environments. At the sensing layer, cameras, lidar, radar, and global positioning systems produce heterogeneous observations with different sampling rates, noise profiles, and failure modes. At the representation layer, object detectors, trackers, map encoders, and scene graph builders convert raw signals into structured state histories. At the prediction layer, sequence models,

graph neural networks, and generative decoders estimate future positions and intentions [3], [4], [5]. At the deployment layer, planners, safety monitors, fleet operators, regulators, and insurers interpret prediction outputs. Data curation cuts across all these layers because labels, maps, scenario tags, and evaluation splits determine what the prediction model can learn and what it will ignore. A system that performs well on a benchmark may still fail in a city with different intersection geometry, informal traffic norms, or vulnerable road user behavior.

The evolution of trajectory datasets illustrates this dependence. Datasets such as nuScenes, Argoverse, the Interaction dataset, and the Waymo Open Dataset have expanded the scale and multimodal richness of available data [7], [8], [9], [10]. They provide calibrated sensors, tracking annotations, map elements, and scenario diversity that enable reproducible research. Yet scale alone does not guarantee representativeness. Geographic coverage is uneven, with many datasets concentrated in a small number of cities and driving cultures. Temporal coverage may miss seasonal effects, construction zones, emergencies, and rare weather conditions. Annotation categories may privilege vehicle-centric metrics while underrepresenting pedestrians, cyclists, wheelchair users, and children. These gaps are not merely technical; they reflect choices about funding, access, privacy, and the perceived value of different road users. Curation therefore becomes an ethical and political act as much as an engineering task.

The socio-technical context also includes organizational incentives. Data collection is expensive, labeling is labor-intensive, and deployment timelines reward rapid iteration. This creates pressure to reuse existing datasets, automate annotation, and defer hard questions about bias and robustness. Technical debt accumulates when data pipelines are patched without documentation or validation [16]. Human-centered design approaches emphasize that reliable systems require transparency, contestability, and meaningful human control [17]. In trajectory prediction, these principles imply that curation tools should help reviewers understand why a scenario was labeled, how confident the system is, and what alternative interpretations exist. Vision-language assistance can support this goal only if it is embedded in workflows that preserve human oversight and make uncertainty visible.

3. Data Curation Challenges for Trajectory Prediction

The first major curation challenge is semantic grounding. Trajectory data often records positions and velocities but not the social meaning of an interaction. Was a vehicle yielding, hesitating, or preparing to turn? Was a pedestrian waiting, jaywalking, or responding to another road user? These distinctions matter for prediction because similar geometric paths can arise from different intentions. Vision-language models can generate rich descriptions of scenes, propose labels for maneuvers and interactions, and link visual cues to textual categories. However, semantic grounding requires validation because language can be ambiguous and culturally situated. A term such as aggressive driving may have different meanings across regions, road types, and observer perspectives. Automated labels must therefore be treated as hypotheses that require calibration against human annotation protocols and domain expertise.

The second challenge is distributional imbalance. Trajectory datasets are dominated by ordinary, low-risk scenarios, while rare events such as near misses, sudden pedestrian crossings, or unusual vehicle interactions are precisely the cases that matter for safety. A density-balanced curation and evaluation framework for trajectory prediction has been proposed to address long-tail imbalance and evaluation fragmentation [15]. This work illustrates a broader lesson: curation must actively seek, preserve, and weight rare scenarios

rather than merely sampling more data. Vision-language models can help retrieve rare events by searching for textual descriptions, clustering visually similar anomalies, and generating candidate scenario tags. Yet retrieval depends on the vocabulary and priors of the model. If the model has not been exposed to a particular environment or road-user behavior, it may fail to recognize it as anomalous. Curation pipelines must therefore combine automated discovery with community knowledge, field audits, and targeted data collection.

The third challenge is cross-modal consistency. Trajectory prediction systems often rely on camera, lidar, radar, and map data that can disagree because of calibration errors, synchronization issues, occlusion, or sensor degradation. Vision-language models can compare textual scene descriptions against visual evidence, flag inconsistencies, and prompt human review. This is valuable because inconsistencies that remain hidden during training can become catastrophic during deployment. However, cross-modal consistency checking introduces its own risks. A vision-language model may confidently describe an object that is not present, or it may miss a small but critical road user. When such outputs are used to discard or relabel data, they can silently shrink the effective training distribution. Robust curation therefore requires uncertainty quantification, provenance tracking, and escalation paths for ambiguous cases.

The fourth challenge is evaluation fragmentation. Different datasets use different coordinate systems, map formats, prediction horizons, and metrics. This makes it difficult to compare systems and to understand whether improvements generalize. Vision-language assistance can help harmonize metadata by generating natural-language summaries of scenarios, identifying equivalent interaction types across datasets, and supporting semantic search. But harmonization can also erase meaningful differences. A unified label such as unprotected left turn may cover radically different intersection designs and traffic rules. Curation must preserve local context while enabling cross-dataset analysis. The goal is not a single universal vocabulary but an interoperable set of descriptions that remain traceable to their original sources.

4. Vision-Language Assistance as Curation Infrastructure

Vision-language models offer several capabilities that are directly relevant to curation. They can generate captions, answer questions about scenes, retrieve images or trajectories from natural-language queries, classify interactions, and propose structured metadata. In a curation pipeline, these capabilities can be placed at multiple points: during ingestion to create initial tags, during review to surface uncertain cases, during dataset construction to balance scenario categories, and during evaluation to generate diagnostic reports. The value of vision-language assistance is greatest when it reduces the cognitive burden of routine labeling while preserving human attention for genuinely ambiguous or high-stakes cases. This division of labor is consistent with human-centered artificial intelligence, which emphasizes reliable automation, human control, and clear accountability [17].

However, treating vision-language models as infrastructure rather than as tools changes the design requirements. Infrastructure must be maintainable, versioned, and governed. It must handle changes in model versions, data schemas, and organizational priorities. If a vision-language model is updated, its outputs may shift even when the input data does not. A curation system that depends on such a model must document version changes, revalidate affected labels, and provide rollback mechanisms. These requirements resemble the hidden technical debt that accumulates in machine learning systems when data dependencies, configuration, and feedback loops are not managed explicitly [16]. Without such discipline,

vision-language assistance can create a false sense of automation while increasing long-term maintenance costs.

Privacy and access are also central. Trajectory data can reveal individual movement patterns, home locations, workplace routines, and social relationships. Vision-language models may generate descriptions that combine visual and textual cues in ways that make re-identification easier. Curation infrastructures must therefore incorporate privacy-preserving practices such as aggregation, differential privacy, access control, and purpose limitation. They must also address the labor conditions of annotators and reviewers, who may be asked to evaluate disturbing crash scenarios or to apply ambiguous guidelines at high speed. Ethical curation requires fair compensation, mental health support, and channels for workers to contest automated labels. These concerns connect to broader debates about the ethics of algorithms and the distribution of power in data-intensive systems [20].

5. Architecture and Structural Trade-offs

A vision-language assisted curation architecture must balance centralization and federation. A centralized pipeline can enforce consistent standards, reduce duplication, and simplify auditing. A federated pipeline can respect local data sovereignty, accommodate regional differences, and reduce the risk of concentrating sensitive data. In practice, hybrid architectures are likely: a central governance layer defines schemas, evaluation protocols, and audit requirements, while local nodes perform collection, annotation, and model adaptation. Vision-language models can operate at both levels. Central models can provide general-purpose retrieval and labeling, while local models can be fine-tuned for regional road culture, language, and sensor configuration. This approach mirrors infrastructure studies, which show that large information systems succeed when they balance standardized interfaces with local flexibility [18].

Another trade-off is between automation and interpretability. Highly automated curation can process enormous volumes of data, but it may produce labels whose rationale is opaque. Interpretable curation may require slower, more expensive human review, but it can generate better documentation and stronger accountability. A practical architecture might use vision-language assistance to rank cases by uncertainty, novelty, or potential safety impact, and then route only the most consequential cases to human experts. This tiered approach can improve efficiency while maintaining quality. It also creates a need for audit trails that record which model version, prompt, threshold, and reviewer contributed to each decision. Such records are essential for internal algorithmic auditing and external accountability [19].

A third trade-off concerns evaluation. Next-generation trajectory prediction systems must be evaluated not only on average displacement error but also on rare-event recall, calibration, interaction awareness, and fairness across road-user groups. Vision-language assistance can support richer evaluation by generating scenario-specific test suites, summarizing failure modes, and comparing predicted behaviors against textual norms. However, evaluation itself can become a target for optimization. If curators know which scenarios are used for evaluation, they may inadvertently tune the data pipeline toward those cases. Evaluation datasets should therefore be protected, rotated, and supplemented with hidden test sets. Vision-language models can help create synthetic or augmented scenarios, but synthetic data must be clearly labeled and validated to avoid feedback loops that degrade real-world performance.

6. Governance, Fairness, and Policy

Governance is the set of rules, roles, and processes that determine how curation decisions are made and who is accountable for them. In trajectory prediction, governance must address data collection consent, annotation guidelines, model selection, bias testing, incident reporting, and redress. Vision-language assistance adds new governance questions. Who is responsible when a vision-language model mislabels a pedestrian and that error propagates into a safety-critical system? How can affected communities challenge a label that misrepresents their behavior? What documentation is required when a model version changes? These questions cannot be answered by technical metrics alone. They require policy frameworks that connect data curation to broader regulatory expectations for artificial intelligence, including transparency, risk management, and human oversight.

Fairness is both a technical and a social property. Trajectory prediction systems can disadvantage certain road users if training data underrepresents their behavior or if evaluation metrics prioritize vehicle flow over pedestrian safety. Vision-language models can help identify underrepresented groups by searching for rare attributes in image and text data, but they can also reproduce stereotypes embedded in their training corpora [12]. Fairness-aware curation therefore requires disaggregated evaluation, participatory design, and continuous monitoring. It is not enough to include more examples of a group; the examples must be contextually rich and free from biased labels. For instance, adding more pedestrian trajectories from a single intersection does not address the diversity of pedestrian behavior across ages, abilities, and cultural contexts.

Policy implications extend to deployment and liability. Regulators may require evidence that trajectory prediction systems were trained on representative data and evaluated against relevant edge cases. Insurers may require documentation of known limitations. Municipalities may require privacy impact assessments before data collection. Vision-language assisted curation can support these requirements by producing structured metadata and audit trails, but only if the outputs are reliable and contestable. A governance framework should therefore mandate versioning, human review for high-stakes labels, external audits, and public summaries of data limitations. It should also clarify that automation does not transfer responsibility away from the organizations that deploy the system.

7. Deployment, Sustainability, and Robustness

Deployment environments are dynamic. Roads change, traffic laws evolve, vehicle fleets turn over, and sensor suites age. A curation infrastructure that is designed only for initial training will quickly become obsolete. Sustainable curation requires continuous data collection, drift detection, re-annotation, and model retraining. Vision-language models can assist by monitoring incoming data for semantic drift, flagging scenes that do not match existing categories, and generating summaries for operations teams. However, continuous curation also increases energy consumption, storage costs, and organizational complexity. Sustainability must therefore be considered alongside accuracy. Not every data point needs to be stored forever, and not every scene needs to be processed by a large vision-language model. Tiered storage, sampling strategies, and model distillation can reduce environmental and economic costs.

Robustness is closely tied to curation. A trajectory prediction system may fail when it encounters adversarial behavior, sensor occlusion, unusual weather, or rare road geometry. Vision-language assistance can improve robustness by helping curators identify gaps in coverage and by generating stress tests that target weak spots. Yet robustness cannot be guaranteed by data alone. Systems must also include runtime monitoring, fallback behaviors,

and safe-stop capabilities. Curation should provide the evidence needed to validate these safeguards. For example, if a system claims to handle pedestrian-heavy environments, the curation record should show that such environments were sampled, annotated, and evaluated with appropriate metrics. Without this linkage, robustness claims remain aspirational.

8. Evaluation and Case Illustrations

Consider an autonomous shuttle deployed on a university campus. The initial dataset may contain many vehicle and bicycle interactions but few encounters with wheelchair users, service animals, or children. Vision-language assistance can search campus camera footage and generate candidate scenes for annotation. Human reviewers then validate the scenes and add contextual notes about crossing behavior, visibility, and social norms. Over time, the curation infrastructure monitors for new edge cases and updates the evaluation suite. This illustrates how vision-language assistance can support a continuous improvement cycle. It also shows why governance matters: who decides which edge cases are important, how are reviewers trained, and how are privacy concerns addressed.

A second case involves a fleet operating across multiple cities. Regional driving cultures differ in ways that are difficult to capture with fixed labels. Vision-language models can generate descriptions of local behavior and help compare scenarios across regions. However, automatic harmonization may flatten meaningful differences. A better approach is to maintain local taxonomies linked through a shared metadata layer, allowing cross-city analysis without erasing context. This requires investment in documentation, version control, and inter-organizational agreements. It also requires evaluation metrics that can detect whether a model trained in one region transfers fairly to another.

A third case concerns incident response. When a prediction failure occurs, investigators need to reconstruct the data provenance, annotation history, model version, and deployment context. A well-designed curation infrastructure can provide this evidence quickly. A poorly designed one may leave engineers unable to determine whether the failure originated in the sensor, the label, the model, or the deployment environment. Vision-language assistance can help by generating natural-language summaries of the incident and retrieving similar cases, but these outputs must be treated as investigative aids rather than definitive conclusions. The ultimate goal is accountability, not just explanation.

9. Future Directions

Future research should develop curation architectures that are modular, auditable, and participatory. Modularity allows components such as vision-language labeling, human review, and evaluation to be replaced or upgraded without disrupting the entire pipeline. Auditability ensures that every label and split can be traced to its source and rationale. Participatory design brings affected communities into decisions about what data is collected and how behavior is categorized. These directions connect technical design to democratic governance. They also suggest new benchmarks for curation quality, including measures of provenance completeness, label contestability, and coverage of underrepresented scenarios.

Another direction is the development of uncertainty-aware vision-language models for curation. Such models should communicate when they are unsure, when their training data is insufficient, and when human review is necessary. They should also be evaluated for calibration, robustness, and bias in curation-specific tasks. Finally, the field needs shared infrastructure for documenting and sharing curation workflows. Standards analogous to datasheets and model cards can be extended to curation pipelines, describing data sources,

annotation protocols, model assistance, and known limitations [13], [14]. These standards would help regulators, researchers, and operators compare systems and identify risks before deployment.

10. Conclusion

Vision-language assisted data curation offers a promising path toward more reliable, representative, and adaptable trajectory prediction systems. It can improve semantic labeling, rare-scenario discovery, cross-modal consistency, and evaluation design. Yet these benefits depend on how assistance is integrated into the broader socio-technical system. If vision-language models are treated as black-box replacements for human judgment, they may introduce new biases, obscure provenance, and create unsustainable dependencies. If they are treated as components of a governed curation infrastructure, they can augment human expertise while preserving accountability. The difference lies in architecture, documentation, oversight, and policy.

Next-generation trajectory prediction systems require curation that is continuous, fair, robust, and sustainable. This means investing in data pipelines as long-term institutions rather than temporary preprocessing steps. It means measuring quality not only by scale but by coverage, provenance, and contestability. It means connecting curation decisions to deployment monitoring and incident response. It also means recognizing that data about human movement is deeply social and must be governed with care. Vision-language assistance can support these goals, but only when it is designed with humility, transparency, and a commitment to the publics affected by prediction technologies. The future of trajectory prediction will be shaped as much by curation infrastructures as by model architectures, and the field should plan accordingly.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.
3. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 961–971). IEEE.
4. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2255–2264). IEEE.
5. Salzmann, T., Ivanovic, B., Chakravarty, P., & Pavone, M. (2020). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision – ECCV 2020* (pp. 683–700). Springer.
6. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., & Urtasun, R. (2020). Learning lane graph representations for motion forecasting. In *Computer Vision – ECCV 2020* (pp. 541–556). Springer.

7. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11621–11631). IEEE.
8. Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., & Hays, J. (2019). Argoverse: 3D tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8748–8757). IEEE.
9. Zhan, W., Sun, L., Wang, D., Shi, H., Clausse, A., Naumann, M., Kummerle, J., Konigshof, H., Stiller, C., de La Fortelle, A., & Tomizuka, M. (2019). Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*.
10. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., ... Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2446–2454). IEEE.
11. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
12. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
13. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.
15. Yang, R., Xu, Y., Fu, Y., & Su, L. (2026). Den-tp: A density-balanced data curation and evaluation framework for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10632-10641).
16. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
17. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504.
18. Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7(1), 111–134.

19. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). ACM.
20. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.