Exploring Practitioner Perspectives On Training Data Attribution Explanations

Elisa Nguyen
Tübingen AI Center
University of Tübingen

Evgenii Kortukov
Tübingen AI Center
University of Tübingen

Jean Song
DGIST

Seong Joon Oh
Tübingen AI Center
University of Tübingen

Abstract

Explainable AI (XAI) aims to provide insight into opaque model reasoning to humans and as such is an interdisciplinary field by nature. In this paper, we interviewed 10 practitioners to understand the possible usability of training data attribution (TDA) explanations and to explore the design space of such an approach. We confirmed that training data quality is often the most important factor for high model performance in practice and model developers mainly rely on their own experience to curate data. End-users expect explanations to enhance their interaction with the model and do not necessarily prioritise but are open to training data as a means of explanation. Within our participants, we found that TDA explanations are not well-known and therefore not used. We urge the community to focus on the utility of TDA techniques from the human-machine collaboration perspective and broaden the TDA evaluation to reflect common use cases in practice.

1 Introduction

The suite of explainable AI (XAI) encompasses models and explanation methods that aim at uncovering the rationale behind black-box model behaviour for humans [1]. XAI methods are usually attribution methods, which can be categorised into feature and instance attribution. While the former finds explanations for model predictions within the features of an input (e.g. SHAP [2]), the latter explains model predictions at the instance level (e.g. Influence functions [3]).

This study focuses on an instance attribution approach called training data attribution (TDA). TDA gives insight by attributing model behaviour to training samples [4, 5]. The ground truth attribution of the model prediction on test sample z to a training sample z_j is the change in loss after leave-one-out retraining:

$$\mathrm{TDA}(z_j, z) := \mathcal{L}(z; \theta_{\setminus j}) - \mathcal{L}(z; \theta) \tag{1}$$

where the model parameters θ are trained with the loss $\mathcal{L}$. As such, TDA views the model as an output of the learning algorithm and attributes model behaviour to parts of the training set.

Explanations of machine learning (ML) models are sociotechnical in nature [6]. Efforts in human-centred XAI emphasise this side of XAI and aim at a deeper understanding of the explainee because it is essential for the effective adoption of XAI in practice [7]. Yet, we find that the human factor of XAI is underexplored for TDA.

To address this gap, we present a qualitative interview study with ML practitioners in application areas of high-risk systems according to [8] (e.g. healthcare, employment, law enforcement). ML applications in such areas will require assessment throughout their product lifecycle. We therefore expect XAI to be particularly relevant in such areas.

By interviewing practitioners, we take a human-centered perspective which gives us an impression of how ML models and explanation methods are put into practice and how practitioners view

Preprint. Under review.

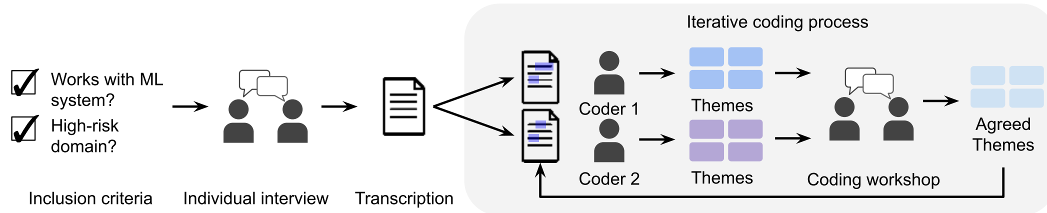

Figure 1: Interview and data analysis process.

the idea of TDA. Through an inductive thematic analysis, we find: (1) End-users are interested in training data attribution methods that could facilitate human-machine collaboration. Model developers find value in methods that enable them to improve the dataset quality. (2) Though the idea of TDA is generally positively perceived, within our participant pool, TDA is not utilised. XAI tools are only used as out-of-the-box functionality. We therefore anticipate that TDA tools can deliver practical value if they are easy to implement.

2 Related Work

Interview studies provide insights into human factors in explainable AI (XAI) and can inform the design of human-centred XAI technology [9]. Previous work has conducted semi-structured interviews with XAI practitioners of different technical expertise to study how people understand the problem of XAI [10], people's preferences regarding interactivity of explanations [11] and user needs, intentions and perceptions of explanations [12]. They found that user needs and XAI vocabulary vary across users [10] but interactivity [11] and actionability [12] are desired. These studies result in concrete recommendations about XAI in practice, i.e. a call for consistent vocabulary to facilitate clear communication and progress in XAI [10], the case for interactive dialogue systems [11] and the need for considering an explanation's actionability in the design process [12]. However, they base their studies mainly on feature attribution explanations while our work focuses on training data attribution (TDA) explanations. We therefore expand on existing literature about user perspectives on XAI.

TDA aims at capturing a training sample's attribution to a model decision on a test sample. Several methods exist [3, 13–16] which focus on accurately approximating the counterfactual change in a model's output on a test sample when a training sample is removed from the dataset (cf. Eq. 1). Applications of TDA methods are focused on topics from data-centric AI i.e. aiming at model improvement by improving the data (e.g. cleaning faulty data labels [17] or detecting model biases [18, 19]). We find that the study of user needs and perspectives is underexplored for TDA. Our study presents a first step in addressing this gap.

3 Interview methodology

This study aims to explore practical perspectives on training data attribution (TDA). Since we study subjective experiences, we opt for a qualitative analysis through interviews. We conducted semi-structured interviews to balance the interview structure and the freedom of conversational flow [20] and analysed the transcripts in an inductive thematic analysis (cf. Figure1). [1]

Participants. We define inclusion criteria to ensure participants align with our research aims: They should (1) have at least one month of experience in working with ML systems and (2) either work in a high-risk application area [8] (e.g. health care, law enforcement, employment. Full list in Appendix A). This criterion serves to focus our studies on application areas that are likely to be subject to further regulations and governance in the future [1, 21]. Recruiting participants poses a challenge, especially in high-risk application areas. Hence, we use purposive sampling [22] and approach potential participants from the authors' network individually. We recruit 10 participants from diverse domains and degrees of experience (cf. Table 1).

[1] The IRB approval process is currently ongoing. We expect a decision in November 2023.

Table 1: Participant information. HR = Human resources, AV = autonomous vehicles, TC = telecommunications, CV = Computer vision for automation. P5 did not meet the inclusion criteria.

ID	Country of work	Domain	Type	Job experience/with ML	Type of ML
P1	Germany	HR	End-user	3 yrs/1 mo.	Chatbot
P2	USA	AV	Developer	2 yrs/7 yrs	Prediction model
P3	Netherlands	TC	Developer	3 yrs/5 yrs	Prediction model
P4	Finland	CV	Developer	4 yrs/6 yrs	Prediction model
P6	Switzerland	Health	End-user	2 yrs/2 yrs	Prediction model
P7	Netherlands	Health	Developer	1 yr/3 yrs	Prediction model
P8	Belgium	Health	Developer	2 yrs/6 yrs	Prediction model
P9	Pakistan	Health	Developer	5 yrs/2 yrs	Prediction model
P10	Germany	HR	End-user	3 yrs/1 yr	Chatbot
P11	Germany	Health	End-user	10 yrs/6 yrs	Clustering, Chatbot

Interview process. The interviews were conducted during June - September 2023, either in person or remotely via video call. All interviews are one-on-one conversations in English, except with P10 in German. The participants were first briefed on the objective of the study and data processing using the informed consent form (cf. Appendix B). Upon receiving informed consent, we started the interview recording. Overall, the interviews lasted between 30 to 60 minutes. In each interview, the following topics are addressed (full interview guide in Appendix C):

- **Job-related information.** Perspectives may vary between different domains and levels of seniority as well as experience with the ML tool.
- **Interviewee's workflow with ML systems.** By asking about the workflow with the ML tool, we wish to understand the patterns of usage and challenges participants encounter.
- **Perspectives on training data.** Since we investigate TDA explanations, we explicitly ask participants about the role training data plays in their tasks.
- **Perspectives on data-driven XAI.** We address the participant's perspectives on XAI and particularly on TDA.

Interview transcription. The interviews are first transcribed automatically using Whisper [23] and then cleaned up manually. The transcript is then pseudonymised. We translated P10's German transcript to English using DeepL [24].

Analysis. We analyse the transcripts through an inductive thematic analysis by two coders (cf. Figure 1). The analysis is iterative: The interview transcript of P1 is first analysed jointly in an initial coding workshop. Afterwards, coders independently code five transcripts, extending on the themes and codes found in the initial analysis. During an intermediate coding workshop, agreements and disagreements between the coder's themes and codes are discussed. The workshop resulted in a new, merged definition of themes and codes which are used for the remaining transcripts. At the intermediate coding workshop, the interrater agreement is 77.3% measured by the percentage of agreement participants coded to themes. The final coding workshop serves the same purpose - after both coders reviewed the remaining transcripts, the overlap and gaps are discussed and the final themes are agreed upon. The final interrater agreement is 80.3%. Full analysis instructions in Appendix D.

4 Findings

The result of the thematic analysis is shown in Figure 2. We identified six main themes which are related to the current use of ML systems, perspectives towards explainable AI (XAI) and training data attribution (TDA). Two groups of interviewees have provided noticeably different perspectives - end-users and model developers. We thus discuss their perspectives separately.

4.1 End-user perspective

An end-user makes use of ML tools and is not involved in the model-building process. We find that end-users often face challenges related to trust calibration when using ML systems and identify a possible use of TDA in facilitating human-machine collaboration.

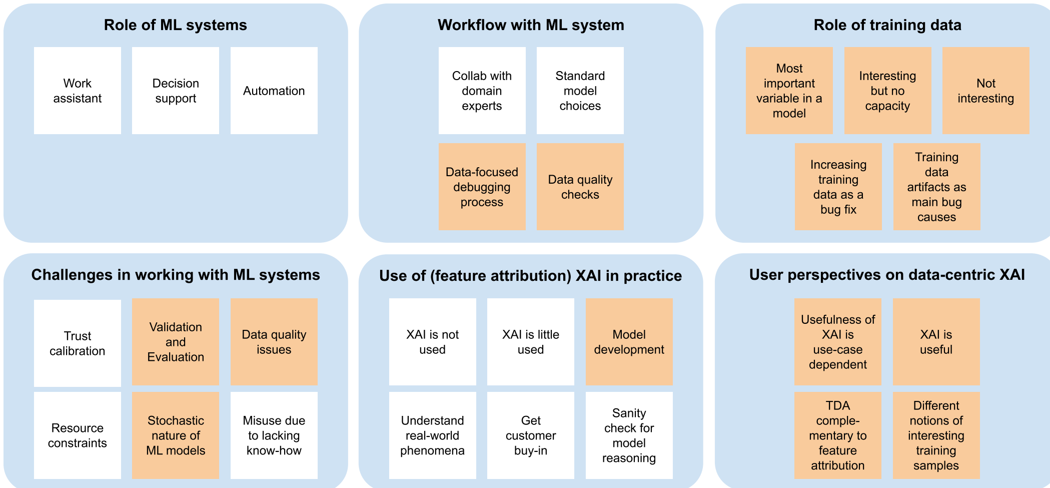

Figure 2: Theme overview as a result of the thematic coding process. Themes directly related to training data and TDA are highlighted in orange.

Role of ML system. End-users use ML systems for work assistance and decision support. Chatbots generally fill the role of a work assistant that "[takes work off of the participant's hands and makes their work easier]" (P10) and is "available around the clock" (P10). Participants use chatbot systems to improve their writing in English (P1, P11), to search for information where previously they would "[ask] Google" (P1), or to ideate research ideas (P11). Moreover, P10 trusts their company-internal chatbot enough to redirect simple employee questions to the chatbot. As decision support systems, ML systems deliver information that acts as a basis for decisions taken by end-users, e.g. diagnosis support (P6).

Workflow with ML system. End-users rely on ML systems when they deliver helpful suggestions. If the ML system generates unhelpful results (P6, P10), users take over and turn away from ML support. We also find that ML systems often lack feedback loops, particularly when ML systems are purchased as a product from the market, leading end-users to voice concerns mainly when bugs accumulate (P6).

Role of training data. End-user participants were unaware of which training data may have been used to train the models (P1, P10). P6 (a medical doctor) mentions that they would be curious about training data but "[it] is a luxury that [requires time]", highlighting the practical constraint of time pressure.

The greatest challenge of ML system end-users is trust calibration. Our findings agree with Kim et al. [12]: It is unclear how much and when a system can be trusted. P1 sometimes finds themselves in a dilemma in which they wish to learn something from the chatbot, but are unable to calibrate their trust in the response due to missing knowledge: "I don't know everything regarding this topic. I [don't even] know what he's replying to me." (P1). P10 also mentions inadequate know-how in ML system usage as a challenge: "the employees often don't manage to ask the chatbot the right questions".

Use of XAI in practice. Not all participants have used XAI. Some participants were unaware of XAI since explanations are not a part of the ML tools they use (P1, P10, P11). P6 reports that XAI tools they used in radiology images so far (i.e. heatmaps) do not deliver a full answer to the why question, as counterfactual information is missing: "[If] I just get an overall highlight in these basal lung regions and the prediction that is atelectasis, I still don't know why this is atelectasis but not pleural effusion or consolidation." Moreover, P6 highlights their time constraints: They would hardly be able to look at explanations even when available. Therefore, the end-user's challenges in using XAI are not only a lack of awareness and availability but also limited time.

Perspectives on data-centric XAI. End-users are not familiar with the idea of TDA explanations. When asked for their opinion about the concept of TDA, chatbot users (P1, P10, P11) were interested in training samples which give additional information related to their request or samples which help them improve their interaction with the model and ask better questions ("if [the chatbot] can also sometimes formulate: [for stupid questions] you [cannot expect a good answer because] [...] then maybe you understand [how to ask the question better] and can ask it more precisely again" - P10). P6 would be interested in samples similar to the test sample to calibrate their trust. P6 also emphasised that explanations can only be helpful if there is time to spend on an explanation. Back-and-forth interactions with the system are "absolutely unrealistic" (P6). The above findings agree with the insights in Kim et al. [12]: End-users want explanations that help them improve collaboration with the ML system. End-users wish to overcome the challenge of trust calibration and showed a positive sentiment towards the idea of TDA.

4.2 Model developer perspectives

Model developers are concerned with the building of ML systems. We find that model developers often face challenges related to data quality and identify potential use cases for TDA.

Role of ML systems. Model developers work on decision-support (P3, P5, P7, P8, P9) and automation systems (P2, P4). They build ML systems according to the customer's needs. P3 uses ML systems to identify and explain the contributing factors to product issues: "If we can predict it, we can also have an idea what are the factors mostly creating this phenomenon."

Workflow with ML systems and the role of training data. Developers and end-users collaborate closely for building and evaluating ML models, where bugs are reported to the developers by the end-users (P2, P3, P4, P9). This shows a clear separation of domain knowledge: "Because personally, I cannot know if the model is doing the correct thing [...] business have to tell me" (P3). The model-building workflow is focused on data and developers spend a considerable amount of time with data curation (P2, P4, P3, P7, P8, P9). Participants explicitly stated that they use standard model architectures and the majority of the work is dataset curation (P3, P4, P8): "[What] drives your model is your data. [...] [If] it's already an established problem, you're probably not going to do better than an algorithm that's already been laid out to solve that problem for you." (P8). Data quality checks are a set part of the data preprocessing pipeline (P2, P3, P4, P8). P2 and P9 reported that they first assess data quality before inspecting the model in debugging. Furthermore, P2 explained that collecting more data is a common way to overcome model shortcomings in autonomous driving. This shows that development work in practice is centred around data. Consequently, model developer participants consistently view training data as the most important variable in a model, e.g. "[we] [...] believe that [...] the models can only be as good as the [...] data that you feed in." (P4).

Challenges in working with ML systems. Data quality issues are often the root cause of model malfunction. Participants report distribution shifts (P3, P4), data collection artefacts like missing data or labels (P2, P4, P3, P7, P8, P9), wrong labels (P2, P4), wrong data formats due to aggregation of different data sources (e.g. dates being interpreted as integers or wrong ordering of temporal data) (P8, P9), and historical data (P9). Issues with data quality impact model validation; for example, participants encounter difficulties due to absent labels. Furthermore, P2 mentions that the validation itself is a challenge due to multiple requirements that the ML system should fulfil. P2 also sees a challenge in the stochastic nature of ML models: "[The] same data set, same model, you train multiple times, you can get [different] results." In addition, memory and compute constraints are relevant to P2 and P4 as they work with ML systems on the edge. Our analysis shows that data plays a substantial role both in the challenges faced by model developers and in the development process itself.

Use of XAI in practice. Participants use XAI for different purposes, most commonly as a tool for model development (P2, P3, P8). As such, XAI tools offer explanations for per-example debugging of e.g. wrong predictions or act as a sanity check for model reasoning. Furthermore, P8 states that XAI tools are useful in getting customer buy-in and convincing the customers of the model's decision suggestion. P3 described the use of XAI as a tool to understand phenomena represented by the ML model: "[Building] the model, the whole purpose is to get some explainability. Because [...] we know that [a problem is] happening and predicting doesn't really add value. But if we can predict it, we can also have an idea what are the factors mostly creating this phenomenon." While XAI and therefore explanations have different purposes, we note that participants use XAI tools mainly as an out-of-the-box functionality. P3 and P8 reported using a SHAP [2] library, whereas P2 visualises attention maps. We find that implementation thresholds must be low for the adoption of XAI in practice.

Perspective on (data-centric) XAI. Within our participants, we find that model developers are not familiar with TDA explanations. However, when asked about their intuition on what important training data could be, participants talked about out-of-distribution samples (P3, P8), mislabelled samples (P2), and samples close and far from the model's decision boundary (P7, P8). Developers seek to understand the data distribution and find ways to improve the data quality, and participants are interested in how TDA enables this. However, some participants specified that the usefulness of XAI depends on certain conditions: P3 and P8 who use explanations to present models to their business, state that in their experience, model performance must be high for explanations to serve their purpose. Additionally, P8 mentions that finding an individual training sample is unlikely to be informative in a large dataset and relevant data on a "collection level" would be more interesting. Our analysis shows that the idea of TDA is positively perceived by model developers. Furthermore, TDA as a data-centric XAI approach could fit well into the work of a model developer, which is strongly centred around the data itself.

5 Implications for future TDA research

Status quo of TDA research. Training data attribution (TDA) explains model behaviour by finding relevant training data to a model prediction, where "relevant" is defined by the change in loss after leave-one-out retraining (LOO) (cf. Eq. 1) [4, 5]. As mentioned in section 2, recent TDA research is focused on studying efficient and accurate approximations of Eq. 1 (e.g. [16]) or the application of TDA methods to particular use cases in data-centric AI (e.g. [18]). The human factor in TDA is underexplored and our study takes a first step in addressing this gap.

Some of the ideas from our study are actively researched. Our analysis of participants' ML workflow and perspectives on XAI has shed light on the required features for TDA methods. Some are being actively studied in the community: P8 mentions that the attribution of a single training sample is unlikely to be informative, which has been studied in e.g. [25, 26] and motivated TDA approaches like [27, 28]. Also, model developers' intuition that mislabeled data are important training data is addressed in TDA research through existing evaluations using mislabel identification tasks as in Koh and Liang [3].

Some are yet to be studied further. Other perspectives could add to TDA research: Participants mention several types of data quality issues beyond mislabels, such as missing data (P3, P8, P9), wrong data formats (P8, P9), distribution shifts (P3, P4), which are currently not often considered in evaluation. Furthermore, questions related to TDA in human-machine collaboration, like interaction and usability (P1, P6, P10, P11), are not explored in TDA research.

Future directions in TDA research. It is important to consider the user and human factors in the development of XAI technology like TDA, whether it addresses model developers or end-users [6]. We find that participants are generally unaware of TDA and therefore do not apply it even in suitable use cases. To improve accessibility, TDA researchers should understand and address user needs better. This includes, for example, expanding the current evaluation practices to cover diverse use cases. Practical constraints like time pressure (P6) and low implementation thresholds (P3, P8) should also be actively formulated as one of the research goals in the future.

6 Conclusion

We present a qualitative interview study with ML practitioners from various high-risk application areas to investigate the human factor of training data attribution (TDA) explanations. Through an inductive thematic analysis, we find that priorities and perspectives differ between end-users and developers but the idea of gaining insights into the model through training data is positively perceived overall. Our research reveals possible research directions in TDA to bridge the gap from research to practice: TDA for human-machine collaboration and expanding the evaluation of TDA to diverse data-centric use cases. Further, we highlight that simple and intuitive implementations of TDA methods are key.

Acknowledgments and Disclosure of Funding

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Elisa Nguyen. This work was supported by the Tübingen AI Center.

References

[1] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[3] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

[4] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *ArXiv*, abs/2212.04612, 2022. URL https://api.semanticscholar.org/CorpusID:254535627.

[5] Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962.

[6] Upol Ehsan and Mark O. Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones, editors, *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pages 449–466, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60117-1.

[7] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, 2021.

[8] European Parliament. Ai act, annex iii, 2023.

[9] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376590. URL https://doi.org/10.1145/3313831.3376590.

[10] Andrea Brennen. What do people really want when they say they want "explainable ai?" we asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–7, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3383047. URL https://doi.org/10.1145/3334480.3383047.

[11] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv e-prints*, art. arXiv:2202.01875, February 2022. doi: 10.48550/arXiv.2202.01875.

[12] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581001. URL https://doi.org/10.1145/3544548.3581001.

[13] Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. FastIF: Scalable influence functions for efficient model interpretation and debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10333–10350, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.808. URL https://aclanthology.org/2021.emnlp-main.808.

[14] Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8179–8186, Jun. 2022. doi: 10.1609/aaai.v36i8.20791. URL https://ojs.aaai.org/index.php/AAAI/article/view/20791.

[15] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.

[16] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023.

[17] Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, and Andrea Passerini. Interactive label cleaning with example-based explanations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=T6m9bNI7C__.

[18] Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron Wallace. Combining feature and instance attribution to detect artifacts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1934–1946, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.153. URL https://aclanthology.org/2022.findings-acl.153.

[19] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

[20] Omolola A. Adeoye-Olatunde and Nicole L. Olenik. Research and scholarly methods: Semi-structured interviews. *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY*, 4(10):1358–1367, 2021. doi: https://doi.org/10.1002/jac5.1441. URL https://accpjournals.onlinelibrary.wiley.com/doi/abs/10.1002/jac5.1441.

[21] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[22] Greg Guest, Emily E. Namey, and Marilyn L. Mitchell. Collecting Qualitative Data: A Field Manual for Applied Research. SAGE Publications, Ltd, 55 City Road, September 2023. doi: 10.4135/9781506374680. URL https://methods.sagepub.com/book/collecting-qualitative-data.

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

[24] Deepl translator. https://www.deepl.com/translator.

[25] Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xHKVVHGDOEk.

[26] Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. A bayesian perspective on training data attribution, 2023.

[27] Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32, 2019.

[28] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

A Inclusion criteria: High-risk application areas

We refer to the definition of high-risk application areas according to Annex III of the European Union's AI Act [8]. For readability, we include an overview:

- AI applications in products that require a specific level of safety:
  - Toys,
  - Aviation,
  - Cars,
  - Medical devices,
  - Lifts.
- Biometric identification and categorisation of natural persons.
- Management and operation of critical infrastructure.
- Education and vocational training.
- Employment, worker management and access to self-employment.
- Access to and enjoyment of essential private services and public services and benefits.
- Law enforcement.
- Migration, asylum and border control management.
- Assistance in legal interpretation and application of the law.

B Participant information and informed consent

A censored version of the participant information and informed consent form is attached in the following:

Participant information

Dear participant,

We are asking you to take part in a scientific study. You will find everything you need to know about the study in this information sheet for participants.

Please read this information carefully. If you have any questions, please do not hesitate to contact us.

Around 10 - 15 people are to take part in the study.

This study is planned and conducted by XXXX XXXX (XXXX@XXXX).

Participation in the study is voluntary. If you do not wish to participate or if you later withdraw your consent, you will not suffer any disadvantages as a result.

Why is this study being conducted?

The objective of this study is to ensure that research efforts in training data attribution methods for AI models are human-centered and focused on real use cases. These methods aim to find relevant training data that can explain the predictions and provide insights into the model's behavior. However, the theoretical definition of "relevant" may not align with practical needs.

In the field of explainable AI, researchers are actively working on providing explanations that help humans understand how the model works. Explanations come in different formats, but this study specifically focuses on relevant training data.

To gain a practical understanding of what constitutes "relevant" training data, we plan to conduct expert interviews. These interviews will involve individuals who work with or are familiar with AI systems, particularly in "high-risk" domains where transparency in the decision process is often mandated by legislation, such as the AI Act and GDPR.

What is the study process?

For this study, you will be interviewed on your current use of AI systems and your view on training data. The interview is semi-structured, meaning that some questions are prepared but the conversation can drift off into a natural conversation. The interview will take roughly 30 minutes to one hour. The interview will be recorded and afterward transcribed. After transcription, the original voice recording will be deleted for data protection reasons.

Would you like to learn more about the subject of the study?

If you would like to learn more about the field of explainable AI, the following material may be helpful:

- Machine Learning Explainability Workshop by Hima Lakkaraju at Stanford (https://www.youtube.com/playlist?list=PLoROMvodv4rPh6wa6PGcHH6vMG9sEIPxL)
- Interpretable Machine Learning book by Christoph Molnar, in particular, chapter 10.5 Influential Instances (https://christophm.github.io/interpretable-ml-book/influential.html)

Who can I contact in case of questions?

If you have any further questions, please contact:

XXXX XXXX
XXXX@XXXX

Information on data protection

In this study, XXXX XXXX (XXXX@XXXX) is responsible for data processing. The legal basis for processing is personal consent XXXX. The data will be treated confidentially at all times.

The data will be collected exclusively for the purpose of this study described above and will only be used within this framework.

The data collected also includes personal identifying data such as names and your voice.

All data by which you could be directly identified, e.g. your name or date of birth, are replaced by an identification code (pseudonymized). This makes it almost impossible for unauthorized persons to identify you.

The data will be stored at the XXXX.

We only keep the personal data for as long as it is required for the above-mentioned purpose. The data will be deleted at the latest after 10 years after the study.

For the study, we will conduct an interview with you. The interview will be recorded with a recording device. The audio recording of the interview is first stored at the XXXX. The interview will be written down verbatim (transcribed) within three months. The audio recordings are then deleted so that only the interview transcript exists. All information that would allow third parties to draw conclusions about you is changed in the transcription so that it is no longer possible to draw conclusions. The pseudonymized transcribed text is stored for 10 years in XXXX and then deleted.

We do not transfer the collected data to other institutions in XXXX.

Consent to the processing of your data is voluntary. You can revoke your consent at any time without giving reasons and without disadvantages for you. After that, no more data will be collected. The lawfulness of the processing is carried out on the basis of the consent until the revocation is not affected by this.

In the event of revocation, you can request the deletion of the collected data. The data can also be further used in the anonymized form if you agree to this at the time of your revocation.

You have the right to obtain information about the data, also in the form of a free copy. In addition, you can request the correction, blocking, restriction of processing or deletion as well as, if applicable, a transfer of the data.

In these cases, contact:

XXXX XXXX (XXXX@XXXX)

You also have the right to complain to any data protection supervisory authority.

You can reach the supervisory authority responsible for you at:

XXXX

Consent form

Consent to participation

I have been informed about the study by *XXXX XXXX*. I have received and read the written information and consent form for the above study. I was informed in detail in writing and verbally about the purpose and the course of the study, the opportunities and risks of participation, and my rights and obligations. I had the opportunity to ask questions. These were answered satisfactorily and completely. In addition to the written information, the following points were discussed:

____

____

My consent to participate in the study is voluntary. I have the right to withdraw my consent at any time without giving reasons and without incurring any disadvantages.

I hereby consent to participate in the above study.

____

Name of the participating person in block capitals

____

Place, date Signature of **participating person**

____

Name of the person providing the information in block capitals

____

Place, date Signature of **informing person**

Consent to data processing

The processing and use of personal data for the above-mentioned study will only take place as described in the information about the study.

I hereby consent to the processing of my personal data as described.

____

Place, date Signature of **participating person**

____

Place, date Signature of **informing person**

C Interview guide

C.1 Interview meeting process

1. Welcome the participant
   - Greeting.
   - Thank them for their time.
2. Briefing and informed consent.
   - Jointly discuss the participant information sheet and informed consent form.
   - Ask participants for any open questions.
   - Sign the informed consent form.
3. Conduct interview.
4. Closing.
   - Ask participants for any open questions.
   - Thank participant for their time.

C.2 Interview questions

Table 2: Prepared question areas and questions.

Question topic	Intention
Job-related information	
Domain of work	Different domains may have different needs.
Field of work	Different countries may have different requirements for transparency.
Years of expertise	The level of expertise in the task may affect the need for explanations.
Understanding the current process	
Purpose of ML model	Understand the downstream task/application domain.
Model functionality	Understand how the ML model assists the downstream task.
Model interaction	Understand the level of interaction with the system.
Evaluation of ML model	Relevant for ML developers; Understand requirements for ML models in practice and how much developers trust the evaluations.
Model malfunction process	Understand the current process of dealing with unexpected model behaviour and reasons for the process.
Perspectives on training data	
Type of training data	Understand the role of training data; Is the participant aware of the training data and process?
Training data quality	Understand state of real-world datasets.
Perspectives on TDA	
Participant opinion on relevant training data for unexpected model output	Relevant for model developers; Understand what artefacts participants usually encounter in datasets, what is important data and why.
Participant opinion on the helpfulness of knowing important data	Understand perspectives on helpfulness of data-centric information to the downstream task.
Participant opinion on TDA definition	Understand perspectives on TDA.

D Thematic analysis instructions

Instructions for data analysis

The data will be analyzed in an **inductive thematic analysis.** In this process, common themes are identified by the coder without any predefined themes. The objective is to identify common themes in participant's answers regarding:

The workflow with ML systems in practice: e.g.
- What role does the ML system play in their work?
- How do participants interact with the system?
- What are the objectives of the participant's workflows with the ML system?
- Do participants already work with explanations?

The expectations on data-driven explanations in practice: e.g.
- What is the role explanations in the participant's work?
- What are the participants' views on XAI in practice?
- What kind of training data is interesting to the participant (if it is interesting)?
- How do participants view TDA vs feature attribution?

Step 1: Reading and familiarization.

Read the transcript(s) carefully and if needed, also multiple times. The objective is to familiarize yourself with the answers given and perhaps get initial ideas about common themes. The data has been cleaned before (raw transcripts were formatted to separate utterances from interviewer and interviewee, "ums" and "ahs" were removed).

Step 2: Inductive coding.

There may be common themes in how participants work with ML systems and how they view training data as a means of explanation. In this step, you should:
- Identify themes from the transcripts and name them according to what you find,
- Provide quotes from the transcript that you find important for the themes.

A good way to do this could be the direct annotation of text with the topics.

Step 3: Repeat and converge.

From the granular topics you found, try to converge to <10 themes by finding abstractions that encompass different topics. Read through the transcripts again until you are happy with your code.

Please feel free to change codes from previous transcripts if you analyze the data sequentially if you think that a different code fits better. The end picture matters most.