Technical Considerations for XAI in AI Governance

Susanne Dandl*
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Zurich, Switzerland
susanne.dandl@uzh.ch

Fiona Katharina Ewald*
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, Germany
fiona.ewald@stat.uni-muenchen.de

Enrique Valero-Leal*
Department of Artificial Intelligence
Technical University of Madrid
Madrid, Spain
enrique.valero@upm.es

Bernd Bischl
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, Germany
bernd.bischl@stat.uni-muenchen.de

Kristin Blesch
Faculty of Mathematics and Computer Science
University of Bremen
Bremen, Germany
blesch@uni-bremen.de

Abstract

This paper highlights crucial technical considerations when applying explainable artificial intelligence (XAI) methods in AI governance to explain black-box supervised machine learning models. We emphasize that their application in AI governance involves technical nuances that, if overlooked, can yield misleading interpretations. We highlight key factors to consider in AI governance for a non-technical audience, using a conceptual example: Feature importance methods explain an AI model that automatically invites job interview candidates based on the applicant's CV. By highlighting common pitfalls, we aim to better align the demands of AI governance with XAI methods.

1 Introduction

AI governance is evolving rapidly, driven not only by recent advances in AI technology itself, but also by developments in regulatory and private AI governance frameworks (see [1] for a literature review), which emphasize principles such as transparency and explainability, e.g., as in the EU AI Act [2]. Addressing these concepts requires considering legal, ethical, and technical perspectives together. Explainable AI (XAI) methods are frequently proposed to support AI transparency [3].

*Shared first authorship.

1st EurIPS Conference (EurIPS 2025) Workshop: Beyond Regulation – Private Governance & Oversight Mechanisms for AI (PAIG Workshop).

However, the technical nuances of using XAI are underexplored in the context of AI governance frameworks [4, 5]. Some preliminary approaches discuss properties of XAI methods which may be useful to fulfill legal requirements on a conceptual level [6], but practical implementations are currently lacking. As a result, potential pitfalls when applying XAI methods might be overlooked.

As a group of method-oriented researchers working on XAI for supervised machine learning with tabular data, we aim to overcome this gap by framing technical considerations in our research area to non-technical readers working on AI governance. We illustrate crucial considerations and potential pitfalls when drawing on XAI for AI governance through the following example.

2 Conceptual example

Consider a company that uses an automated decision-making process to invite applicants to job interviews. Assume that a fitted supervised machine learning model automatically decides to "invite" or "reject" applicants based on several input variables (features) provided by the applicant's CV, such as "highest educational degree", "years of work experience", or "age".

The company aims to implement AI governance measures for this model, particularly because the system is classified as a high-risk AI system under Article 6 of the EU AI Act [2] (see also Annex III, point 4(a)). Striving to implement AI governance, particularly to fulfill the explainability requirements for high-risk systems stated in the EU AI Act Article 13, point 3(b), XAI may be used [3].

In our example, the company's main AI governance goal is to identify the aspects of the applicants' CVs that are most relevant for the model decision with XAI. Hence, methods that allocate importance values to features according to the degree of the model's reliance are of particular interest. The company faces the challenge of choosing the right feature importance (FI) method from the many proposed ones. Selecting an inappropriate FI method for a specific use case or misinterpreting its output can yield misleading conclusions [7], hence we outline a few exemplary problems below.

Surveying FI methods shows that many of them quantify how much the predictive performance of the AI model is influenced by a feature. A well-studied method is, e.g., permutation feature importance (PFI) [8], which shuffles a feature's values to destroy its predictive information. If this procedure leads to a drastic difference in the model's performance, one can conclude that the feature impacts the model to a major extent.

Let us assume that in our example, applying PFI indicates that "age" is the most influential feature. Concerned about potential age bias, the company consults an XAI expert and learns that the method reflects so-called marginal FI, which assesses FI independently, i.e., irrespective of other features. In contrast, conditional FI, e.g., as in [9–11], accounts for other features and their dependencies, revealing that given "highest educational degree", "age" no longer has a dominant importance. This seems reasonable since people with a higher educational degree typically access the job market at later ages. This dependency is captured by conditional FI, but not marginal FI.

Evidently, XAI methods, including FI methods, can yield considerably different insights; see details in [12]. From a technical perspective, differing methods' outputs are expected and legitimate, as they tackle slightly different questions [13]. A "disagreement problem" can also arise when the same method is applied multiple times with different random seeds (e.g., to shuffle features in PFI) or varying parameter settings. Such variability may, indeed, undermine user trust in the results [14].

Recapitulating our example, we can note another challenge, which arises because a large number of features can be derived from the applicants' CVs. Such high-dimensional data comes with multiple challenges for XAI methods: Many XAI methods are unsuitable for high-dimensional data, as they face computational constraints (e.g., when determining dependency structures for conditional FI) and interpretability challenges introduced by dimensionality reduction (e.g., when transforming features through principal component analysis) [15, 16].

Finally, it is challenging for the company to provide clear and careful communication of XAI results (in text or visuals) to avoid overwhelmed or misled audiences [7]. Especially the diverse audience of the XAI findings proves a challenge, since recipients may have different kinds of expertise, e.g., technical vs. legal.

3 Conclusion and Discussion

In this paper, we have touched on key considerations for using XAI methods in AI governance. We emphasize that explanations based on unnoticed technical nuances can lead to misguided interpretations, which contradicts the core principles of transparency and explainability. We therefore aim to promote awareness of the need to respect the underlying technicalities of XAI methods.

To facilitate navigating XAI methods for AI governance, regulators (e.g., from the EU) and XAI experts could co-produce a transparency template that companies can use for their AI models. This may circumvent the need for companies to consult XAI experts directly when trying to avoid the above-mentioned challenges, even though it remains a case-by-case decision to choose XAI methods. In particular, we emphasize that a collaboration between XAI researchers, private governance stakeholders and regulators may be particularly fruitful in drafting the transparency template. A starting point may be the use of model cards [17] to specify model details and extend them with information from XAI methods that increase model transparency [18]. Datasheets for datasets [19] can further extend the transparency template to give details on the data used during model training.

In addition to applying XAI methods, we want to stress that model builders should ensure sufficient model performance, alongside evaluating and reporting the model's uncertainty to offer a thorough depiction of the model. Furthermore, we encourage user studies to evaluate that the insights obtained from XAI methods align with the stakeholders' expectations of what "explainability" means; see, e.g., [20, 21].

We acknowledge the limitations of our reflections in this paper due to the fact that all authors are solely grounded in the technical field of XAI. Having cautiously engaged with legal and ethical aspects of AI governance, we found it challenging to have a technical interpretation of the transparency desiderata, e.g., in the EU AI Act, despite our expertise. This highlights the need for constant knowledge exchange between legal, ethical, and technical experts, resulting in clear guidelines, frameworks, and software tools to support the practical implementation of AI governance.

Acknowledgments and Disclosure of Funding

The project was initiated at the INTERACT workshop supported by the Munich Center for Machine Learning (MCML). KB declares to work on the development of lucin (`www.lucin.tech`), a planned XAI software startup. KB is funded by EXIST FKZ 03EGTHB007, a scholarship, which supports the transfer of research into commercial software products that is funded by the German Federal Ministry for Economic Affairs and Energy and the European Union through the European Social Fund Plus (ESF Plus).

References

[1] Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki. AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

[2] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/eu, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act) (text with EEA relevance), June 2024. Official Journal of the European Union.

[3] Georgios Pavlidis. Unlocking the black box: Analysing the EU artificial intelligence act's framework for explainability in AI. *Law, Innovation and Technology*, 16(1):293–308, 2024.

[4] Luca Nannini, Agathe Balayn, and Adam Leon Smith. Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1198–1212, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3593013.3594074.

[5] Joshua Krook, Peter Winter, John Downer, and Jan Blockx. A systematic literature review of artificial intelligence (AI) transparency laws in the european union (EU) and united kingdom (UK): A socio-legal approach to AI transparency governance. *AI and Ethics*, pages 1–22, 2025.

[6] Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F. Huber, and Christian Horz. How should AI decisions be explained? Requirements for explanations from the perspective of european law. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 438–450, 2024. doi: 10.1609/aies.v7i1.31648.

[7] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 39–68. Springer, 2020.

[8] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[9] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

[10] Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Mining and Knowledge Discovery*, 38(5):2903–2941, 2024.

[11] Kristin Blesch, Niklas Koenen, Jan Kapar, Pegah Golchian, Lukas Burk, Markus Loecher, and Marvin N Wright. Conditional feature importance with generative modeling using adversarial random forests. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):15596–15604, 2025. doi: https://doi.org/10.1609/aaai.v39i15.33712.

[12] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transaction of Machine Learning Research*, 2024.

[13] Fiona Katharina Ewald, Ludwig Bothmann, Marvin N Wright, Bernd Bischl, Giuseppe Casalicchio, and Gunnar König. A guide to feature importance methods for scientific inference. In *World Conference on Explainable Artificial Intelligence*, pages 440–464. Springer, 2024.

[14] Christoph Molnar, Timo Freiesleben, Gunnar König, Julia Herbinger, Tim Reisinger, Giuseppe Casalicchio, Marvin N. Wright, and Bernd Bischl. *Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process*, page 456–479. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-44064-9_24.

[15] Quay Au, Julia Herbinger, Clemens Stachl, Bernd Bischl, and Giuseppe Casalicchio. Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery*, 36(4):1401–1450, 2022. doi: 10.1007/s10618-022-00840-5.

[16] Alexander Brenning. Interpreting machine-learning models in transformed feature space with an application to remote-sensing classification. *Machine Learning*, 112(9):3455—-3471, 2023. doi: 10.1007/s10994-023-06327-8.

[17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[18] Susanne Dandl, Marc Becker, Bernd Bischl, Giuseppe Casalicchio, and Ludwig Bothmann. mlr3summary: Concise and interpretable summaries for machine learning models. In *Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI 2024)*. CEUR-WS.org, 2024. URL `https://ceur-ws.org/Vol-3793/paper_36.pdf`.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[20] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2104–2122, 2024. doi: 10.1109/TPAMI.2023.3331846.

[21] Enrique Valero-Leal, Manuel Campos, and Jose M. Juarez. Simple explanations to summarise subgroup discovery outcomes: A case study concerning patient phenotyping. In Irena Koprinska, Paolo Mignone, and Riccardo Guidotti, et al., editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 434–451, Cham, 2023. Springer Nature Switzerland.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions focus on pitfalls for using XAI in AI governance through a practical example. We clearly state the scope towards supervised machine learning and tabular data in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The introduction and discussion state that we are method-oriented researchers, such that our knowledge on the legal aspects is limited, but we think a technical perspective and discussion are required for AI governance in practice.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: Our paper focuses on a non-technical audience, such that we refrain from theoretical or methodological details.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA] .

Justification: We do not have experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA] .

   Justification: We do not implement any code or use data in our paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: We do not have experimental results, i.e., no training or test sets.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: We do not have experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We ensured to respect the aspects highlighted in the NeurIPS Code of Ethics throughout our work, in particular in our example.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed positive impacts of respecting technical nuances in XAI methods for AI governance and regulations, such as the EU AI Act, which impact society. Also, we outlined negative issues that may arise when not respecting technical assumptions.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper poses no such risks because we do not use any of these technologies or datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA] .

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.