My Voice, Your Voice, Our Voice: Attitudes Towards Collective Governance of a Choral AI Dataset

Jennifer Ding
The Alan Turing Institute
jding@turing.ac.uk

Eva Jäger
Serpentine Arts Technologies
evaj@serpentinegalleries.org

Victoria Ivanova
Serpentine Arts Technologies
victoriai@serpentinegalleries.org

Mercedes Bunz
King's College London
mercedes.bunz@kcl.ac.uk

Abstract

Data grows in value when joined and combined; likewise the power of voice grows in ensemble. With 15 UK choirs, we explore opportunities for bottom-up data governance of a jointly created Choral AI Dataset. Guided by a survey of chorister attitudes towards generative AI models trained using their data, we explore opportunities to create empowering governance structures that go beyond opt in and opt out. We test the development of novel mechanisms such as a Trusted Data Intermediary (TDI) to enable governance of the dataset amongst the choirs and AI developers. We hope our findings can contribute to growing efforts to advance collective data governance practices and shape a more creative, empowering future for arts communities in the generative AI ecosystem.

1 Introduction & Related Work

Current concerns about AI and creativity are often grounded in artists' fear of losing control over their work when it becomes training data for AI models. While current technical and legal discourse on this topic concentrates on enabling individual opt in and opt out, there are other dimensions of empowerment worth exploring that may be possible through collective approaches to governance that can enable further distribution of power between contributors to AI training datasets and AI developers. We introduce the "Choral Data Trust Experiment" as a case study, in particular our work surveying the attitudes of artists who contributed to the project in order to guide the design of collective governance infrastructure for the jointly created Choral AI Dataset.

Models of data governance have been explored by generative AI initiatives [5] such as the BLOOM Large Language Model (LLM) [8] and StarCoder LLM [6], but at present, it is not considered for many model building efforts. As a result, data contributors often remain an unacknowledged and disempowered group in the model building pipeline. This problem compounds at the intersection of arts and AI, as the work of artists has been used to train generative AI models that reproduce their work often without their knowledge, consent, or benefit. While examples are emerging for ways to expand transparency about the inclusion of art in training datasets such as Spawning's Have I Been Trained and to enable creator opt out such as BigCode's Am I in The Stack?, governance tools that focus on individual opt in or opt out put the burden to act on the individual and do not create affordances to shape the overall model building process or output.

This project aims to challenge this status quo in AI and build upon prior efforts to capture collective rights and preferences in governance mechanisms such as ethical charters and licences [7, 4] and legal entities such as data cooperatives, data trusts [3, 1] and Trusted Data Intermediaries [9, 2] to empower data contributors to shape the process and outcomes of generative AI projects.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

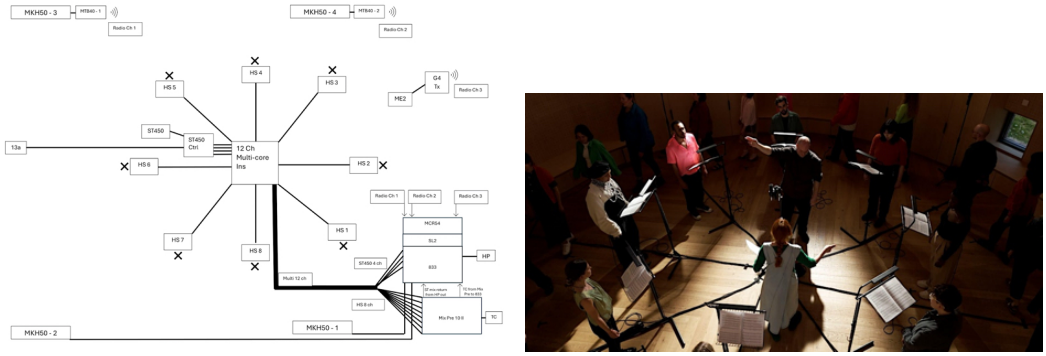


Figure 1: Schematic (left) and image (right) depicting the recording setup for collection of the Choral AI Dataset, with a multi-microphone array capturing 8 close-range microphones for soloists, 4 room microphones and a first-order ambisonic microphone

2 The Choral Data Trust Experiment

15 community choirs from across the UK were invited to record performances of a songbook composed by artists Holly Herndon and Mat Dryhurst. The compositions and recording methods were optimised for the collection of a Choral AI Dataset, purpose-built for training Choral AI models. Herndon and Dryhurst worked alongside researchers from IRCAM, a French music research institute, and engineers at Stability AI to train state-of-the-art models for the exhibition *The Call*, which opened at Serpentine in Fall 2024. This model building process is inspired by previous work by Herndon and Dryhurst on Holly+, a voice AI model trained using recordings of Herndon's own voice. To collect the Choral AI Dataset, the artists traveled to each choir for the recording session, which included use of an ambisonic microphone to capture higher quality data than stereo sound to "future-proof" the dataset (see: Figure 1). The Data Card for the Choral AI Dataset documents further technical information about the data collection, processing, and use considerations.

Alongside the technical challenge of scaling up data collection and model development to accommodate hundreds of different voices, this project also presents the challenge of scaling up data governance with hundreds of choristers with different backgrounds and preferences. Inspired by the Data Trusts Initiative and the investment in human infrastructure such as Data Trustees to facilitate governance for large groups of diverse data subjects, we test the development of a Trusted Data Intermediary (TDI) to assess the opportunity for collective governance of the Choral AI Dataset.

Building Capacity for Collective Data Governance The TDI is composed of a team of Serpentine art curators, legal experts, and an independent data steward, who served as the primary point of contact for the choirs. The data steward began by hosting several Data Conversations on Zoom open to all choristers to share information about the process of training models from the Choral AI Dataset and explore points in the model building pipeline where choristers were interested in more information or agency (see: Figure 6). Afterwards, a Choral Data Preferences survey was released, which received over 100 responses from each of the 15 choirs, as well as a Licence Preferences Polis, which received over 700 anonymous votes. The responses were used to synthesise overall preferences and identify distinct preference groups towards use and governance of the Choral AI Dataset.

Data Preferences Survey Findings A key discovery was the recognition of the discomfort around using the term "data" when referring to choral performance. Choristers described this as "dehumanising", "disembodied", and "disempowering" while also exciting and opening up new possibilities. It also highlights the often disjointed motivations between AI developers and artists (especially live performers), where the former are focused on leveraging digital traces of the art as raw material and the latter are often focused on the ephemeral process of creating or experiencing the art itself.

This concern of misalignment was reflected in survey outcomes, as over 25% of participants responded that they were not comfortable with use of their data by users beyond the project to train an AI model. This stood in contrast to only 4% of participants responding that they were not comfortable with Herndon and Dryhurst creating the Choral AI model for the exhibition. This indicates that

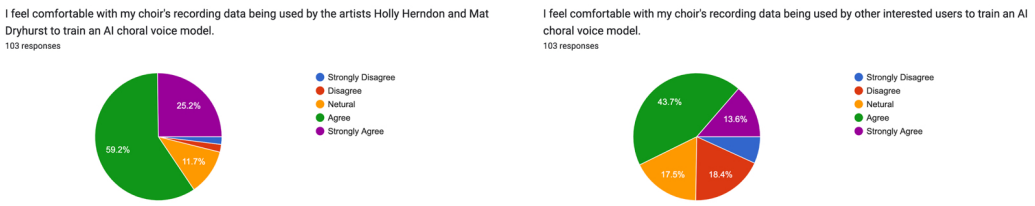


Figure 2: Changes in comfort levels around the use of the Choral AI Dataset to train models by the exhibition artists (left) and other potential users (right)

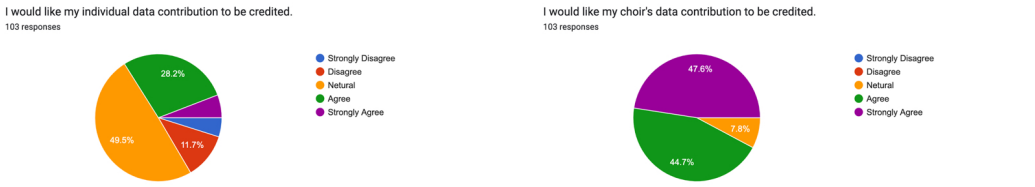


Figure 3: Changes in preferences around crediting for individual contribution (left) and choir contribution (right) by future users of the Choral AI Dataset

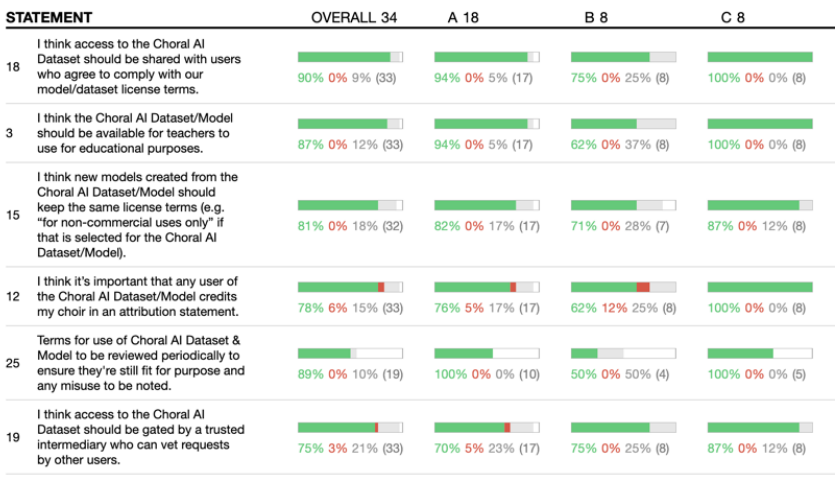
transparency about data practices can mitigate unease and distrust surrounding datafication and model building processes. This also indicates that blanket opt in and opt out do not capture critical nuances of consent preferences, as context of use and user intentions may matter more than the act itself of sharing data for generative AI. The survey also highlights that participants were more interested in group recognition, with over 92% of participants indicating interest in choir level credit, while only 34% indicating interest in individual credit (see: Figure 3). This is further evidence that governance mechanisms should be able to interface with semantic groups rather than solely individuals.

Licence Preferences Polis Findings Across the choirs and individuals, there were differences in risk tolerance and openness to data sharing, important considerations for setting licence terms for the Choral AI Dataset and models. To surface distinct preference groups, a Polis with 20 seed statements was shared with participants to vote on (Agree, Disagree, Pass/Unsure). These statements described different scenarios for potential dataset users and use cases. Over anonymous 700 votes were cast by 37 voters which resulted in 3 opinion groups (A, B, C). While Group C was more permissive in their views towards sharing and wider reuse of the Choral AI Dataset and models, Groups A and B were more cautious, against public sharing and commercial and profit-generating use cases (see: Figure 4).

	STATEMENT	OVERALL 34	A 18	B 8	C 8
8	I think the Choral AI Dataset/Model should be available for public use for any use case.	9% 66% 24% (33)	0% 76% 23% (17)	0% 75% 25% (8)	37% 37% 25% (8)
10	I think the Choral AI Dataset/Model should be available for commercial and profit-generating use cases.	21% 60% 18% (33)	0% 88% 11% (17)	12% 50% 37% (8)	75% 12% 12% (8)
16	I think the Choral AI Dataset should be shared publicly for use by other users.	30% 42% 27% (33)	23% 41% 35% (17)	0% 75% 25% (8)	75% 12% 12% (8)

Figure 4: Polis statements with the highest levels of disagreement among preference groups

However, 90% of voters agreed that the Choral AI Dataset should be shared with users who comply with the licence terms and around 80% are interested in sharing for non-commercial use cases that re-licence under the same terms and credit the choirs for their contribution (see: Figure 5). These findings indicate that if the Choral AI Dataset is released, a licence like the Creative Commons Attribution Non-Commercial Share Alike (CC-BY-NC-SA) would be a good fit to meet group preferences. This outcome and the Polis report data will be used in future negotiations with the AI developers when selecting licences and release strategies for the Choral AI Dataset and models.



	STATEMENT	OVERALL 34	A 18	B 8	C 8
18	I think access to the Choral AI Dataset should be shared with users who agree to comply with our model/dataset license terms.	90% 0% 9% (33)	94% 0% 5% (17)	75% 0% 25% (8)	100% 0% 0% (8)
3	I think the Choral AI Dataset/Model should be available for teachers to use for educational purposes.	87% 0% 12% (33)	94% 0% 5% (17)	62% 0% 37% (8)	100% 0% 0% (8)
15	I think new models created from the Choral AI Dataset/Model should keep the same license terms (e.g. "for non-commercial uses only" if that is selected for the Choral AI Dataset/Model).	81% 0% 18% (32)	82% 0% 17% (17)	71% 0% 28% (7)	87% 0% 12% (8)
12	I think it's important that any user of the Choral AI Dataset/Model credits my choir in an attribution statement.	78% 6% 15% (33)	76% 5% 17% (17)	62% 12% 25% (8)	100% 0% 0% (8)
25	Terms for use of Choral AI Dataset & Model to be reviewed periodically to ensure they're still fit for purpose and any misuse to be noted.	89% 0% 10% (19)	100% 0% 0% (10)	50% 0% 50% (4)	100% 0% 0% (5)
19	I think access to the Choral AI Dataset should be gated by a trusted intermediary who can vet requests by other users.	75% 3% 21% (33)	70% 5% 23% (17)	75% 0% 25% (8)	87% 0% 12% (8)

Figure 5: Polis statements with the highest levels of agreement that informed recommendations for Choral AI Dataset licence terms and further investment in the Trusted Data Intermediary

3 Prototyping Novel Governance Mechanisms

Guided by the findings from the group conversations, survey and Polis, the TDI team worked with legal experts to prototype novel governance mechanisms that aimed to encode contributor preferences into actionable and accountable legal structures. These are described below and in Table 1:

- Formalising Serpentine LLC as the legal entity for the **Trusted Data Intermediary**, which can enter into contracts, act as the administrative hub for further sub-licensing of the dataset, and be responsible for enforcing the terms set out in the Performance Rights Agreement and Data Rights Mandate
- Consolidating choristers' preferences in the terms of the **Performance Rights Agreement**, entered into between the Trusted Data Intermediary and choristers. This Agreement leverages individual performers' rights to set terms for downstream usage of the dataset, which includes permissible uses and types of users, expectations around data security, crediting and compensation practices
- Creating a **Data Rights Mandate** that enables solo singers whose voices (personally identifiable information) are captured in the dataset to mandate the Trusted Data Intermediary the exercise of their GDPR UK data rights

4 Future Work

While the experiment is still underway, our findings raise questions about how enabling foundational components of collective data governance such as providing transparency, building trust, and accounting for diverse preferences can be managed at scale for datasets with many contributors. Alongside the development of automated tools, we propose further investigation into the development and deployment of Trusted Data Intermediaries to navigate these complex challenges.

For the Choral Data Trust Experiment, the TDI has played an important role in capturing, synthesising, and translating preferences across the choirs into practice. By leveraging Serpentine LLC as a trusted legal entity for the TDI, we have the ability to enter into and uphold legal agreements to sustain ongoing gating, maintenance, and governance of the dataset. Whether in the form of an individual representative or team, a TDI can bring flexibility to the process of data governance and a human touch to an otherwise confusing, nonhuman process of transforming art into raw material for producing AI models. We hope to collaborate with more arts and AI communities to advance our shared understanding and best practices for collective and empowering approaches to data governance in the generative AI ecosystem.

Acknowledgments and Disclosure of Funding

The Choral Data Trust Experiment is led by Victoria Ivanova and Jennifer Ding, who acted as the data steward, with Eva Jäger, Ruth Waters and Mercedes Bunz. It is a research and development project associated with Holly Herndon and Mat Dryhust's commission and solo exhibition at Serpentine, *The Call* and incubated by the Future Art Ecosystems project at Serpentine Arts Technologies.

Jennifer Ding was supported by the Ecosystem Leadership Award under the Engineering and Physical Sciences Research Council Grant [EP/X03870X/1], the Turing Institute, and Boundary Object Studio.

Further research support for this work by the Centre for Data Futures at King's College London (Sylvie Delacroix), RadicalxChange (Matt Prewitt) and the Creative AI Lab, a collaboration between Serpentine Arts Technologies and King's College London Digital Humanities Department.

Legal counsel for data empowerment of choir members by AWO Agency (Alex Lawrence-Archer and Ravi Naik) and Keystone Law (Alasdair Taylor).

References

[1] Ada Lovelace Institute (2021, 3). Exploring legal mechanisms for data stewardship.

[2] Centre for Data Ethics and Innovation (2021, 7). Unlocking the value of data: Exploring the role of data intermediaries.

[3] Delacroix, S. and N. D. Lawrence (2019). Bottom-up data trusts: Disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law 9*(4), 236–252.

[4] Hughes, S., H. de Vries, J. Robinson, C. M. Ferrandis, L. B. Allal, L. von Werra, J. Ding, S. Paquet, and Y. Jernite (2023). The bigcode project governance card.

[5] Jernite, Y., H. Nguyen, S. Biderman, A. Rogers, M. Masoud, V. Danchev, S. Tan, A. S. Luccioni, N. Subramani, I. Johnson, G. Dupont, J. Dodge, K. Lo, Z. Talat, D. Radev, A. Gokaslan, S. Nikpoor, P. Henderson, R. Bommasani, and M. Mitchell (2022). Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA, pp. 2206–2222. Association for Computing Machinery.

[6] Li, R., L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim, Q. Liu, E. Zheltonozhskii, T. Y. Zhuo, T. Wang, O. Dehaene, M. Davaadorj, J. Lamy-Poirier, J. Monteiro, O. Shliazhko, N. Gontier, N. Meade, A. Zebaze, M.-H. Yee, L. K. Umapathi, J. Zhu, B. Lipkin, M. Oblokulov, Z. Wang, R. Murthy, J. Stillerman, S. S. Patel, D. Abulkhanov, M. Zocca, M. Dey, Z. Zhang, N. Fahmy, U. Bhattacharyya, W. Yu, S. Singh, S. Luccioni, P. Villegas, M. Kunakov, F. Zhdanov, M. Romero, T. Lee, N. Timor, J. Ding, C. Schlesinger, H. Schoelkopf, J. Ebert, T. Dao, M. Mishra, A. Gu, J. Robinson, C. J. Anderson, B. Dolan-Gavitt, D. Contractor, S. Reddy, D. Fried, D. Bahdanau, Y. Jernite, C. M. Ferrandis, S. Hughes, T. Wolf, A. Guha, L. von Werra, and H. de Vries (2023). Starcoder: may the source be with you!

[7] Pistilli, G., C. Muñoz Ferrandis, Y. Jernite, and M. Mitchell (2023). Stronger together: on the articulation of ethical charters, legal tools, and technical documentation in ml. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, New York, NY, USA, pp. 343–354. Association for Computing Machinery.

[8] Scao, T. L., A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, and et al. (2022). BLOOM: A 176b-parameter open-access multilingual language model. *CoRR abs/2211.05100*.

[9] Stanford Center on Philanthropy and Civil Society (2018, 5). Workshop summary: Trusted data intermediaries.

A Appendix / supplemental material

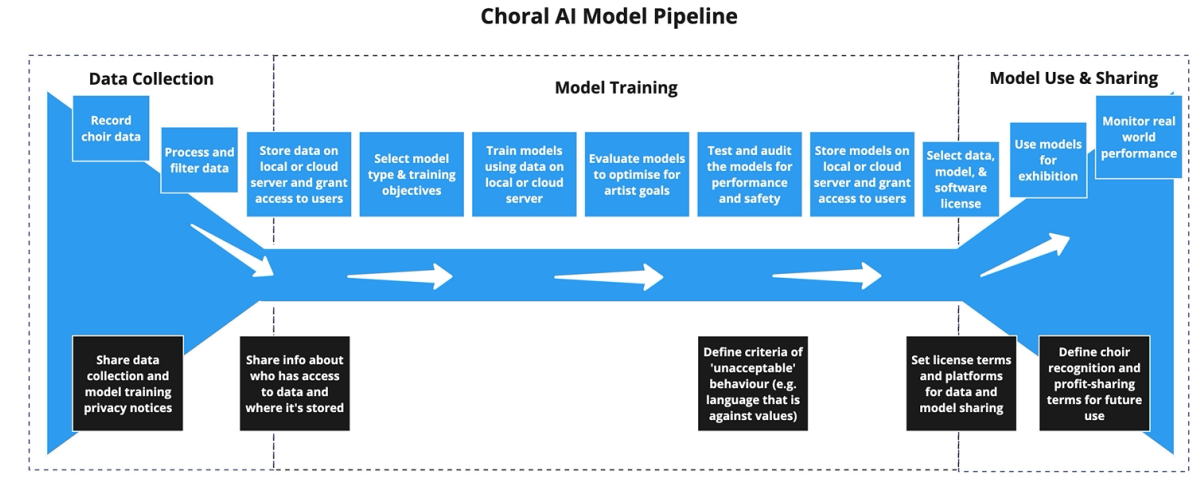


Figure 6: Priority interventions identified by contributors across the Choral AI Model Pipeline

Table 1: Mapping data contributor preferences to governance mechanisms

Governance Mechanism	**Contributor Preferences Source**	**Purpose**
`Trusted Data Intermediary`	Licence Preferences Polis	Formalises Serpentine LLC as TDI so there is a legal entity that can enter into contracts (sub-licensing), act as an admin hub and be responsible for enforcing the terms of the Performance Rights Agreement and Data Rights Mandate.
`Performance Rights Agreement`	Data Preferences Survey	Set terms between TDI and choristers that collectivises the leverages individual performers' rights for downstream usage of the dataset, which includes permissible uses and types of users, expectations around data security, crediting, and compensation practices.
`Data Rights Mandate`	Licence Preferences Polis	Enable solo singers whose individual voices (personally identifiable information) are captured in the dataset to mandate TDI to exercise data rights on behalf of the group.

Choral Data Project - Data Preferences Survey

As part of the Choral Data Project, your choir recording data ("the data") will be used to train an AI model. One output of that model will be an AI choral voice, trained by using the choir recordings dataset. This AI choral voice will be exhibited alongside some of the original choir recordings as well as used to produce new choral music.

The purpose of this online survey is to better understand your preferences for how the choral recording data is used and how each choir and choir member would like decisions to be made about the recordings. The feedback will be used to shape ways of working in the project and shape legal/governance mechanisms that aim to uphold your data rights and preferences.

Figure 7: Screenshot of instructions for the Data Preferences Survey

Choral AI Dataset/Model Preferences

The purpose of this poll is to gather anonymous feedback about preferences about future users & use cases for the Choral AI Dataset and Model (definitions below). This feedback will be used to set licence terms and legal agreements surrounding these resources.

1. Choral AI Dataset: dataset of the recording data prepared for use beyond the project/exhibition
2. Choral AI Model: initial model trained by IRCAM and Holly & Matt from all the recording data produced and prepared for wider release beyond the project/exhibition

Welcome to a new kind of conversation - vote on other people's statements.

Anonymous wrote: *23 remaining*

Folks involved in the project should be kept up-to-date on usage of their data.

Agree | Disagree | Pass / Unsure

*Are your perspectives or experiences missing from the conversation? If so, **add them** in the box below.*

What makes a good statement?

- *Stand alone idea*
- *Raise new perspectives, experiences or issues*
- *Clear & concise (limited to 140 characters)*

Please remember, statements are displayed randomly and you are not replying directly to other participants' statements.

Share your perspective...

Submit

Figure 8: Screenshot of instructions for the Licence Preferences Polis

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: Section 2 describes the set up of the project and initial findings on the claims made 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 paper notes in the **Prototyping Novel Governance Mechanisms** in Section 3 and **Future Work** in Section 4 the limitations of the work performed due to the early stage of the project and open questions regarding impact and sustainability of the approach, which will require more time and further study beyond this single qualitative case study.

   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: The paper does not include theoretical results.

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: The paper does not include experiments. Though the project is called "Choral Data Trust Experiment" that refers to the process of testing the creation of a Data Trust in practice.

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: The paper does not include experiments requiring code.

   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: The paper does not include experiments. Though the project is called "Choral Data Trust Experiment" that refers to the process of testing the creation of a Data Trust in practice.

   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: The paper does not include experiments. Though the project is called "Choral Data Trust Experiment" that refers to the process of testing the creation of a Data Trust in practice.

   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. Though the project is called "Choral Data Trust Experiment" that refers to the process of testing the creation of a Data Trust in practice.

   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: Yes the paper conforms to the code of ethics. Taking UK GDPR into account and to preserve the anonymity of the community who participated in the survey, raw survey data containing PII is not released, but rather what is publicly released is a general report synthesising this data, as described in **Building Capacity for Collective Data Governance** in section 2.

   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: The paper discusses the negative impact on art communities when their work ends up in AI training datasets in **Introduction & Related Work** in Section 1. The paper also discusses potential positive social impacts through empowerment for artists so they can have more say in how their artwork is used in AI projects as described in **Data Preferences Survey Findings** in Section 2, **Licence Preferences Polis Findings** in Section 2, **Prototyping Novel Governance Mechanisms** in Section 3 and **Future Work** in Section 4.

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: This paper does not involve release of data and models so it does not pose such risks.

    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 licence and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not involve the use of 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 licence (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 licence, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licences for some datasets. Their licensing guide can help determine the licence of a dataset.
- For existing datasets that are re-packaged, both the original licence and the licence 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 involve introducing new assets, but rather the process of setting up data governance around potential new assets that will be created in the project. The governance set up process is documented in Section 2, including direct links to the survey and report synthesising the survey findings.

    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, licence, 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: [Yes]

    Justification: There was no crowdsourcing for this project. The participating choirs involved were paid for their time to be part of the art project. The survey and Polis, which were part of this project, with full text of instructions is described and linked in the **The Choral Data Trust Experiment** in Section 2 and screenshots (Figure 7, Figure 8) included in supplemental materials in A.

    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: [Yes]

Justification: Ethical clearance for the research is obtained through Mercedes Bunz's "AI art beyond the Gallery" study obtained on 17 April 24 is valid for five years. The survey and work with the choir members has been classified as 'minimal risk', which was confirmed by the Research Ethics Office, King's College London, reference MRA-23/24-42618.

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.