# DATA COLLECTION IN MUSIC GENERATION TRAINING SETS: A CRITICAL ANALYSIS

**Fabio Morreale**
University of Auckland
f.morreale@auckland.ac.nz

**Megha Sharma**
University of Tokyo
meghas@g.ecc.u-tokyo.ac.jp

**I-Chieh Wei**
University of Auckland
iwei022@aucklanduni.ac.nz

## ABSTRACT

The practices of data collection in training sets for Automatic Music Generation (AMG) tasks are opaque and overlooked. In this paper, we aimed to identify these practices and surface the values they embed. We systematically identified all datasets used to train AMG models presented at the last ten editions of ISMIR. For each dataset, we checked how it was populated and the extent to which musicians wittingly contributed to its creation. Almost half of the datasets (42.6%) were indiscriminately populated by accumulating music data available online without seeking any sort of permission. We discuss the ideologies that underlie this practice and propose a number of suggestions AMG dataset creators might follow. Overall, this paper contributes to the emerging self-critical corpus of work of the ISMIR community, reflecting on the ethical considerations and the social responsibility of our work.

## 1. INTRODUCTION

The quest to generate music with AI (Automatic Music Generation, AMG) is undergoing crucial but overlooked ontological, artistic, and political transformations. Originally confined to academic labs and employed in niche music genres, this quest is gaining traction mostly among commercial companies [1] aiming at automatically generating music in all genres. These transformations are enabled by a combination of socio-technical novelties, including i) the growing influx of money in the field [2, 3]; ii) advanced in Deep-Learning (DL) techniques, such as Transformers [4]; and iii) the increase of cheap computational power. While, from a purely musical perspective, the quality of the music created with AI is undoubtedly rising, from a socio-political perspective, a new gold rush resulting from efforts to outperform competitors and make the best AMG model is following the typical blueprint of capitalist innovation [5–7]: corners are being cut; critical questions have not been asked; short-term gains are prioritised; *permission* is not being sought.

These ethically questionable practices are causing increased concerns. A group of artists [2] recently released a manifesto that identifies one of the most urgent ethical issues arising from AI-generated art: the exploitation of artists' work in training AI generation systems. Similarly, Holly Herndon, a musician famous for popularising AI-generated music, recently criticised OpenAI for not asking living performers' permission to use their music in their AI model, JukeBox [8]. Most systems that generate artistic content using Machine Learning (ML) indeed often indiscriminately populate their training datasets by accumulating original material that is available online [9–12].

Within the ISMIR community, occasional fiery calls requested the community to reflect on the ethical implications [13–15] of, and demanded accountability [2] for the work we produce. However, no specific work investigated the potentially exploitative nature of the datasets we use, and no ethical consideration has been given to how data has been generated. We argue that such investigation is long overdue, especially as the publication of AMG models proceeds undisturbed - and actually, as we will show in the paper, is steadily increasing.

To fill this gap, we aimed to assess the extent to which training sets used in ISMIR papers that propose new AMG models are affected by this issue. We first identified all papers presented at the last ten editions of the conference, from 2013 to 2022, that introduced a new music generation model or a pertinent dataset. Then we identified all dataset(s) that have been used in these papers. Finally, we surveyed information for each dataset, including how data was populated and the extent to which musicians wittingly contributed to its creation.

The contribution of this paper is threefold. First, we provide descriptive statistics about the datasets that are mostly used at ISMIR in AMG applications and how they are populated. Second, we report the ideologies that are embedded in them and outline a lack of adequate engagement with musicians and carelessness on ethical matters. Third, we offer suggestions for dataset creators interested in following responsible practices in their work.

The rest of the paper is structured as follows. We first review literature in Critical Data set Studies and report discussions on ethical issues within MIR. We then describe

[1] A list from Water & Music [1] includes, as of July 2023, companies like Microsoft, Facebook, Google, Spotify, Deezer, and ByteDance.

[2] The European Guild for Artificial Intelligence Regulation, https://www.egair.eu/

the research process, report the results, identify the values that are inscribed in the datasets, and offer suggestions for dataset creators. We conclude the paper with a summary of the study and directions for future work.

## 2. BACKGROUND

Deep-learning (DL), which nowadays is the most commonly adopted method to generate music automatically [16–19], significantly relies on the quality and volume of vast training data. Despite this reliance, dataset development remains an underappreciated element in DL practice.

### 2.1 Critical Data Set Studies

A growing literature on *critical data set studies* [20] aims at identifying the ethical issues and hegemonic power structures of datasets, in particular when used to train ML models [10,12,21]. One of the most urgent issues concerns the exploitation of user labour in AI systems: datasets are populated with data generated by "unwitting labourers" [9] and scraped from the Internet "without context and without consent" [11]. The question around consent is particularly convoluted: consent may have been given unwittingly, for a specific use only, and "some people may never have been given the chance to offer their consent at all" [20]. This concern is not limited to the ivory tower of academia. Musicians are gaining awareness of this issue, and an increasing number of complaints arise from the unfair or unconsented use of original material in AI-generated art. [3]

Most critical work on dataset creation addressed Computer Vision (CV) sets [11, 24, 25] like ImageNet and MS-Celeb, which contain tens of millions of digital images uploaded by platform users. [25] identified the values embedded into these datasets and their formation: evaluating *model work* is prioritised to the detriment of careful *data work*. Another case study that received attention is that of reCAPTCHA [26–28]. Disguised as a *human authentication tool*, reCAPTCHA can be seen as a capture-machine that exploits unpaid individuals' perceptual abilities and micro-labour to train AI datasets [27, 29].

The very way in which most datasets are created embeds specific neoliberal values, like extractivism and deregulation, as exemplified by OpenAI's argument that "IP should be free to use for AI, with training constituting fair use" [30, p. 54]. The all-you-can-scrap ideology dismisses individuals' contributions to dataset creation, which can be met by their creators with a *laissez-faire* attitude [31] that overlooks the ethical implication and liability of scraping the whole Internet [21,25]. In fact, when concerns are voiced, they are specifically aligned with libertarian values and related to how data privacy and data ownership are barriers to collecting data [25].

The practices and routines of data accumulation are not secret. The opposite is true. Among dataset creators, they have become widely accepted, unquestioned, and unchallenged following a process of *dataset naturalisation*: "the contingencies of dataset creation are eroded in a manner that ultimately renders the constitutive elements of their formation invisible" [24]. Notably, the values and ideologies are not only inscribed in how technology is used but also in how it is taught. The lack of interest in how datasets are constructed can indeed be found in the lack of guidance in typical ML textbooks or syllabi [24, 32].

While many dataset creators do not consciously attempt to hide their data accumulation practices, they do not try to fully disclose them either. Dataset naturalisation is indeed exacerbated by ill documentary practices: as reported by [33], ML communities pay little attention to documenting data creation and use. [24] proposes that the lack of information on dataset creation (e.g. how datasets have been created, and whether and how much annotators have been paid) is *structural* - thus ideological - rather than accidental. Every decision and every step in dataset development that is left unaccounted and unarticulated from documentary practices has a political meaning as these steps and decisions are related as "not important" [25]. We will return to this point in the discussions.

### 2.2 Critical turn in MIR and AIM

Several technology communities are undergoing a *critical turn* [34–37] that challenges existing knowledge production methods and political positions as well as ethical and political thoughts within a field. This turn is ethico-onto-epistemological [38, 39] insofar as it questions what kinds of work, knowledge, and social commitment is pursued within and by the community.

While most criticisms of MIR research come from outside the community [3, 40–42], recent academic production within MIR [2, 3, 14, 43–45] and the development of a workshop series on Human-Centric MIR [46] testify that we might be close to a *Critical MIR* - i.e. MIR scholarship devoted to critically analysing the work produced in the field. However, the sort of work that is (not) published at ISMIR (less than 0.5% of the ISMIR submissions engage with any sort of ethical discussions [2]) indicates that the response of the field on ethical issues is still inadequate.

With respect to AMG, the ethical issues that have been identified include copyright issues [15, 47], a narrow and Western-centered understanding to music [43, 45], the risk of *musician redundancy* [2, 14] or the *crisis of proliferation* [44, 48], diversity issues [49], colonialist and extractive practices [2], and assumptions and bias that are embedded in the AI systems [13–15]. To the best of our knowledge, the potentially exploitative nature of AMG datasets remains uncharted territory.

## 3. METHODOLOGY

### 3.1 Researcher Positionality and Motivation

Positionality statements are common in critical studies and serve as a foundation for critical work to understand the research context and the authors' interpretation of the results. Since the outset of the research process, we have

[3] Notable cases include GettyImages suing Stable Diffusion's creators [22] and audiobook narrators complaining against Apple for using their voices to train AI [23].

strived to maintain objectivity and reflexivity by acknowledging our unique positions and backgrounds. All three authors are actively involved in MIR. The first author is formally trained in computer science and is expert in critical theory and technology studies; the second author has a background in computer science; and the third author has a background in electronic engineering and is specialised in machine learning algorithms.

The motivation for undertaking this study is twofold. First, we aimed to support the growth of ISMIR community by contributing to the corpus of self-reflective work, identifying ideologies that might be latent but, once surfaced, can be considered problematic by community members. Second, the development of the suggestions for dataset creation derived from the personal experience of one of the authors, who was involved in dataset creation for AMGs and acknowledged the importance of community guidance on the best ethical practices to adhere to.

### 3.2 Analysis of ISMIR publications

We conducted a systematic review of the last ten editions of ISMIR (2013-2022). A total of 1078 publications were sourced from the conference proceedings. Two of the authors manually filtered the papers adopting two inclusion criteria. First, we included all papers presenting a new music generation model. We included all models that generate new compositions or performances, including in-painting, style transfer, and improvisation. Second, we included all papers that introduced a new dataset that could potentially be utilised as training material for AMG models but rejected works that did not contain symbolic or raw audio music files. For example, we did not include the NSynth Dataset [50], which contains sampled notes from different instruments, but we included MedleyDB [51], which contains annotated multitrack audio.

The analysis proceeded in two phases. First, for each paper, we identified whether authors employed existing datasets (i.e. datasets released or introduced before the publication of the ISMIR paper) or created new ones (i.e. datasets created or introduced as part of the original research reported in the paper). We also examined the presence of any discussions of ethics and permission for using data entries training data for AMG models.

In the second phase, we examined the datasets identified in the first phase. For papers that used an existing dataset, we retrieved dataset information from the original paper (whether or not it was published at ISMIR) in which it was introduced. When we could not find sufficient information in the paper, we checked dataset release links, which were found either in the original paper or by a web search of the dataset name. The information we collected included i) data format (symbolic or audio), ii) how datasets were populated; iii) whether data contained original performances, compositions, or arrangements; iv) the data type; v) the extent to which musicians were involved in the dataset creation and whether they were aware of the intended purposes for the dataset; and vi) whether ethical concerns were discussed.

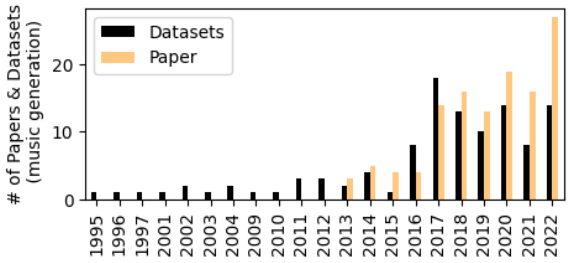



**Figure 1**: Distribution of selected papers and datasets over the years (only papers after 2003 were considered).

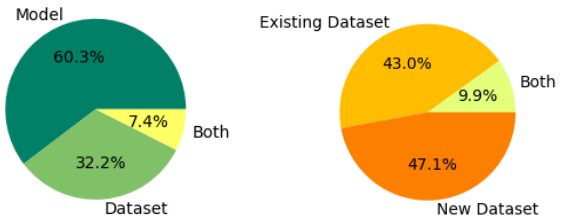
| Dataset Name | Format | Occurrence |
|---|---|---|
| POP909 | Symbolic | 11 |
| Nottingham | Symbolic | 9 |
| Lakh MIDI | Symbolic | 5 |
| HTPD3 | Symbolic | 3 |
| Yamaha e-Competition | Symbolic | 3 |
| Lakh Pianoroll | Symbolic | 3 |
| MusicNet | Both | 3 |
| URMP | Both | 3 |
| AILABS17k | Both | 3 |
| Bach Music21 | Symbolic | 3 |
| Bach Chorales | Symbolic | 3 |
| RWC | Both | 3 |

**Table 1**: The most popular datasets and their occurrences. Each dataset comprises data in symbolic form, but three of them also include audio files.

From a methodological point of view, most of these investigations involved checking the aspect under scrutiny (e.g. whether ethics was discussed) from the dataset sources. The task of identifying how datasets were populated was not as straightforward. In order to streamline the analysis and facilitate the report of the findings, we aimed to cluster datasets into categories that reflected different ways of populating datasets. Two of the authors performed this categorisation following a deductive approach. As they analysed more datasets, they introduced new categories and deleted or merged existing ones. The analysis spreadsheet is available at `https://github.com/Sma1033/amgdatasetethics`.

## 4. FINDINGS

A total of 121 papers survived the filtering. Fig. 1 shows the significant rise of interest in AMG in recent years. Three fourth of the articles (82) introduced a new model, which was either introduced on its own or with a new dataset (Fig. 2a). From this list of papers, we identified 115 datasets (Fig. 2b). When only considering the 82 papers that introduced a new model, most (62 papers, 75.6%) ISMIR researchers use, at least in part, existing datasets to train their AMG models. Tab. 1 shows the 12 most frequently used datasets in our survey, along with data format and their occurrence in our survey. The remaining 104 datasets were only used in one or two papers.




(a) (b)

**Figure 2**: **(a) What is introduced:** new model (73 papers), new dataset (39), both (9). **(b) Dataset Originality:** new datasets (58), existing datasets (53), both (12).

### 4.1 Dataset Creation

We clustered datasets into nine categories to reflect the different ways in which entries were collected (Tab. 2). The categories are non-orthogonal: datasets could be associated with more than one category. In most cases, datasets were populated without creators incurring any costs. Only five datasets used paid online sources or compensated the involved musicians. With the exception of those belonging to the 'Involved musicians' and 'Synthesised music' categories, all datasets were populated with existing music data. New music data accounted for 16.5% of all datasets. We found evidence of poor documentary practices for 18 datasets (15.7%): in 10 cases, there was no information on how data was collected; in 8 cases, we could not find any documentation reporting how datasets were created.

### 4.2 Musicians' involvement

Only 17 datasets (14.8%) involved musicians in any capacity (category 'Involved musicians'). In 11 cases, musicians performed or arranged existing compositions. Three of these datasets (ASF-4, HP-10, and AIST) were entirely created from novel compositions. The remaining three datasets from this category did not contain compositions or performances created specifically for the dataset, or at least, it was not explicitly mentioned. The Irish Traditional Dance Music dataset [61] used the recordings of one of the authors' own performances. For the remaining two datasets [51, 62], the creators mentioned that professional musicians had created those recordings. However, it is unclear whether these recordings are specifically created for the dataset. The other two datasets that included new music belonged to the 'Synthesised music' category and algorithmically generated monophonic melodies [59] or polyphonic MIDI sequences [63].

### 4.3 Musicians' permission and awareness

We checked whether explicit permission was sought from musicians to use their creations to train an AMG model. Only three datasets creators reported having asked such permission. The authors of ASF-4 and HP-10 datasets [64] explicitly mentioned that the musicians involved were made aware of the purpose of the dataset for AMG. Jazz players participating in the creation of the FILOSAX dataset [54] signed a document that provided explanations about the goals of the dataset. However, it is not clear whether these goals included AMG: whereas the authors mention "music generation" in the Abstract, AMG was not included in the list of potential applications of the dataset. In the Mozart Piano Music Dataset [65], pianists gave permission to use their performances for the intended use of the dataset (music analysis), but they were probably not aware and did not consent to have their performances used to train the AMG model introduced in [66].

Two cases were particularly problematic. The MAST Dataset [67], which was introduced for automatic rhythm assessment, was sourced from student entrance exams without seeking consent from the students. Another popular dataset, the Yamaha e-Competition dataset [4] features MIDI files of piano performances obtained from the entries of the piano competition. Although Yamaha claims ownership over all data generated during the event, competitors are unlikely aware their performances are used to train AMG models, as seen in [68]. The lack of permission sought from the musicians clashes with the several comments offered by dataset creators that often acknowledged the valuable contributions made by these musicians, which allows the dataset to existing in the first place.

### 4.4 Discussions on Ethical Issues

Our analysis revealed a lack of engagement with ethical issues, corroborating findings from [25] in their analysis of CV datasets. Only four datasets included any ethical considerations, and only two of them contained an explicit ethics statement. The authors of [69], which presented a new GuitarPro dataset, listed several questions, some of which are particularly relevant to this paper: "*How to acknowledge, reward and remunerate artists whose music has been used to train models?*" and "*What if an artist does not want to be part of a dataset?*". While their spontaneous engagement with these issues is commendable, it is not clear to which extent the authors used these questions in the development of their dataset.

In [70], the authors raised concerns about the impact of AMG for "human musicians of the future". They also stated "care have (*sic*) to be given regarding the fair use of existing musical material for model training" but did not further explain what sort of care and what constitutes unfair use. [57] included an analysis concerning plagiarism issues and observed that their model demonstrated a potential tendency for plagiarism. This issue was also recently highlighted in [47], similar to the level exhibited by a human musician.

## 5. DISCUSSIONS

By leveraging our findings, this section first reports and discusses the values embedded in the datasets used at ISMIR for AMG models. Then, we move to offer practical suggestions to AMG dataset creators.

[4] https://www.piano-e-competition.com/

| Category | Description | Occurrence |
|---|---|---|
| Scraped online | Existing music data collected online from websites [52] or databases [53] | 49 |
| Existing datasets | Existing music data collected from existing datasets [16] | 26 |
| Involved musicians | New music data was created by involving musicians in some capacity [54] | 17 |
| Private data | Existing music data was collected from private databases [55] | 5 |
| Book collection | Existing music data collected from printed books [56] | 5 |
| Online store | Existing music data collected from an online commercial website [57] | 4 |
| CD collection | Existing music data collected from published CD recordings [58] | 3 |
| Synthesised music | New music synthesised using rule-based heuristics or other methods [59] | 2 |
| Not mentioned | No explicit information about how data was obtained [60] | 18 |

**Table 2**: Categorisation of how data was collected in the datasets. For each category, we included exemplary references.

### 5.1 The Values Embedded in Our Datasets

Our analysis extends what [24, 25, 33] have suggested for other ML applications areas: *data work* and data collection practices are de-prioritised and de-valued in AMG datasets used at ISMIR. Most datasets ($\sim$60%) had been populated either by scraping songs from the Internet or by accumulating data from existing ones. Considering original music as a *terra nullius* that is free for the taking means addressing dataset creation with expediency. This approach follows the hegemonic narrative that compares *data* to *oil*. This comparison is highly ideological [71–73] as it disguises the origin (and ends) of data [11], and de-penalises and justifies extractive practices using neo-colonial rhetoric that data is something waiting to be discovered [71, 73].

This narrative underestimates or blatantly neglects the human labour necessary for its development - which includes writing, performing, transcribing, and recording music. The majority of datasets were created by amassing musical compositions initially intended for purposes other than AMG. In these cases, the original labour that was put into the creative acts of composition or performance is simply neglected. Relatively few datasets included original material, and in only two cases, specific permission was asked to use musicians' work to train datasets for AMG purposes. This discussion point resonates with objections to the unfair and exploitative practices of capturing individuals' labour and *humanness* [9] when creating data for digital platforms [6, 74–76] and training AI systems [10–12]. As proposed by [77], human labour is *structurally obfuscated* in ML applications to the benefit of profit and innovation. Similarly, [78] proposes that hiding the labour in this context is crucial to attracting capital investments.

Our direct knowledge and lived experience of MIR offers us a vantage point that we can employ in our reflexive inquiry. We propose that dataset creators might have prioritised safety over criticality and followed common, albeit questionable, procedures simply because these are the procedures that are typically employed in AMG research. This comment is not intended to absolve dataset creators from the responsibilities that come with their work. Rather, it is an invitation to self-assess one's alignment with the exploitative ideologies we surfaced in this section. Yet, we unequivocally found a lack of *data work* - including a limited interest in creating one's own data, exploitation of the labour of unwitting musicians (e.g. in the e-piano competition) and students [67] in dataset curation, and poor documentary practices regarding the source of data [60, 79]. We argue that this lack is ideological. What we leave unaccounted for or unspoken in dataset creation and documentation signs what we consider important or irrelevant [25].

Our findings indicate that the rights and demands of musicians are not prioritised by dataset creators and that the degree to which new models and datasets advance or curb a fair model for musicians is largely ignored. This comment resonates with a note from [80], who explained that streaming services overlook "the rights of musicians or users because their decisions are made based on wholly other problems". It is thus essential that ISMIR researchers and practitioners reflect on the *problems* they drive their decisions on and the *agendas* they implicitly or explicitly follow. Answering questions like "what is the agenda we are following and who benefits from it?" [81] requires community discussions that are difficult, uncomfortable, and controversial but nevertheless necessary. Avoiding engaging with these questions is not a political absence but rather a political tacit acceptance of the status quo [36, 37] as datasets do not exist in a political void [20, 82].

### 5.2 Suggestions

In this section, we offer suggestions to the broader community and to individual authors interested in creating new datasets or using existing ones to train AMG models. We developed these suggestions by integrating results from our analysis with findings from other academic contributions, including ethical CV datasets recommendations [25]. These suggestions are not intended to be meticulously followed as a recipe book. Rather, we devised them as probes, navigation tools, or structured conversations whose development should continue in a participatory way with the rest of the community.

**Develop one's own dataset**. While exploiting musicians' labour in AI dataset creation is a questionable practice [9], expecting dataset creators to seek and obtain consent from all humans involved in AMG datasets is unrealistic [30]. Thus, we recommend creating, whenever possible, one's own dataset and hire musicians for as many tasks as possible (i.e. composing, performing, arranging songs). A small but important amount of datasets in our

investigation followed this practice. We acknowledge that this suggestion might lead to equity issues. If it were to be enforced, only big companies and top university labs would have the economic means to develop such datasets. However, rather than dismissing this issue as unsolvable and continuing business as usual, we propose that the community interrogates itself and finds strategies to tackle it. As an alternative, efforts might be made to develop models that are trainable on small or procedurally-generated datasets following recent successful examples like [30, 83].

**Receive consent from musicians and remunerate them**. Dataset creators should inform musicians about the specific goals of the dataset. It is possible that musicians would willingly consent to train a dataset for several MIR tasks but not for training AMG. When possible, dataset creators should consider paying musicians for their labour and disclose the amount [24], as found in [54]. Given the equity problem discussed above, when paying musicians is not feasible, that should be reported [25], and musicians should be at least acknowledged. When AMG systems are integrated into commercial products, a technical infrastructure might be implemented to distribute royalties to dataset contributors. This suggestion shares Holly Herndon's vision for a novel IP framework "compensates me for my likeness when (and only when) money is made from it" [8].

**Document the process of dataset development**. Our analysis revealed a general lack of care not only in *doing* but also in *documenting* data work. For instance, POP909 dataset's creators did not mention the source or selection process of the "909 popular songs" used to generate piano arrangements [84] and the Lakh dataset's creators simply mentioned that they extracted songs from "publicly-available sources on the internet" [5] website. Careless documentary practices, which we believe were mostly involuntary and caused by an undervaluing of this process in the field [24, 25], implicitly reveal that *how a dataset is developed* and *whose labour goes in it* is not important. We suggest the community develop protocols, guidelines, or templates offering *fair practice* suggestions for dataset creators to follow.

**Report the intended use of the dataset**. Our findings indicate that it is a common practice among AMG dataset creators to reuse existing datasets. We suggest that dataset creators should report the original intended use of their dataset and list the potential 'allowed' applications, following the example of [54]. This practice would prevent, or at least dissuade, future dataset creators from using that data for purposes other than the ones envisioned by the creators and that musicians agreed on. This suggestion is grounded on the observation that technologies are often interpreted, used, and appropriated in ways that their creators cannot foresee or control (what [85] terms *designer's fallacy*). As new applications of datasets are discovered, measures should be taken to ensure that permission from involved musicians is obtained to use their work for uses other than the ones they agreed on.

**When borrowing data, maintain the purpose of the original datasets**. Connected to the above suggestion, creators should maintain their original purpose when borrowing entries for new datasets and avoid misappropriation. This is particularly important when dealing with culturally relevant and sensitive music. This is, for instance, the case of the dataset on the Australian Aboriginal language used by [86]. The author reported: "These datasets were public domain and encouraged for use by the creator as a way to share the sound of the language. Even so, it is not clear that the creators of the dataset from the late nineties could predict this (AI generation) 'future use' case" [30].

**Volunteer ethical considerations**. Our analysis revealed that almost the entirety of the papers did not engage in any form of ethical considerations. Authors can show commitment to advancing more just practices in dataset creation by reflecting on potential ethical limitations in their datasets. Preferably, they should also include documents approved by an Ethics board, if applicable, that were given and signed by the participating musicians.

## 6. CONCLUSIONS AND FUTURE WORK

We identified the dominant approaches to dataset creation within ISMIR and analysed them with critical lenses to understand their ideological substrate. Most authors seem to handle dataset creation with neoliberal attitudes and expediency. However, a small - yet significant - number of dataset creators showed that other attitudes and values are at play within ISMIR when creating datasets for AMG. Our analysis did not explain the motivations for dataset creators to engage, or not engage, with ethical issues in their work, and this investigation is left for future work. Finally, we aim to extend the analysis to papers other than the ones published at ISMIR and to conduct an ethnographic study with AMG dataset creators to give voice to their perspectives on the topic. To conclude, ISMIR has been playing a significant role in the growth of ML models for AMGs but the lack of an ethical infrastructure may facilitate an exploitative industry. It is our responsibility as the main academic hub of AMG to recognise the need to engage in discussions around the matters raised in the article and to establish ISMIR as the home of this debate.

## 7. ETHICAL STATEMENT

In this study, we only used secondary data (desk research) that is publicly available online. We reflect that analysing existing information does not incur ethical issues.

## 8. CONFLICTS OF INTEREST

We have no relevant financial or non-financial interests to disclose and no financial or proprietary interests in anything discussed in this paper.

## 9. DATA AVAILABILITY

The spreadsheet with our analysis is available at `https://github.com/Sma1033/amgdatasetethics`.

[5] `https://colinraffel.com/projects/lmd/`

## 10. REFERENCES


[1] D. Edwards and D. McGlynn, "Creative AI for artists: Track 80+ tools," https://www.waterandmusic.com/data/creative-ai-for-artists/, [Accessed: 12-Apr-2023].

[2] F. Morreale, "Where does the buck stop? Ethical and political issues with AI in music creation." in *Transactions of the International Society for Music Information Retrieval*, 2021, pp. 105–114.

[3] E. Drott, "Copyright, compensation, and commons in the music AI industry," in *Creative Industries Journal*, 2021, pp. 190–207.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017.

[5] D. Harvey, *A brief history of neoliberalism*. Oxford University Press, USA, 2007.

[6] S. Zuboff, "The age of surveillance capitalism: The fight for a human future at the new frontier of power," *PublicAffairs*, 2018.

[7] E. Morozov, *To save everything, click here: The folly of technological solutionism*. Public Affairs, 2013.

[8] M. Clancy, "The Artist: Interview with Holly Herndon," in *Artificial Intelligence and Music Ecosystem*, 2023, pp. 44–51.

[9] F. Morreale, E. Bahmanteymouri, B. Burmester, A. Chen, and M. Thorp, "The unwitting labourer: extracting humanness in ai training," *AI & SOCIETY*, pp. 1–11, 2023.

[10] P. Tubaro, "Learners in the loop: Hidden human skills in machine intelligence," in *Sociologia Del Lavoro*, 2022, pp. 110–129.

[11] K. Crawford, "The atlas of AI: Power, politics, and the planetary costs of Artificial Intelligence," *Yale University Press*, 2021.

[12] N. Dyer-Witheford, A. M. Kjøsen, and J. Steinhoff, "Inhuman power: Artificial Intelligence and the future of capitalism," *Pluto Press*, 2019.

[13] A. Holzapfel, B. Sturm, and M. Coeckelbergh, "Ethical dimensions of music information retrieval technology," in *Transactions of the International Society for Music Information Retrieval*, 2018, pp. 44–55.

[14] G. Born, J. Morris, F. Diaz, and A. Anderson, "Artificial Intelligence, music recommendation, and the curation of culture," *Schwartz Reisman Institute for Technology and Society White Paper*, 2021.

[15] B. L. Sturm, M. Iglesias, O. Ben-Tal, M. Miron, and E. Gómez, "Artificial Intelligence and Music: Open questions of copyright law and engineering praxis," in *Arts*, 2019, p. 115.

[16] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential Generative Adversarial Networks for symbolic music generation and accompaniment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[17] W. Chen, J. Keast, J. Moody, C. Moriarty, F. Villalobos, V. Winter, X. Zhang, X. Lyu, E. Freeman, J. Wang, S. Cai, and K. M. Kinnaird, "Data usage in MIR: History & future recommendations," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 25–30.

[18] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the ACM International Conference on Multimedia*, 2020.

[19] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "MusicLM: Generating music from text," *arXiv:2301.11325*, 2023.

[20] N. B. Thylstrup, "The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains," *Media, Culture & Society*, vol. 44, no. 4, pp. 655–671, 2022.

[21] R. Van Noorden, "The ethical questions that haunt facial-recognition research," *Nature*, vol. 587, no. 7834, pp. 354–359, 2020.

[22] J. Vincent, "Getty images is suing the creators of AI art tool Stable Diffusion for scraping its content," https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit, [Accessed: 11-Apr-2023].

[23] S. Agarwalt, "Audiobook narrators fear apple used their voices to train AI," https://www.wired.com/story/apple-spotify-audiobook-narrators-ai-contract/, [Accessed: 11-Apr-2023].

[24] E. Denton, A. Hanna, R. Amironesei, A. Smart, and H. Nicole, "On the genealogy of machine learning datasets: A critical history of imagenet," in *Big Data & Society*, 2021.

[25] M. K. Scheuerman, A. Hanna, and E. Denton, "Do datasets have politics? Disciplinary values in computer vision dataset development," in *Proceedings of the ACM on Human-Computer Interaction*, 2021, pp. 1–37.

[26] V. Avanesi and J. Teurlings, "I'm not a robot, or am i?: Micro-labor and the immanent subsumption of the social in the human computation of recaptchas," in *International Journal of Communication*, 2022, p. 19.

[27] B. T. Pettis, "reCAPTCHA challenges and the production of the ideal web user," in *Convergence*, 2022.

[28] R. Mühlhoff, "Human-aided Artificial Intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning," in *New media & Society*, 2020, pp. 1868–1884.

[29] J. O'Malley, "Captcha if you can: How you've been training AI for years without realising it," https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it, [Accessed: 12-Apr-2023].

[30] R. Savery and G. Weinberg, "Robotics: Fast and curious: A CNN for ethical deep learning musical generation," in *Artificial Intelligence and Music Ecosystem*, 2022, pp. 52–67.

[31] E. S. Jo and T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2020, pp. 306–316.

[32] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *MIT press*, 2016.

[33] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, p. 86–92, nov 2021. [Online]. Available: https://doi.org/10.1145/3458723

[34] N. E. Gold, R. Masu, C. Chevalier, and F. Morreale, "Share your values! Community-driven embedding of ethics in research," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–7.

[35] S. Benford, C. Greenhalgh, B. Anderson, R. Jacobs, M. Golembewski, M. Jirotka, B. C. Stahl, J. Timmermans, G. Giannachi, M. Adams *et al.*, "The ethical implications of HCI's turn to the cultural," in *ACM Transactions on Computer-Human Interaction*, 2015, pp. 1–37.

[36] F. Morreale, A. Bin, A. McPherson, P. Stapleton, and M. Wanderley, "A NIME of the times: Developing an outward-looking political agenda for this community," in *New Interfaces for Musical Expression*, 2020.

[37] O. Keyes, J. Hoy, and M. Drouhard, "Human-computer insurrection: Notes on an anarchist HCI," in *Proceedings of the CHI conference on human factors in computing systems*, 2019, pp. 1–13.

[38] K. Barad, "Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning," *duke university Press*, 2007.

[39] C. Frauenberger, "Entanglement HCI the next wave?" in *ACM Transactions on Computer-Human Interaction*, 2019, pp. 1–27.

[40] J. W. Morris, "Curation by code: Infomediaries and the data mining of taste," in *European journal of cultural studies*, 2015, pp. 446–463.

[41] N. Seaver, "Computing taste: Algorithms and the makers of music recommendation," *University of Chicago Press*, 2022.

[42] J. Sterne and E. Razlogova, "Machine learning in context, or learning from LANDR: Artificial Intelligence and the platformization of music mastering," in *Social Media + Society*, 2019.

[43] G. Born, "Diversifying mir: Knowledge and real-world challenges, and new interdisciplinary futures," in *Transactions of the International Society for Music Information Retrieval*, 2020.

[44] M. Clancy, "Reflections on the financial and ethical implications of music generated by Artificial Intelligence," Ph.D. dissertation, Trinity College Dublin. School of Creative Arts. Discipline of Music, 2021.

[45] R. Huang, B. L. Sturm, and A. Holzapfel, "Decentering the west: East asian philosophies and the ethics of applying Artificial Intelligence to music." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2021, pp. 301–309.

[46] L. Porcaro, C. Castillo, and E. Gómez Gutiérrez, "Music recommendation diversity: a tentative framework and preliminary results," in *Workshop on Designing Human-Centric Music Information Research Systems.*, 2019.

[47] Z. Yin, F. Reuben, S. Stepney, and T. Collins, "Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation," in *Machine Learning*, 2023, pp. 1–38.

[48] J. Attali, "Noise: The political economy of music," *Manchester University Press*, 1985.

[49] L. Porcaro, C. Castillo, and E. Gómez Gutiérrez, "Diversity by design in music recommender systems," in *Transactions of the International Society for Music Information Retrieval*, 2021.

[50] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proceedings of the International Conference on Machine Learning*, 2017.

[51] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2014, pp. 155–160.

[52] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017.

[53] M. Pesek, P. Godec, M. Poredos, G. Strle, J. Guna, E. Stojmenova, M. Pogacnik, and M. Marolt, "Introducing a dataset of emotional and color responses to music." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2014, pp. 355–360.

[54] D. Foster, S. Dixon *et al.*, "Filosax: A dataset of annotated jazz saxophone recordings," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2021.

[55] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019.

[56] E. Parada-Cabaleiro, A. Batliner, A. Baird, and B. W. Schuller, "The seils dataset: Symbolically encoded scores in modern-early notation for computational musicology." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017, p. 575–581.

[57] S. H. Hakimi, N. Bhonker, and R. El-Yaniv, "Bebopnet: Deep neural models for personalized jazz improvisations." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020, pp. 828–836.

[58] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra, "Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2018, pp. 483–490.

[59] A. Pati, S. Gururani, and A. Lerch, "dmelodies: A music dataset for disentanglement learning," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[60] L. N. Ferreira and J. Whitehead, "Learning to generate music with sentiment," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019.

[61] P. Beauguitte, B. Duggan, and J. D. Kelleher, "A corpus of annotated irish traditional dance music recordings: Design and benchmark evaluations." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 53–59.

[62] L. Crestel, P. Esling, L. Heng, and S. McAdams, "A database linking piano and orchestral midi scores with application to automatic projective orchestration," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017.

[63] A. Ycart, E. Benetos *et al.*, "A study on lstm networks for polyphonic music sequence modelling," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017.

[64] L. Angioloni, V. Borghuis, L. Brusci, and P. Frasconi, "Conlon: A pseudo-song generator based on a new pianoroll, wasserstein autoencoders, and optimal interpolations." in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020, pp. 876–883.

[65] G. Widmer, "Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries," in *Artificial Intelligence*, 2003, pp. 129–148.

[66] S. Lattner, M. Grachten, and G. Widmer, "A predictive model for music based on learned interval representations," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2018.

[67] F. Falcao, B. Bozkurt, X. Serra, N. Andrade, and O. Baysal, "A dataset of rhythmic pattern reproductions and baseline automatic assessment system," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019.

[68] H. H. Tan and D. Herremans, "Music fadernets: Controllable music generation based on high-level features via low-level feature modelling," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[69] P. Sarmento, A. Kumar, C. Carr, Z. Zukowski, M. Barthet, and Y.-H. Yang, "DadaGP: A dataset of tokenized guitarpro songs for sequence models," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2021.

[70] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 178–186.

[71] L. Gitelman, "Raw data is an oxymoron," *MIT press*, 2013.

[72] J. Sadowski, "When data is capital: Datafication, accumulation, and extraction," *Big data & society*, vol. 6, no. 1, p. 2053951718820549, 2019.

[73] L. Stark and A. L. Hoffmann, "Data is the new what? popular metaphors & professional ethics in emerging data culture," *Journal of Cultural Analytics*, 2019.

[74] H. R. Ekbia and B. A. Nardi, "Heteromation, and other stories of computing and capitalism," *MIT Press*, 2017.

[75] B. Brown, "Will work for free: The biopolitics of unwaged digital labour," in *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 2014, pp. 694–712.

[76] M. Pasquinelli, "Google's pagerank algorithm: A diagram of cognitive capitalism and the rentier of the common intellect," in *Deep search: The politics of search beyond Google*, 2009, pp. 152–162.

[77] J. Sadowski, "Planetary potemkin AI: The humans hidden inside mechanical minds," in *Digital Work in the Planetary Market*, p. 229.

[78] L. Irani, "Difference and dependence among digital workers: The case of amazon mechanical turk," in *South Atlantic Quarterly*, 2015, pp. 225–234.

[79] C. Ó Nuanáin, H. Boyer, S. Jordà Puig *et al.*, "An evaluation framework and case study for rhythmic concatenative synthesis," in *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. ISMIR 2016. Proceedings of the 17th International Society for Music Information Retrieval Conference; 2016 Aug 7-11; New York City (NY).[Canada]: ISMIR; 2016. p. 67-72.* International Society for Music Information Retrieval (ISMIR), 2016.

[80] J. W. Morris, "Selling digital music, formatting culture," *University of California Press*, 2015.

[81] C. Cath, "Governing artificial intelligence: ethical, legal and technical opportunities and challenges," in *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2018.

[82] J.-P. Deranty and T. Corbin, "Artificial intelligence and work: a critical review of recent research from the social sciences," *AI & SOCIETY*, pp. 1–17, 2022.

[83] T. Moore and J. Brazeau, "Serge modular archive instrument (smai): Bridging skeuomorphic machine learning enabled interfaces," in *New Interfaces for Musical Expression*, 2023.

[84] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, X. Gu, and G. Xia, "Pop909: A pop-song dataset for music arrangement generation," *arXiv preprint arXiv:2008.07142*, 2020.

[85] D. Ihde, "The Designer Fallacy and Technological Imagination," in *Defining Technological Literacy: Towards an Epistemological Framework*, 2006, pp. 121–131.

[86] R. Savery, R. Rose, and G. Weinberg, "Establishing human-robot trust through music-driven robotic emotion prosody and gesture," in *International Conference on Robot and Human Interactive Communication*, 2019, pp. 1–7.