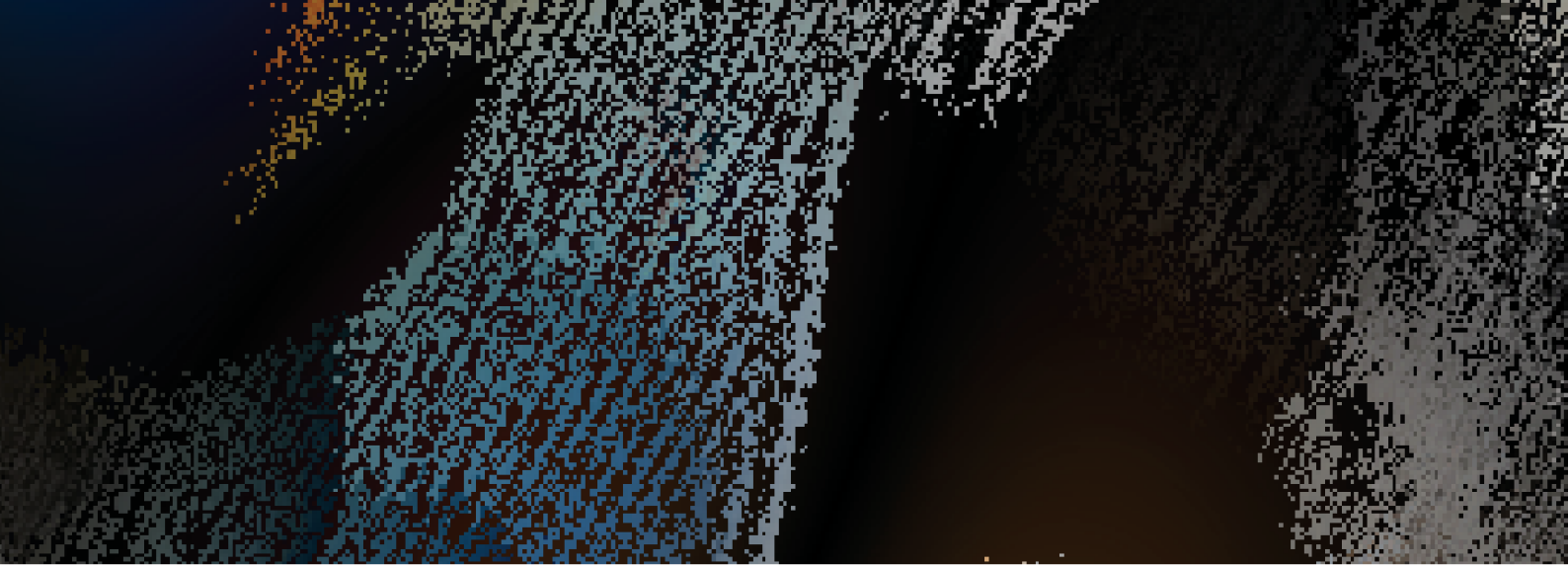


JUNE 2025

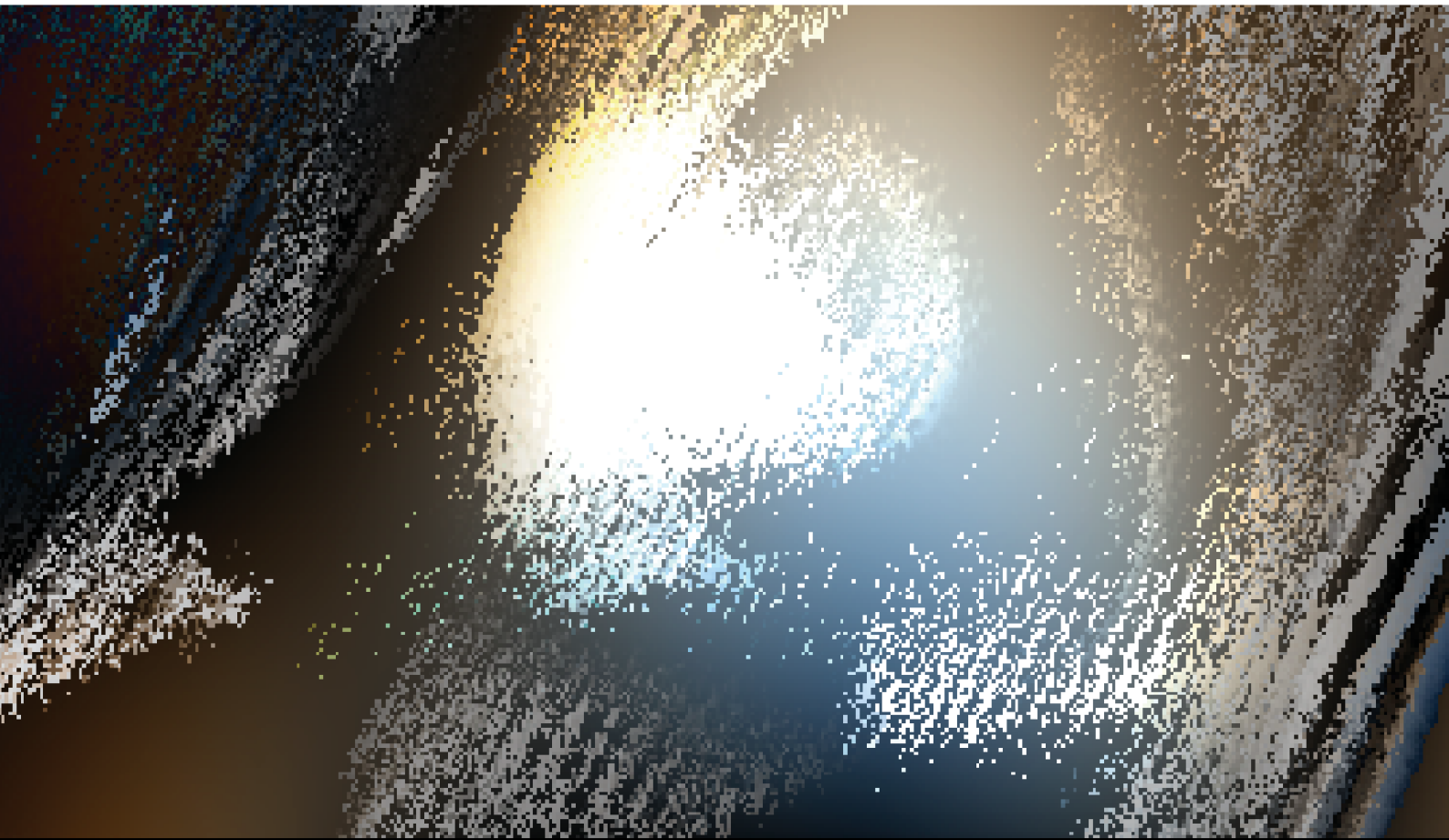
RAI UK INTERNATIONAL PARTNERSHIP X CREATIVE COMPUTING INSTITUTE

BRINGING PEOPLE INTO AI

POLICY REPORT:
ETHICAL APPROACHES TO SMALL DATASETS AND LOW-RESOURCE AI MODELS IN ARTS AND SCIENCES

NICK BRYAN-KINNS,
OLGA SUTSKOVA,
REBECCA FIEBRINK,
PHOENIX PERRY,
ELIZABETH WILSON,
ANNA WSZEBOROWSKA

THE CREATIVE COMPUTING INSTITUTE,
UNIVERSITY OF THE ARTS LONDON



ual: creative computing institute

RAI

Executive Summary

There are substantial concerns about the unethical use of creative content to train large Artificial Intelligence (AI) models. With the increase of under-regulated AI use, the £124 billion UK Creative Industries [19] are at risk of losing revenue to large international tech monopolies, and UK-based creatives are losing their trust in the UK Government's ability to protect and nurture our globally recognised talent [31].

This report builds on a series of stakeholder workshops with Creative Industry professionals and practitioners [6] who identified concerns with the use of AI and ways forward for the legislation in this area. The workshops explored the potential value of low-resource AI models and small datasets for music making, which itself has been heavily influenced by recent AI developments. With expert input, this report identifies key issues of using large AI models, highlights ethical and creative concerns, and proposes approaches to address these concerns. Opportunities to apply these recommendations beyond the music industry, for example, in scientific fields, are also introduced.

The report concludes that a small dataset and low-resource AI approach can bring nuance and character into AI-mediated creative practice while allowing creators more control and recognition for their work. This topic is especially timely as AI researchers and creative practitioners are pushing back [28] against the UK Government's proposed changes to Copyright Law in 2025 to allow for AI model training using copyrighted creative output [13].

The report outlines steps that should be taken for future ethical policy development with creative datasets in mind:

- Improving access to low-resource AI and reducing barriers to its use
- Improving access to small datasets and improving their metadata
- Championing repository access and usability
- Improving dataset security and attribution
- Building financial sustainability for small dataset approaches to AI

The research in this report was conducted as part of the 12-month project, Responsible AI International Community to reduce bias in AI music generation and analysis (2024-2025). Funded by Responsible Artificial Intelligence (RAI) UK International Partnerships (UKRI EPSRC grant reference EP/Y009800/1).

Project lead

Prof. Nick Bryan-Kinns (University of the Arts London, UK; UAL)

Prof. Rebecca Fiebrink (UAL)

Dr. Phoenix Perry (UAL)

Anna Wszeborowska (UAL)

Dr. Olga Sutskova (UAL)

Dr. Elizabeth Wilson (UAL)

Project partners

Prof. Zijin Li (Central Conservatory of Music, China; CCoM)

Dr. Nuno Correia (Tallinn University, Estonia; TU)

Dr. Alex Lerch (Georgia Tech, USA; GT)

Prof. Sid Fels (University of British Columbia, Canada; UBC)

Dr. Gabriel Vigliensoni (Concordia University, Canada; CU)

Dr. Andrei Coronel and Dr. Raphael Alampay (Ateneo de Manila University, Philippines; AdMU)

Prof. Rikard Lindell (Dalarna University, Sweden; DU)

Music Hackspace (UK)

DAACI (UK)

Steinberg (Germany)

Bela (UK)

Title page artwork

Dr. Elizabeth Wilson (UAL)

To share and cite this document

Bryan-Kinns, N., Sutskova, O., Fiebrink, R., Perry, P., Wilson, E., & Wszeborowska, A. (2025). Bringing People Into AI. Project Report. *University of the Arts London*.
DOI: https://doi.org/10.58129/9mpb-h282

Contents

1

Bringing People into AI

The UK Government's consultation on Copyright, Generative AI (GenAI), and the use of Creative Content (December 2024) [13], has led to several in-depth responses from the academic stakeholders in AI developments in Creative Industries. Some responses express concerns around artistic rights [28], whilst others embrace the more lenient, GenAI-favourable frameworks [12]. Currently, the debate is still polarised. Technology companies and governments argue that more lenient regulations around AI training using creative content can only benefit the economy, bringing in revenue from new AI developments [32]. We and others [28] argue that the revenue loss from disincentivising the UK creative community through lenient AI regulation is significantly underappreciated. Moreover, the trade unions (TUC) raise significant concerns over the future of the whole Creative Industry sector if artistic rights are not protected against non-consensual data processing by tech companies [31]. These concerns are consolidated by the recent international AI-regulation protest by the creatives at the Annecy Animation Festival 2025 [2] and the emergency declaration signed by numerous international creative unions to protect creative rights agains unethical AI use[7].

Creative Industry stakeholders are concerned that the training of large AI models using unethical data scraping not only takes attribution rights away from the original creators and intellectual property owners but also destroys the nuance and the essence of the original work. For example, millions of pirated books and papers were used to train Llama 3 [3] and lawsuits have been launched against StabilityAI from artists claiming work was taken and used without consent [18]. Moreover, the structure of large AI models means that they often average out crucial details, diminishing the vigour and individuality of the creative works they are trained on, as well as the historical value and provenance of the original creations. This is especially important for communities whose art embodies and represents cultural heritage and historical nuance.

Music is an excellent example of a vital cultural art form at risk from unethical use of GenAI. Music is the bedrock of every culture and has a significant impact on personal, cultural and societal trends. Music, especially cultural and experimental, is often created, kept and shared by independent artists, with few, if any, creative rights in place as far as AI data processing regulations are concerned [28]. Historically, it has been this type of marginalised or underground music that inspires new artistic styles and genres, and introduces new artists into the multibillion-pound UK music industry and the music industry worldwide. Concerns about intellectual property rights with respect to AI and music are not limited to independent artists, as the music industry as a whole has raised an alarm over the current poor AI regulations and artistic intellectual property [28]. Without relevant regulations, we are at risk of losing creative practitioners' trust in relation to the effective governance and support of artistic cultural integrity and innovation, compromising inspiration and future innovation in one of the most societally impactful industries in the world - an industry which the UK remains one of the few net exporters.

2

Concerns around Large AI Models in Music Making

The professionals in the Music Industry have already raised alarming reports about how every step of music making and sharing has, in essence, been hijacked by AI. Indeed, deep learning AI systems are now used professionally to generate high-quality music outputs merely through text prompts alone [5].

Genre and Style Bias

Deep learning AI models rely on gigantic musical datasets [8], typically trained on mainstream music datasets that are already available, such as Western classical and pop music. Although there is growing interest in creating and using less mainstream and smaller datasets, they are rarely sufficient in scale to train a large deep-learning model. With less than 6% of non-Western mainstream music being used to train AI models [21]. AI creations are inherently biased towards mainstream Western genres.

Creativity Marginalisation

As deep learning-based GenAI becomes integral to music making [37], the trained AI bias will not only lead to more monotonous and uninspired creations but also risk excluding more experimental and non-mainstream creations from emerging. The lack of access to up-to-date production tools for music beyond the mainstream in turn leads to further marginalisation of experimental and minority culture music creators.

Big-Tech Monopolisation

The problem of marginalisation risks affecting not only a specific music style and artists but also the Music Industry itself. Most advanced large AI models and tools are owned by multinational tech companies which have access to the substantial resources required for training models of that size. These organisations scrape the internet to collect massive datasets for AI music training, with limited or no rights attribution to the artist [11]. Without the necessary regulations around large AI models and the use of music as training data, the UK music industry is losing its decision-making powers to “Big Tech”, and in turn losing value, recognition and revenue. Immediate legislative guidelines are essential to protect the UK music industry’s world-leading reputation for innovation and independence. The recent UK government consultation on copyright advocated for policies tackle support the creative rights holders [5], however, these new legislative frameworks might still favour the technology companies over individual creative practitioners. For example, the opt-out scheme advocated for by the UK Government, in which artists can opt out from their data being used in AI training, places the onus of rights management and enforcement on individuals to the benefit of technology companies [28].

Environmental Impact

The impact of large AI models is both computationally and resource-intensive. There are growing concerns around the significant energy consumption associated with training and running models, which rely on large server infrastructures [29]. Despite the scale of this issue, it remains difficult to fully quantify the magnitude of AI processing energy consumption due lack of transparent reporting. Some estimates create a stark picture: in a worst-case scenario, the annual energy consumption of generative AI models used by Google alone could equate to the energy consumption of an entire nation the size of Ireland [9]. A single large-scale AI model training and development session can generate nearly five times more carbon emissions than the lifetime exhaust emissions of an average car [29]. Although researchers working with companies such as Google, Open AI, and Microsoft argue that actual emissions may be lower than the more extreme predictions, they also agree on the importance of the necessary processing and environmental changes to lower the environmental and energy consumption footprint [22]. Tooling such as the ML Co2 Impact calculator [17] gives estimates of the energy use of AI models in kilowatt-hours (kWh), offering greater transparency. The trend is, however, clear—AI's demands are escalating rapidly. Researchers worldwide are advocating for responsible solutions and approaches to AI applications to avoid further negative impacts on the climate and the environment as a result of AI use [23].

A key aspect of addressing these environmental impact challenges lies in differentiating between applications that are essential for societal advancement and those that are driven by commercial novelty or competitive strategy. Implementation of such guidelines poses a philosophical question out of the scope of the current report.

More practical solutions are beginning to emerge. One such approach is the adoption of smaller, more efficient models trained on carefully curated datasets. These models can be optimised to perform specific tasks, with significantly reduced resource consumption. Moreover, not all stages of creative practice require high-fidelity GenAI—early stage ideation often works with sketches or rough cuts of music, which would require less energy to generate.

Evaluating AI systems through multi-criteria lenses, including energy usage, ethical alignment, and societal benefit, could support more balanced decision-making [4]. Instead of focusing solely on performance metrics such as accuracy and speed, a broader and more holistic evaluation would guide developers and policymakers towards choices that treat computing as a finite resource [33] and prioritise environmental responsibility.

3

Small – The Next Big Thing?

Recent developments in responsible and ethical AI solutions direct attention towards the potential of using so-called “low resource” AI [6] in combination with the power of small datasets [24], [14], [6]. Researchers argue that these types of models could be the solution for bringing character and artistic integrity back into AI-mediated creations, questioning the need for large generative AI models in the creative fields [16], [34]. Such low-resource AI models also offer a solution for regions of the world that are energy and compute-scarce or with poor internet infrastructures to support large cloud-based processing [35]. A recent demonstration by the research team behind the *Common Pile v0.1* dataset has shown that, with compatible AI models, small, ethically sourced (open-domain) datasets can be used as efficiently as, and can match the performance of, large AI models used by companies like Google—without relying on unethical practices such as the use of unlicensed or proprietary data without creator consent[15].

With the right guidelines and support in place, a low-resource AI model approach could empower artistic attribution, reduce bias, and give creative agency back to creators. Besides empowering the creative industry and creatives, a low-resource approach is also a more environmentally friendly solution, requiring fewer resources for training and the use of AI models.

As with most innovative solutions, some obstacles need to be overcome to ensure the success of a low-resource approach. To understand creative practitioners’ views on AI-generated creativity, the Responsible AI (RAI) Music project organised an international hybrid workshop in 2024 [6]. The workshop gauged the current climate of AI use in the music industry and whether the stakeholders see small datasets-oriented AI as a realistic and preferable way forward, and if so, what resources and frameworks are needed to make it a reality.

4

Workshop - Experts in AI and Music

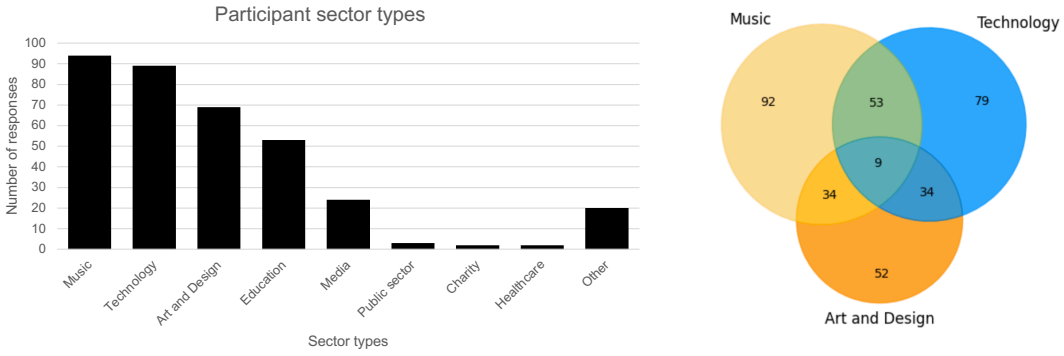
An expert workshop, followed by a stakeholder data analysis session, was an initiative of the RAI International project on Music (MusicRAI), including partners from Canada, China, Estonia, Germany, the Philippines, Sweden, the UK and the USA. The expert workshop was held in a hybrid form (online and in-person) by MusicRAI UK representatives in **London, UK, in July 2024**. The workshop consisted of:

- **Panel Discussions** on the responsible use of GenAI in music (2 total)
- **Case Studies** on GenAI in non-mainstream making (11 total)
- **Focus Group** activities on the topic of responsible GenAI in music (2 total).

The detailed description of in-depth workshop proceedings and conclusions can be found in our academic report [6].

Workshop Attendees

The expert workshop was attended by a total of **148** participants, with the majority being from Music, Technology, Art and Design, and Education backgrounds 4.1.



(a) (b)

Figure 4.1: Attendees' background by sector based on the number of responses.

5

Reflection on Small Datasets and Low-Resource AI

Discussions and focus groups in the expert workshop were analysed in the stakeholder analysis workshop to identify key themes of contemporary discourse around GenAI and music making using thematic analysis. See [6] for a detailed analysis. Themes emerged around the definitions, applications and value of small datasets and the supporting AI models, discussing the benefits of such approaches, see Fig. 5.1. The participants **raised urgency** when discussing the framework to support these approaches, such as **ethical, legal and security measures**, as well as the **accessibility** of the **relevant tools and repositories** to creatives with no technical knowledge of AI and programming.

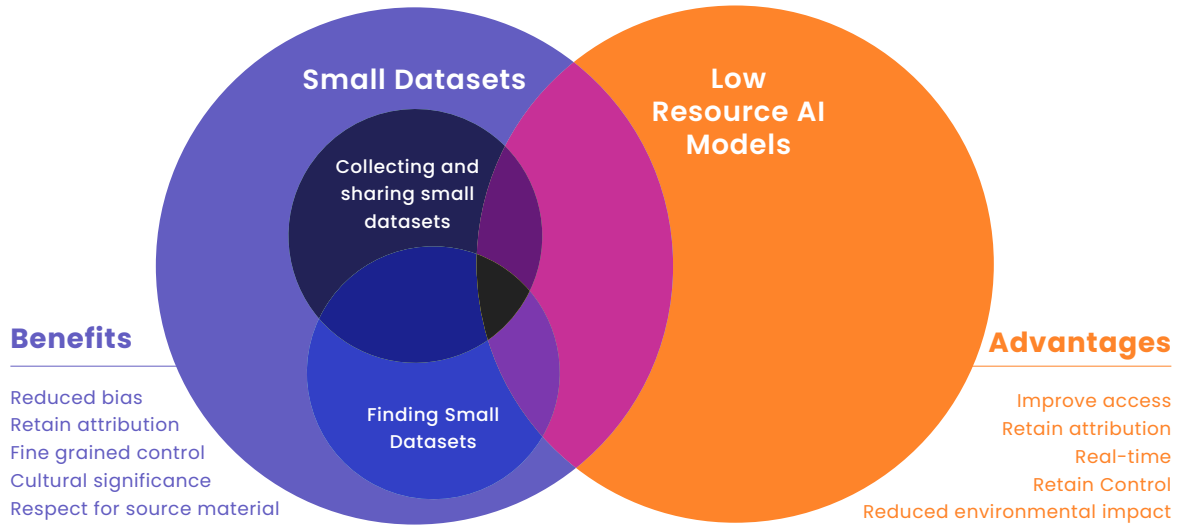


Figure 5.1: The benefits and advantages of small datasets low low-resource models from [6].

Small Datasets: Definition, Application, Value

Experts in the workshop reflected on the current and future role of small datasets of music. The definition, collection and application of small datasets were debated through the prism of data ownership, ethics and supporting systems while preserving music production and sharing values and traditions.

Defining Small Datasets

The conclusions drawn from the workshop were that the small datasets in music should not be defined

by genre alone, but rather classified through the more granular features and specificities of the music, the **metadata** around the creation. These more **granular features** include recording location, time, music mood, or artist name. If music styles are the essence of training intent, then the metadata of music data should encapsulate the cultural context, types of instruments and their articulation, as well as performance technique (live music, exhibition, studio recording, etc). This level of representation could provide sufficient **context** for the specificity of AI training techniques, focusing only on the necessary **selective detail**, bringing context and music character back into the AI-mediated creation.

The pressing issue with clearly defining small datasets and their relationship to GenAI is establishing which music attributes are most salient to capture and how to encapsulate these for multidimensional yet ethical AI use in the future.

Collecting and Sharing Small Datasets

The primary concerns of experts regarding storing and sharing small datasets were around data ethics, data protection and creators' rights, circulating the topics of artistic **consent, agency and cultural appropriation**.

The most straightforward solution to these concerns is that **small datasets remain within the communities** where they are produced and are then used only within these communities. This approach, however, restricts one of the most significant aspects of music creation and innovation, the act of inspiration from other artists and music styles. On the other hand, sharing such datasets outside of the community comes with concerns that users outside of the community might reinterpret such datasets for their agenda and profit, without honouring attribution or reimbursement.

One of the solutions offered by experts is to introduce **structured informed consent** with clear attribution to the music sources in the dataset alongside structured datasheets. The consent should have a **timed and retractability option**, with data usage intent stated and timed. Of course, this means that small dataset holders would need to engage with these frameworks, and dataset quality should be optimised in its granularity, inclusive of attribution information and other necessary metadata for the specificity of use. Unfortunately, many non-mainstream, cultural and independent artists might **not be versed in these new ethical frameworks** or **lack access to the tools needed to engage with such practices**. More education and necessary tools should be accessible for independent artists to ease the process of integration into inevitable AI-futures.

> Government initiatives should be introduced to embrace the implementation of ethical frameworks around data collection and sharing and provide easily accessible educational guidelines and tools to the creative community. This is especially important now, considering the present intellectual property regulations are currently more of a problem than a solution for ethical small dataset sharing by independent creators.

Finding Small Datasets

In the ideal scenario, a repository or navigation tool should aid creative practitioners in finding ethical small datasets, whether music or otherwise. The experts noted that presently, such small datasets are **scattered around the internet**, are **difficult to find** and often with **very limited attributional information**. Information about a dataset is often limited to its filename or a small text file. Unfortunately, at this level of documentation, the engagement with artistic data might not permit the ethical use of creative datasets even with the best intentions in mind. Experts offered a potential solution for such issues, such as implementing specialised ethical **unsupervised machine learning** methods that would apply **non-genre-based classification** techniques to locate and categorise datasets based on attributes which reach beyond coarse genre classification. By doing so, **dataset taxonomies** could move away from the status quo Western lens of music genres towards a more fine-grained, diverse, and inclusive view of music based on culture, environment, artist instrument, representation of specific sounds of interest, etc.

Using Small Datasets

There are ethical benefits of using small datasets, yet at the same time there are several issues when

it comes to their use in practice. The main issues of small and non-mainstream dataset use are around **data quality**. Experts noted that due to a lack of professional equipment or a sufficient environment for good-quality recordings, the recordings found in smaller niche datasets are difficult to process without additional pre-processing procedures. For example, some datasets might suffer from excessive noise or lack of sound, data pollution or intrusion from other environmental noise contaminants.

Another concern is the **technical abilities** that are required from the musicians to engage with AI-based processing of small datasets. Many musicians, especially from independent or cultural music backgrounds, have little or no technical training or the interest to participate in such a level of technical engagement. This puts these creators in a situation in which the new ethical AI systems might not be able to access and promote their creations, and the artists would not be able to benefit from such ethical AI tools. Some level of **AI literacy and ease of access** need to be introduced at all levels of music making.

There are also features of many small datasets that **restrict their use for AI training**. Some small datasets are by design restricted to specific sounds or sound patterns, not full melodies, unlike full compositions that might be used for large AI model training. Additionally, some small datasets are created to be quite experimental or avant-garde, which makes them hard to use for generalised music composition and large AI model training. The inability to use such datasets for any commercial value makes it difficult to commercially justify the storage and maintenance costs of databases with such unique datasets for many artists, raising barriers to access and engagement.

> The specificity of small datasets is a barrier to use by large AI models, and yet it is what makes these datasets invaluable for responsible AI.

Value of Small Datasets

The workshop experts agreed that engaging ethically with small datasets could bring independent and marginalised music communities into the limelight, showcasing their work in the new era of AI. This is an incentive for ensuring that small datasets are open to public access and well-maintained. Ethically managed small datasets could enable communities to create their archives and repositories and keep the **cultural significance and meaning** behind such databases intact. Engagement with the community that owns the small datasets also has the potential to create a **greater understanding and respect** for source material and its creation.

> Small and unique datasets, if managed well, open up a new future of AI music innovation with ethical integrity, character, and creators behind the wheel. This approach could bring under-represented creators into the limelight and inspire new waves of music and AI innovation.

The practical benefits of small datasets outweigh many of the shortcomings in quality as far as the AI-mediated creativity is concerned. Small datasets are often **more detailed and structured**, which could offer **more fine-grained control** to create **more meaningful and authentic creations**. This level of nuance would reflect the essence of the original creations rather than the homogenisation typical of large GenAI models.

Low-Resource AI Models for Small Datasets

The necessity to embrace small datasets in the AI era requires **rethinking the current AI practices** and possibly **revisiting the old**. Workshop experts discussed AI models that are currently able to work with small datasets, and the ways forward for the small dataset mindset when it comes to developing new AI models.

One of the most enticing features of smaller machine learning models, as far as AI impact is concerned, is that they do not necessarily require excessive amounts of computing power. For example, low-resource AI models could be used to create commercially viable products using small datasets for a **fraction of the energy consumption** necessary for a large deep learning AI model, making them more accessible to communities with scarce energy resources [35]. When discussing low-resource AI models suitable for small music datasets, the experts identified several AI architectures and models

(Table 5.1) that have already been successfully applied in music making. Most of the AI architectures are general use (e.g. RNN,'s, Diffusion models, etc), whilst some models are more music-specific (e.g. RAVE, SampleRNN).

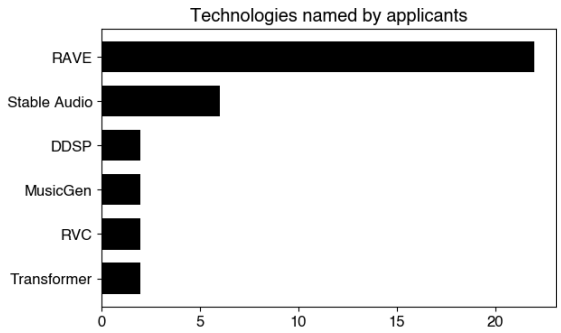
AI Architecture	Music Specific
Recursive Neural Networks (RNN)	No
Long Short-Term Memory (LSTM)	No
Variation Autoencoders	No
Diffusion Models	No
Transformer Models	No
RAVE	Yes
SampleRNN	Yes

Table 5.1: The AI architectures and models for small datasets mentioned by the experts

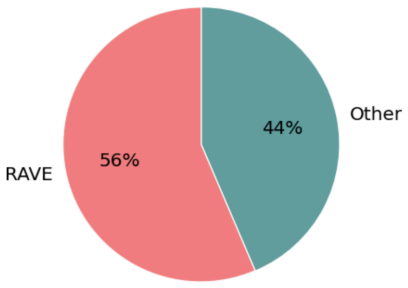
Support Framework for AI models

The technical capabilities of AI models, albeit central to the training outcomes, are not currently the dominant concern for AI-mediated music making. The experts noted that AI model architectures aside, the most defining and often pivotal discussions around AI use are the **support frameworks** around these tools. Such frameworks included **ease of use and access** (workflow integration, data visualisation and representation) and **regulatory frameworks** (policy, data federation, security, licensing of the tools used). Without these in place, it is difficult to ensure a future where everyone benefits from the AI-mediation in an equitable and fair manner.

One approach to facilitate more engagement from the creative community who are not familiar with AI is to **improve access and develop more user-friendly and community-supported tools** for creators of different abilities in programming languages. In music creations, some of such tools are already available and used by many. For example, analysis of applications for the MusicRAI's call for artistic mini-projects, aiming to commission a limited number of music compositions using AI and small datasets [26], revealed that the majority of applicants chose **open-access AI-mediated software that has a strong community support for creative developers** at all levels. Most artists applying for the MusicRAI competition chose the GenAI tool RAVE followed by Stable Audio Open (Fig. 5.2b). Both tools have audio generation frameworks providing support tools, with options to train author models from scratch and use the model within their music composition workflows.



(a) The Frequency of Generative AI models mentioned more than once by the call applicants (participants could name more than one option)



(b) The percentage of applications declaring the use of the RAVE model compared to applications naming other technologies (out of 39 participants)

Figure 5.2: Representation of Most Used GenAI Models by the Music Project Applicants

Value of low-resource AI models

The experts agreed that low-resource AI models bring value from **artistic, ethical and environmental perspectives**. These more restrictive models were argued to be more ethical, offer more control and allow for model specialisation achieved through short training and evaluation cycles. This level of control permits more artistic agency during the training and dataset curation process, enabling more selective and appreciative transformation of the original work.

Finally, due to lower processing needs, the experts noted that low-resource models engaging with small datasets can also be used real-time in **live music performances**. A feat that is currently unimaginable for large language models which take seconds or minutes to generate a piece of audio. This level of live artistic engagement offers the opportunity to bring a whole new form of technology-mediated entertainment into the mainstream, **bringing innovation to the music industry as a whole**.

The main goal now is to ensure that low-resource AI models are readily accessible, understandable, and usable even for the least technically savvy creators.

6

Future Directions and Policy Guidelines

To take concrete steps towards the more widespread use and deployment of small datasets and low-resource AI models, we propose calls to action in this section informing both pragmatic and policy agendas.

Improving access to low-resource AI

To ensure that low-resource AI models are easily findable and accessible, we recommend developing platforms to bring together the fragmented landscape of AI models. To challenge dominant deep-learning technologies, these platforms need to include open and community-editable lists or databases of AI models and which musical features they work best with. This would focus on linking to open-access, non-profit archival repositories and AI models rather than creating another repository of AI models per se.

Reducing barriers to low-resource AI use

Hand-in-hand with the need for more findable low-resource AI models is the need to offer ways to make them more usable. Importantly, we believe that a key opportunity is to create educational tutorials (e.g. videos) for artists about data curation, model training, and ethical approaches to using and customising low-resource AI models. This would also include guides on how to go through the whole GenAI music process from the start with dataset preparation to the finish with AI model selection, training, fine-tuning, and music generation. Such guides empower artists to take more control of GenAI than is possible with closed deep learning models.

We also recommend producing guides for finding the best AI model and data representation for a particular use case, which we found can be challenging for non-technologists. For example, a guide providing an overview of which low-resource AI models are available and how they can be used to generate music, along with guides on how much data would be needed to make the AI models functional for which purposes. And guides for what kind of dataset would be needed for a particular AI model, and what are the pros and cons of each AI model, and what are the privacy and regulatory concerns around their use.

Improving access to small datasets

It is important that communities of users are able to find and access responsibly sourced musical content. To support access, we suggest offering ways to use a subset of large repositories to create curated small datasets, e.g. using Music Information Retrieval (MIR) techniques to find coherent small datasets from large repositories such as SoundCloud. We also suggest developing ways to use AI classification and clustering tools to cluster similar or related datasets together to support dataset curation.

Improving metadata

Current datasets lack consistent metadata such as who composed and performed the music, what genre and style it is, which key it is in, what tempo and rhythm are used, what kinds of instruments are used, how the music is played on those specific instruments, what are the meanings, origins, and cultural context of the music, and so on. In many cases, the metadata is non-existent or of poor quality. We propose that what is needed is a good data dictionary that enumerates and defines each feature found in the datasets linked to by the repository. This metadata then needs to be accompanied by a description of how these musical features relate to features of AI models linked to from the repository.

This could be achieved using a combination of established Music Information Retrieval (MIR) techniques to automatically generate metadata from datasets and offering community-managed taxonomies. Community-managed taxonomies would allow for richer and more culturally meaningful metadata and move beyond colonial epistemologies of genre, where anything beyond the canon of the Global North is labelled as “World Music” regardless of the substantial debt that many genres owe to the Global South.

Championing repository usability

One of the greatest barriers to finding and selecting AI models for use with small datasets is the cumbersome and hard-to-use repository user interfaces (where they exist at all). For example, there is an over-reliance on source control user interfaces such as GitHub (`https://github.com`), which require substantial conceptual technical know-how to successfully navigate. Instead, we recommend developing repository user interfaces which include powerful and usable filters for easier searches, a visualisation-based interface for easy navigation of datasets, and an easy-to-use companion app for capturing and contributing new content.

Improving dataset security and attribution

Key to the success of open, community-based, small datasets and low-resource AI models will be clear community use and access policies. For example, clear methods and governance for adding and removing data post publication that include history deletion, withdrawable consent, and reporting and flagging mechanisms and processes.

Creative attribution needs to be respected and represented across the different levels of data and AI model creation, along with clear attribution governance and explainable terms of use and attribution. Mechanisms also need to be in place to prevent data scraping of datasets whilst also allowing them to be accessible, searchable, workable, and re-mixable. One such example is to apply blockchain based NFT methods to ensure data provenance in a secure manner [25].

Building financial sustainability

The financial and infrastructure costs of deep learning GenAI models and large datasets were repeatedly flagged as barriers to ethical and responsible GenAI with music. To address this issue, we recommend hosting of datasets and AI models on non-profit, open-access archival organisations such as Internet Archive (`https://archive.org`: used for general multimedia, website and literary data sharing and storage), and Zenodo (`https://zenodo.org`: used for scholarly datasets, software and article storage and metadata organisation alongside a persistent identifier for citation, the DOI). The key to supporting the community will be to develop repositories which are either zero or low user cost, whilst ensuring ethical and responsible use of the resources developed and offered by the community.

7

Translational Impact Beyond Music

A well-managed and documented small dataset and low-resource AI approach does not need to be limited to the music industry or even the creative industries. Any kind of **human-centric, people-generated, and legally openly available datasets** used for AI training and ML-based analyses could be approached with a similar responsible and collaborative intent. Potential datasets range from embodied notation and theatrical performance data to multimodal human-cognitive sciences data, such as reaction times, brain and physiological activation recordings.

Human-Centric Science Data

Open data repository practices are now widely endorsed in human-centric sciences, embracing openness and transparency between international researchers from different disciplines. One of the most exemplary instances is the Open Science Framework (OSF: `https://osf.io`), which allows researchers to upload their datasets alongside a description of their research and experimentation intent and descriptions of analyses made to derive their conclusions. The datasets are timestamped, have clear author attribution and affiliation, with author-determined data usage and sharing licensing assigned to the dataset. The **anonymised licensed research data can be easily accessible by other researchers through search options**, without often complex data transfer agreements (DTAs), with an option to reach out to the original author for more in-depth collaboration. Every uploaded dataset is expected to be compliant with local data governance regulations regarding data sharing procedures (such as the GDPR in Europe or UK GDPR in the UK) and stored on domestic OSF servers compliant with such regulations.

The OSF uploaded datasets often have similar features to small datasets of music discussed in this report. They comprise limited intentional information reflecting a certain human phenomenon, such as quantified and categorised behavioural, cognitive or neurophysiological outcomes, to the measures systematically designed by the scientists. The **datasets are often linked to scientific publications that expand on the larger context of acquiring such data**, with publications declaring data sharing with an assigned open data badge and a link to an OSF repository (for example, see [30]).

As with music datasets, if documented well and processed ethically, the merged human-behaviour or psychology-centric data from different OSF sources can be used for meta-analyses and combined machine learning approaches to answer bigger questions, not limited to a single lab or research centre alone. With ethical AI advances and collaborative frameworks in place, these datasets could be an invaluable step to learn more about human nature and biology.

Human-centric science data can be highly sensitive due to the impact of conclusions that could be drawn from the results, and so blindly scraping datasets of human psychology, cognition, and behaviour **for use with large generative AI models would be highly unethical**. Moreover, the conclusions of such an approach will lack the transparency and nuance required when discussing individual

differences and variation in humans. The outcomes risk disproportionately affecting the marginalised communities who often, unfortunately, still have little voice in academic and research circles.

Alongside responsible AI legislation on human-centric science data, more controlled and transparent **low-resource AI models might be the next step towards more ethical AI-led discoveries about people's brains and behaviours**. The ethics and frameworks around large AI models in psychological sciences are already being debated in academic circles [10, 1]. More research and frameworks should be established to support the transition of human-centric psychological science data into the AI era, and communities whose data is being used and processed should be involved in these discussions.

Human Performance Data

Small AI approaches may also be translated into domains such as human artistic performance, including both analysis and synthesis, in practices such as dance, theatre, or other embodied practices. These systems, when trained on small and curated datasets, are often more aligned with values that prioritise re-engagement with human perspectives.

In the context of artistic performance data, small AI models can be trained to recognise, interpret, and co-create movement patterns, whilst retaining significantly lower computational costs [36, 20]. These applications avoid the need for massive datasets or cloud-scale processing and emphasise responsiveness and interpretability to their context.

Smaller AI systems can also be deployed onto local devices or embedded and wearable technologies, either **providing real-time feedback to the performers or enhancing the performance itself** [27, 38]. This could be impactful in educational or community-based contexts, where resources may be limited but the innovation potential is great.

In the context of performance-based practices, the goal of AI usage is often more focused on expression, interpretation and exploration, rather than solving or optimising problem spaces. The result is a new form of translational knowledge, where data from human movement doesn't just inform performance metrics but becomes the basis for shared meaning-making between disciplines. **Ultimately, this approach exemplifies the potential of using AI in a more human way**, where technology can be used to foster connections between the body, data and culture.

8

Conclusions

The UK government consultation, held in December 2024, was a necessary step towards acknowledging the urgency of the legislative changes that need to be made for the UK as the country embraces the AI Era. Some of the options weighed by the Government were realistic and fair solutions for initial discussions. However, it is important to highlight that the final recommendations of the consultation seem dismissive of the years of outcry from the Creative Industry regarding their rights and revenue being "scraped" away by the big tech AI-training procedures.

This report highlights that the UK government should not take the current Creative Industry outcry lightly. By not taking the human-centric approach and setting legislative frameworks favouring more lenient AI-training approaches (e.g. the scrape first, opt-out later solution), the government is risking the talent and revenue of a well-established Creative Industry, in favour of a more innovative yet risky sector of AI.

Establishing Britain's world-leading expertise in AI should not come at the expense of its own citizens' rights, fuelling a feud between the Creative, Technology, and Innovation sectors. Not respecting copyright through non-consensual data scraping will alienate creative talent away from the UK, having immeasurable consequences on the country's global reputation, not to mention diminishing the public's trust in the governing body.

In this report, we proposed several solutions on how to create more ethical, legal, and technologically facilitated frameworks to establish a more symbiotic relationship between the Creative and AI sectors through a human-centric approach. We highlight that although the solutions should be both legislative and technological, they should, most importantly, be accessible to a non-technologically savvy population. These frameworks should be created in collaboration with the creative communities, especially from marginalised and independent creators who are not protected by legal teams and are at the mercy of exploitation of their creativity.

References

[1] Suhaib Abdurahman et al. “Perils and opportunities in using large language models in psychological research”. In: *PNAS nexus* 3.7 (2024), p. 245.

[2] AnimationMagazine. *Protests Against Artificial Intelligence Use in the Industry Planned at Annecy Animation Festival*. Accessed: 17-06-2025. URL: https://www.animationmagazine.net/2025/06/protests-against-artificial-intelligence-use-in-the-industry-planned-for-annecy-animation-fest/.

[3] The Atlantic. *The Unbelievable Scale of AI's Pirated-Books Problem - The Atlantic [Online]*. Accessed: 17-04-2025. URL: https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/.

[4] Lotfi Belkhir and Ahmed Elmeligi. “Assessing ICT global emissions footprint: Trends to 2040 & recommendations”. In: *Journal of cleaner production* 177 (2018), pp. 448–463.

[5] Jean-Pierre Briot and François Pachet. “Deep learning for music generation: challenges and directions”. In: *Neural Computing and Applications* 32.4 (2020), pp. 981–993.

[6] Nick Bryan-Kinns et al. “Leveraging small datasets for ethical and responsible AI music making”. In: *Proceedings of Audio Mostly 2025*. ACM, 2025.

[7] CartoonBrew. *International Coalition Of Worker Unions Declares Emergency Over AI Use In Animation*. Accessed: 17-06-2025. URL: https://www.cartoonbrew.com/artist-rights/international-coalition-of-worker-unions-declares-emergency-over-ai-use-in-animation-247671.html.

[8] Miguel Civit et al. “A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends”. In: *Expert Systems with Applications* 209 (2022), p. 118190.

[9] Alex De Vries. “The growing energy footprint of artificial intelligence”. In: *Joule* 7.10 (2023), pp. 2191–2194.

[10] Dorottya Demszky et al. “Using large language models in psychology”. In: *Nature Reviews Psychology* 2.11 (2023), pp. 688–701.

[11] Pascal Epple et al. “Watermarking Training Data of Music Generation Models”. In: *arXiv preprint arXiv:2412.08549* (2024).

[12] Tony Blair Institute for Global Change. *Rebooting Copyright: How the UK Can Be a Global Leader in the Arts and AI [Online]*. Accessed: 17-04-2025. URL: https://institute.global/insights/tech-and-digitalisation/rebooting-copyright-how-the-uk-can-be-a-global-leader-in-the-arts-and-ai.

[13] GOV.UK. *Copyright and Artificial Intelligence [Online]*. Accessed: 17-04-2025. URL: https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence.

[14] Yun-Ning Hung et al. “Low-resource music genre classification with cross-modal neural model reprogramming”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[15] Nikhil Kandpal et al. *The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text*. 2025. arXiv: 2506.05209 [cs.CL]. URL: https://arxiv.org/abs/2506.05209.

[16] Yana Knight and Mirjam Palosaari Eladhari. “Artificial intelligence in an artistic practice: a journey through surrealism and generative arts”. In: *Media Practice and Education* (2025), pp. 1–18.

[17] Alexandre Lacoste et al. “Quantifying the carbon emissions of machine learning”. In: *arXiv preprint arXiv:1910.09700* (2019).

[18] Matthew Lindberg. “Applying Current Copyright Law to Artificial Intelligence Image Generators in the Context of Anderson v. Stability AI, Ltd.” In: *Cybaris Intell. Prop. L. Rev.* 15 (2024), p. 37.

[19] House of Lords Library. *Creative industries: Growth, jobs and productivity [Online]*. Accessed: 17-04-2025. 2025. URL: https://lordslibrary.parliament.uk/creative-industries-growth-jobs-and-productivity/.

[20] Diego Marin-Bucio. “Dancing Embryo: Enacting Dance Experience Through Human-AI Kinematic Collaboration”. In: *Documenta* 42.1 (2025).

[21] Atharva Mehta et al. “Music for All: Exploring Multicultural Representations in Music Generation Models”. In: *arXiv e-prints* (2025), arXiv–2502.

[22] David Patterson et al. “Carbon emissions and large neural network training”. In: *arXiv preprint arXiv:2104.10350* (2021).

[23] Alison Pease and Arnold Pease. “Computational creativity and the climate crisis”. In: *14th International Conference on Computational Creativity*. Association for Computational Creativity. 2023, pp. 293–297.

[24] Teresa Pelinski et al. “Embedded AI for NIME: Challenges and opportunities”. In: *International Conference on New Interfaces for Musical Expression*. PubPub. 2022.

[25] Mahir Pradana et al. “Revisiting non-fungible token (NFT) research trends: a bibliometric study and future research directions”. In: *Cogent Business & Management* 12.1 (2025), p. 2469764. DOI: 10.1080/23311975.2025.2469764.

[26] MusicRAI Research Project. *Responsible AI international community to reduce bias in AI music generation and analysis [Online]*. Accessed: 17-04-2025. URL: https://music-rai.github.io/.

[27] Diana Serbanescu et al. “Embodied Voice and AI: a techno-social system in miniature”. In: *Artificial Intelligence – Intelligent Art?: Human-Machine Interaction and Creative Practice,* ed. by Eckart Voigts; Robin Markus Auer; Dietmar Elflein; Sebastian Kunas; Jan Röhnert; Christoph Seelinge, pub 2024 (ISBN 9783837669220) (2024).

[28] Anna-Marie Sichani et al. “BRAID researchers response to UK Government copyright and AI consultation”. In: *Zenodo* (2025). DOI: 10.5281/zenodo.14945987.

[29] Emma Strubell, Ananya Ganesh, and Andrew McCallum. “Energy and policy considerations for modern deep learning research”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 09. 2020, pp. 13693–13696.

[30] Olga Sutskova, Atsushi Senju, and Tim J Smith. “Impact of video-mediated online social presence and observance on cognitive performance”. In: *Technology, Mind, and Behavior* 3.2 (2022).

[31] TUC. *“Jewel in the crown” of UK economy at risk without stronger AI protections for creative workers, TUC warn [Online]*. Accessed: 17-04-2025. URL: https://www.tuc.org.uk/news/jewel-crown-uk-economy-risk-without-stronger-ai-protections-creative-workers-tuc-warns.

[32] Tech UK. *AI, creativity, and copyright: finding the right balance [Online]*. Accessed: 17-04-2025. URL: https://www.techuk.org/resource/ai-creativity-and-copyright-finding-the-right-balance.html.

[33] Wim Vanderbauwhede. *Low carbon and sustainable computing [Online]*. Accessed: 17-04-2025. URL: https://www.dcs.gla.ac.uk/~wim/low-carbon-computing/.

[34] Gabriel Vigliensoni, Phoenix Perry, and Rebecca Fiebrink. *A Small-Data Mindset for Generative AI Creative Work*. Online, May 2022. DOI: 10.5281/zenodo.7086327. (Visited on 06/10/2023).

[35] Sai Krishna Revanth Vuruma et al. “From cloud to edge: Rethinking generative ai for low-resource design challenges”. In: *arXiv preprint arXiv:2402.12702* (2024).

[36] Benedikte Wallace et al. “Embodying an interactive AI for dance through movement ideation”. In: *Proceedings of the 15th Conference on Creativity and Cognition*. 2023, pp. 454–464.

[37] Megan Wei et al. “Prevailing Research Areas for Music AI in the Era of Foundation Models”. In: *arXiv preprint arXiv:2409.09378* (2024).

[38] Elizabeth Wilson et al. “MosAIck: Staging Contemporary AI Performance-Connecting Live Coding, E-Textiles and Movement”. In: *Proceedings of the 7th International Conference on Live Coding (ICLC 2023)* (2023).