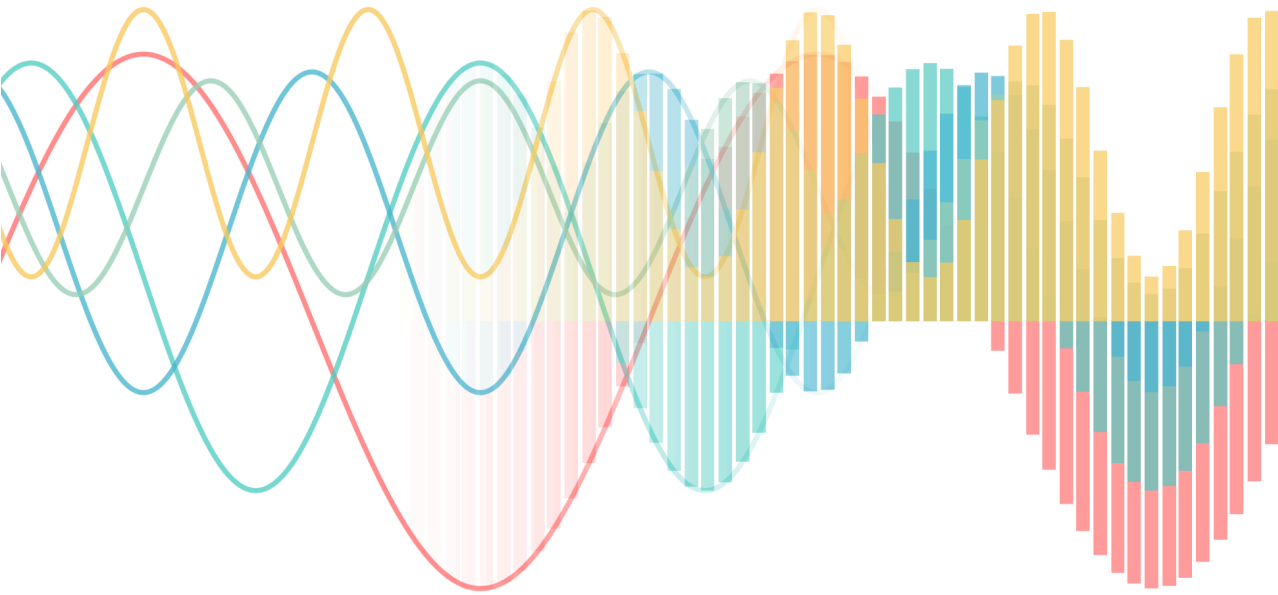
Doctoral Thesis in Speech and Music Communication

Perspectives on AI and Music

Representation, Detection, and Explanation in the Age of AI-Generated Music

LAURA CROS VILA

KTH ROYAL INSTITUTE OF TECHNOLOGY



Perspectives on AI and Music

Representation, Detection, and Explanation in the Age of AI-Generated Music

LAURA CROS VILA

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Monday the 8th of December 2025, at 3:00 p.m. in Q2, Malvinas väg 10, Stockholm.

Doctoral Thesis in Speech and Music Communication / Sound and Music Computing
KTH Royal Institute of Technology
Stockholm, Sweden 2025

TRITA-EECS-AVL-2025:104
ISBN 978-91-8106-460-5

Printed by: Universitetsservice US-AB, Sweden 2025

Abstract

The development and commercial exploitation of AI systems for generating music is impacting how music is created, distributed, and used. This thesis addresses three pressing technical challenges posed by these impacts: (1) how to develop systematic methods for evaluating and representing AI-generated music, including statistical tests and melodic analysis; (2) how to define and build music AI detection systems, revealing vulnerabilities to platform bias and audio manipulation; and (3) how to reframe Explainable AI (XAI) as a communication problem, identifying misalignment risks between AI explanations and human interpretation. The results contribute practical methods for music analysis and detection using AI, and highlight key limitations in current approaches. Overall, this thesis lays a foundation for future work on robust evaluation frameworks and improved explanation methods tailored for music AI systems.

Keywords
Symbolic music representation, AI music detection, AI music, Generative AI, Explainable AI

Sammanfattning

Utvecklingen och det kommersiella utnyttjandet av AI-system för att generera musik påverkar hur musik skapas, distribueras och används. Denna avhandling tar upp tre angelägna tekniska utmaningar som dessa effekter medför: (1) hur man utvecklar systematiska metoder för att utvärdera och representera AI-genererad musik, inklusive statistiska tester och melodisk analys; (2) hur man definierar och bygger AI-detekteringssystem för musik, vilket avslöjar sårbarheter för plattformsbias och ljudmanipulation; och (3) hur man omformulerar Förklarbar AI (XAI) som ett kommunikationsproblem, vilket identifierar risker för feljusteringar mellan AI-förklaringar och mänsklig tolkning. Resultaten bidrar med praktiska metoder för musikanalys och detektering med hjälp av AI, och belyser viktiga begränsningar i nuvarande tillvägagångssätt. Sammantaget lägger denna avhandling en grund för framtida arbete med robusta utvärderingsramverk och förbättrade förklaringsmetoder skräddarsydda för AI-system för musik.

Nyckelord
Symbolisk musikrepresentation, AI-musikdetektering, AI-musik, generativ AI, förklarbar AI

Acknowledgment

I honestly don't know where to start with these acknowledgments, because this thesis wouldn't exist without an incredible group of people who kept me (relatively) sane throughout this journey.

I had never really considered doing a PhD before meeting **Bob L.T. Sturm**. I suppose I always had the image of a PhD student as those I saw in the Telecommunication Engineering (Telecos) school at UPC in Spain, where I did my undergraduate degree. They were some of the smartest people I'd ever known, but they were overworked and underpaid. So, of course, I never quite saw the point in pursuing that lifestyle. However, when I met Bob during the final stages of my Master's at KTH, I started seeing a different side of academia. His passion was infectious. All of a sudden, I found myself overworking because I was doing what I enjoyed most: working at the intersection of Computer Science and Music, something I had no idea was even a research field. I was also underpaid, since as a student I had no income. And I loved it. So, of course, when the MUSAiC project became a reality, I had to at least try to be part of it.

What inspired me at first about Bob was discovering that he shared my passion for signal processing. I still remember how excited I was when I mentioned wavelet transforms and he knew exactly what I was talking about. It was a concept that I did not even fully understand myself back then, but Bob has a remarkable ability to make even the most complex concepts feel accessible and exciting. As a supervisor, he provides thorough, thoughtful feedback and always seems to have the perfect reference for whatever article you are writing. This thesis is actually in its version 14, and he approached each iteration with the same attention to detail. What makes this even more remarkable is that he managed to maintain this level of dedication while supervising three PhD students, four postdocs, and several Master's students simultaneously. I honestly

don't know how he does it, but I am incredibly grateful that he does.

Once the MUSAiC project started, I met my fellow PhD students, **Nicolas Jonason** and **Joris Grouwels**. I already knew Nicolas from our Master's thesis work at Epidemic Sound, but I discovered a whole new side of him over these past four years. He is an incredibly resourceful person, but he is also kind and approachable, which is something I find quite rare among researchers of his caliber. The first time I met Joris was over lunch, where we talked about global economics, a topic that I had found profoundly boring until that very moment. Maybe it is because at heart Joris is a performer, but for the first time in my life, I actually enjoyed listening to someone discuss finance. We were like the three musketeers. None of us really knew what we were doing, especially at the beginning, and our approaches differed quite a lot. Nicolas always tried to stay up-to-date with any and all technologies. He is one of those people who always manages to find the next big development before it becomes mainstream, so once it does become big, he already knows everything about it. Joris would carefully read countless references from any work he found interesting, and thoughtfully try to piece them together. I found myself somewhere in the middle, which in retrospect might have been what helped me stay somewhat sane. Whenever I felt like I was lagging too far behind state-of-the-art progress, I would look at Joris and appreciate the depth and rigour he brought to his work. And whenever I found myself getting lost in endless papers and references, I'd look at Nicolas and realize the value of experimenting and just trying things out to figure out how to narrow the scope. Conveniently, Joris welcomed his wonderful son Toivo a year after starting his PhD, so we were missing a third of our trio for a while. But it wasn't too bad, especially because of how adorable Toivo is. Now, as Nicolas and I are finishing around the same time, Joris has just welcomed his daughter and will take a bit longer to complete his journey. But that just means that we will get to meet up all together again when that time comes.

I had the pleasure to share my thoughts and experiences with so many other people over the years. **André Holzapfel**, my second supervisor, was a perfect counterpart to Bob. While Bob would get excited about many of my research ideas and push me to follow my intuition, André remained sceptical and made sure to challenge them. This helped me develop a more critical view of my own work, and was particularly useful when I had too many research directions I wanted to pursue. **Luca Casini**, my desk neighbor, somehow managed to be genuinely interested in everyone's research, from my DDR obsession to whatever random research rabbit hole anyone else was diving into. He is one of the most generous people I know when it comes to sharing his time, always up for any hackathon idea I had and everyone's go-to person for debugging code. **David Dalmazzo** was a breath of fresh air. He somehow managed to be both a great researcher and musician while also being the most approachable person in any room he entered. He was also the person to ask about how to make your plots pretty, and wasn't afraid to call you out if they looked crappy. And then

there was **Marco Amerotti**, a colleague and friend who provided endless curiosity, great company, and just the right amount of being a pain in the ass to keep things interesting. I inevitably felt like an older sister to him, as he was the youngest of my colleagues, and I am so proud of him for starting his own PhD journey. He might just become the best of us.

The rest of the MUSAiC team — **Elin**, **Ken**, **Rujing**, **Yiren**, **Anna-Kaisa**, **Petra** — are all brilliant, weird, wonderful humans who made this journey infinitely better. I loved hearing about all of their fascinating work in our weekly group meetings. And for that I would like to thank Bob again, for creating a lab that feels more like a family, and all of its members for being such a vital part of it.

My time at TMH introduced me to another wonderful group of people outside of the MUSAiC team, especially **Axel Ekström** whose infectious passion for research was matched only by his enthusiasm for cats, leading to many fascinating conversations that reminded me why I love this work. **Charlotte Stinkeste** had a talent for brightening every room she entered — quite literally, since she loves glitter. Our monthly watercolour painting sessions and ice skating sessions were much-needed distractions from our PhD duties.

There are a few other people whom I admire and who helped me through this journey. One is **Carl Thomé**, who I met when he became my supervisor at Epidemic Sound during my Master's thesis. I think he is in the top five — if not three — of the smartest and most capable people I personally know, as well as dedicated; every time I browse a GitHub repository and look at the Issues, I see his name. How does he have the time and energy for that while simultaneously going to all the after-work events? To this day, I still don't know. Very special thanks also goes to **Emilia Parada-Cabaleiro**, whom I met through the WiMIR mentorship program. She was the most caring mentor I could have asked for. She put up with so many of my insecurities, and we shared many experiences about being women researchers in a foreign country, which really helped me, especially at the beginning of my PhD. I am always so happy whenever I meet her at conferences, and I am lookig forward to seeing her again soon.

Beyond academia, my life was held together by a few key people. **Mar Ouro del Olmo**, la meva germana d'una altra mare, my oldest and dearest friend. Even though her PhD is in a completely different field, we shared the unique experience of being Catalan and doing research in Sweden. She always brings up the most fascinating topics and makes me feel at home. She is also somehow one of the few people who can make me sing and laugh out loud on the floor as if we were still teenagers. I also need to thank **Yvette du Plessis**, my former housemate, friend, and great source of encouragement. She is the kind of person who will always be in your corner, constantly reminding you of your own worth. She was the one who convinced me to ask for more money in my first job after my Master's — and it worked — helping me learn to stand my ground

when I needed to most. **Lorenzo Servedio** is the perfect friend to have if you are doing a PhD. He was always the one to reach out and pull me away from my work, whether it was for a spontaneous trip, a lunch buffet, a workout session, or a movie/series marathon. It is so easy to become completely absorbed in research, but Lorenzo was my constant reminder to actually live life.

Finally, and most importantly, my family: my parents, **Josep** and **Anna**, and my sister, **Núria**, who believe in me unconditionally and have supported every decision that led me here, even when it meant moving to a far-away, cold, dark country. And **Andreas Lundberg**, the most caring, supportive partner I could have asked for. I am excited to see what life will bring us both as we navigate these big changes together.

This PhD research was conducted under the grant ERC-2019-COG No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism.

To everyone who listened to my ramblings about my research, celebrated the small victories, and reminded me that finishing a PhD is, to some extend, refusing to quit long enough — thank you for helping me refuse to quit.

This thesis belongs to all of us.

Sincerely,

Laura Cros Vila
Stockholm, October 31, 2025

List of included papers

Contributions This doctoral dissertation consists of two parts. The first part provides an overview of the research field on which I focus and a summary of my contributions to it. The second part comprises the contributions that I have made to this research field through published works.

Paper A **Statistical evaluation of abc-formatted music at the levels of items and corpora**, Laura Cros Vila and Bob L. T. Sturm In Proceedings of the International Conference on AI and Musical Creativity (AIMC), 2023

Paper B **Cosine Contours: A Case Study with Melodies from Irish Traditional Dance Music**, Laura Cros Vila and Bob L. T. Sturm In Late Breaking Demo in the 24th International Society for Music Information Retrieval Conference (ISMIR), 2023

Paper C **The AI Music Arms Race: On the Detection of AI-Generated Music**, Laura Cros Vila, Bob L. T. Sturm, Luca Casini, and David Dalmazzo In Transactions of the International Society for Music Information Retrieval 8(1), pp. 179–194, 2025

Paper D **(Mis)Communicating with our AI Systems**, Laura Cros Vila and Bob L. T. Sturm In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025

Other contributions by the author not included in the thesis.

Paper E **AI Music Studies: Preparing for the Coming Flood**, Bob L. T. Sturm, Ken Déguernel, Rujing Stacy Huang, Anna-Kaisa Kaila, Petra Jääskeläinen, Elin Kanhov, Laura Cros Vila, David Dalmazzo, Luca Casini, Oliver R Bown, Nick Collins, Eric Drott, Jonathan Sterne, Andre Holzapfel,

and Oded Ben-Tal In Proceedings of the International Conference on AI and Musical Creativity (AIMC), 2024

Paper F **Stochastic Pirate Radio (KSPR): Generative AI applied to simulate commercial radio**, Bob L. T. Sturm, Marco Amerotti, David Dalmazzo, Laura Cros Vila, Luca Casini, and Elin Kanhov In Proceedings of the International Conference on AI and Musical Creativity (AIMC), 2024

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence 3–11, 15, 17–20, 47–49, 51–53, 57–59, 63–65
BPM Beats Per Minute 48, 63
CAEMSI Cross-Domain Analytic Evaluation Methodology for Style Imitation 7, 10, 28, 29
CLAP Contrastive Language-Audio Pretraining 10, 19, 48, 49
CNN Convolutional Neural Network 19
DCT Discrete Cosine Transform 7, 10, 39–41
GDPR General Data Protection Regulation 9, 20
ISMIR International Society for Music Information Retrieval 6, 8
ITM Irish Traditional Music 7, 10, 18, 32, 33, 37, 38, 40
LIME Local Interpretable Model-Agnostic Explanations 6, 22, 59, 61
MIR Music Information Retrieval 17
ML Machine Learning 8, 48
MSD Million Song Dataset 8, 10, 49–51
NCD Normalized Compression Distance 17, 32, 37
PCA Principal Component Analysis 40, 41
SI Style Imitation 17
SpecTTTra Spectro-Temporal Tokens Transformer 19
STFT Short-time Fourier transform 61
t-SNE t-distributed Stochastic Neighbor Embedding xvi, 33, 34, 36
TISMIR Transactions of the International Society for Music Information Retrieval 9, 52
UMAP Uniform Manifold Approximation and Projection xvi, 34, 36, 49
XAI Explainable Artificial Intelligence 6, 15, 20, 22, 23, 57–59, 63–65

Introduction

1 Introduction

This chapter situates the thesis within the rapidly evolving landscape of Artificial Intelligence (AI) applied to music generation, examining both the technical capabilities and cultural implications of current systems. It is structured in three parts: Section 1.1 analyzes why AI and music have converged at this particular moment, exploring the rise of commercial platforms like Suno[1] and Udio[2] alongside the challenges they present; Section 1.2 provides a reflexive account of how this thesis developed through iteration, rejection, and reframing; and Section 1.3 presents the four core research questions that address the evaluation, detection, and explanation of AI-generated music.

1.1 Why AI and Music Now?

The launch of ChatGPT in November 2022 [7] marked a turning point for generative AI. Reporting from UBS Group AG [8] noted that ChatGPT reached an estimated 100 million monthly active users within two months of launch, making it the fastest-growing consumer application in history. Its rapid popularity reflected a technological advance designed for mainstream users, enabling complex creative tasks through simple text prompts and setting a new standard for adoption speed in AI applications. However, this rapid adoption has also revealed significant challenges. Large language models often produce "hallucinations" — false or inaccurate information presented as fact [9] — raising serious concerns about reliability in scientific research and healthcare [10–12]. Additionally, these systems can perpetuate and amplify existing biases, reinforcing racial and social inequities in healthcare and employment decisions [13, 14].

[1] `https://www.suno.com`, last accessed Aug. 19, 2025

[2] `https://www.udio.com`, last accessed Aug. 19, 2025

There now exist numerous companies applying AI to music generation, including Aimi[3], Aiva Technologies[4], Amadeus Code[5], Beatoven.ai[6], Boomy[7], brain.fm[8], Cassette[9], Databass AI[10], Loudly[11], Riffusion[12], Soundful[13], soundraw.io[14], Soundry AI[15], Splash[16], ElevenLabs Music[17], Owl Duet[18], Sonauto[19], Suno and Udio — alongside major corporations exploring similar directions, such as Google, Meta, Microsoft, Open AI, Stability AI and Spotify. These systems are increasingly capable of producing music competitive with that made by humans, as illustrated by an AI-generated track reaching the German singles chart in August 2024 [15], becoming the first AI-generated song to chart in Germany. Music can now also be generated at scales never seen before: Deezer reported in April 2025 [16] that 18% of daily uploads to their platform are fully AI-generated tracks — more than 20,000 AI tracks daily. This number has now, as of September 2025, increased to 28% of daily uploads [17].

The growth of AI-generated music brings important cultural, legal, and ethical challenges [5, 18–21]. While some see these tools as making music creation more accessible, critics warn they could lead to more generic and less diverse music [22, 23]. Many systems are trained on Western classical or mainstream pop music, often ignoring less common genres and non-Western traditions [24–26]. Commercial platforms like Boomy produce millions of tracks, which could overwhelm streaming services and make it even harder for human artists to be heard [5, 27]. Yet most research still focuses on technical improvements rather than these broader effects [28].

There are also important legal questions around ownership and copyright. Some platforms let users own their generated songs, while others keep the rights [18, 29]. The development of commercial AI music systems raises fundamental concerns about exploitation of existing musical works and creators' rights, as high-

[3] `https://aimi.fm` , last accessed Aug. 21, 2025

[4] `https://www.aiva.ai` , last accessed Aug. 21, 2025

[5] `https://amadeuscode.com/en/` , last accessed Aug. 21, 2025

[6] `https://www.beatoven.ai` , last accessed Aug. 21, 2025

[7] `https://boomy.com` , last accessed Aug. 21, 2025

[8] `https://www.brain.fm` , last accessed Aug. 21, 2025

[9] `https://cassetteai.com` , last accessed Aug. 21, 2025

[10] `https://databass.ai` , last accessed Aug. 21, 2025

[11] `https://www.loudly.com` , last accessed Aug. 21, 2025

[12] `https://classic.riffusion.com/` , last accessed Aug. 21, 2025

[13] `https://soundful.com` , last accessed Aug. 21, 2025

[14] `https://soundraw.io` , last accessed Aug. 21, 2025

[15] Recently renamed as Sample Planet. `https://www.sampleplanet.fm/` , last accessed Aug. 21, 2025

[16] `https://www.splashmusic.com` , last accessed Aug. 21, 2025

[17] `https://elevenlabs.io/music` , last accessed Aug. 21, 2025

[18] `https://www.owlduet.com` , last accessed Aug. 21, 2025

[19] `https://sonauto.ai` , last accessed Aug. 21, 2025

lighted in ethical analyses of the field [30]. Models like MusicLM [31] and MusicGen [32] have faced criticism for producing outputs that are too close to their training data [31, 33], raising copyright concerns. Terms of Service of platforms like Suno and Udio often ban automated scraping of their music, but court rulings such as X Corp. v. Bright Data [34] suggest that collecting public data may still be legal in some cases. In Europe, academic researchers are allowed to use publicly available data under EU Directive 2019/790 [35], and similar protections exist under U.S. fair use for non-commercial research [36].

Training datasets of commercial music are another grey area. Sometimes data is collected without clear consent or documentation, which has led to growing criticism. Initiatives such as Fairly Trained[20] and the AI OK[21] campaigns seek to promote the use of datasets with proper permissions and clearer provenance, yet significant legal uncertainties remain. Calls for stronger measures have also come from outside academia and industry, such as the March 2023 open letter [37] urging a pause on training AI systems more powerful than GPT-4. Controversies around large-scale datasets illustrate the problem: the Books3 collection [38], which contained over 190,000 books scraped from the web, was taken down in 2023 after widespread complaints from authors [39]. Meanwhile, companies like OpenAI, Microsoft, and Meta continue to face lawsuits [40] over the ways they collect and use copyrighted material in training.

The advertising of commercial AI music systems often promote a narrative of "democratization." As Mikey Shulman, Suno's CEO, put it, the company's goal is to "democratize music creation and unlock the musical creativity within everyone" [41]. These systems position themselves as tools that remove barriers to creative expression by eliminating the need for training or expertise [42]. However, this framing tends to obscure what some scholars describe as technologically mediated deskilling [43] — a restructuring of creative labour where users shift from active creators to passive selectors or curators. Such transformations also resonate with broader social concerns: recent research shows that anxieties about AI are closely linked to fears of technological unemployment, particularly among younger populations such as university students [44]. Shulman exposes a misalignment between the values embedded in AI tools and those held by artists and creators when he further claims that making music "takes a lot of time, it takes a lot of practice, you have to get really good at an instrument or really good at a piece of production software. I think the majority of people don't enjoy the majority of time they spend making music" [45]. At the same time, recent research [22] shows that generative systems like ChatGPT can lead to a homogenization of ideas: users produce a higher quantity of ideas, but these ideas are less semantically diverse and more influenced by the system's stylistic defaults. In music, where aesthetic variation is central, such tendencies raise concerns about the long-term impact of generative AI on

[20] `https://www.fairlytrained.org/`, last accessed Aug. 19, 2025

[21] `https://www.ai-ok.org/` , last accessed Aug. 19, 2025

creative diversity.

Meanwhile, the commercial acceptance of AI music has rapidly accelerated. Services like Spotify have relaxed restrictions on AI-generated content [46], while music rights organizations raise alarms about its impact. A 2024 survey by music rights groups GEMA and SACEM [47] found that 89% of creators want AI-generated music clearly labelled, and 71% fear it will push human-made art to the sidelines. This concern is not only economic but also cultural: in a future where AI can be used to generate music at massive scales, distinguishing between human and AI-made content becomes increasingly important [3, 16, 48]. This raises practical and epistemic questions about how we evaluate, detect, and explain AI-generated music — questions that are central to this thesis.

1.2 Reflections on the Development of This Thesis

This thesis addresses three interconnected challenges. First, it develops systematic approaches for evaluating AI-generated music, including statistical analysis and melodic pattern modelling [1, 2]. Second, it defines the AI music detection problem and proposes methods for identifying AI-generated music, revealing key vulnerabilities such as platform bias and audio manipulation [3]. Third, it reframes explainability in AI as a communication problem, emphasizing misalignments between AI-generated explanations and human interpretation [4]. This research contributes technical methods and conceptual insights at the intersection of music informatics, human-computer interaction, and critical perspectives on AI. It explores how AI-generated music can be represented and evaluated, how AI-generated content can be detected, and how explanations from music AI systems can be improved through a communication lens.

This thesis evolved not only through research and experimentation, but also through iteration, rejection, and re-framing. A particularly transformative example was the trajectory of the paper that eventually became [4]. It began as a technical paper intended for ISMIR 2023, analysing the behaviour of the XAI method LIME [49] when applied to a pitch recognition model. Our preliminary experiments revealed that small changes in LIME's configuration could dramatically alter the resulting explanation — effectively letting us "tune" the explanations produced by LIME. Although the paper was not completed in time for submission, this line of inquiry prompted me to reconsider the focus of the work: why are we taking these explanations at face value? Why aren't we questioning the assumptions built into XAI tools themselves?

This led me to consider communication theory [50–54] and to the idea that explainability in XAI systems is often treated as a one-way transmission, when in fact it should be analysed as a complex human-agent interaction, such as done by Miller et al. [55, 56]. I reframed the paper as a survey and theoretical analysis, emphasizing interpretability as a form of communication, and submitted

it to ACM FAccT 2024. It was rejected, though the reviews pointed out both its promise and its need for greater focus.

I rewrote the paper as a more concise, targeted argument and submitted it to CHI 2024, where it was accepted [3]. In the process of writing and revising this work, I shifted from treating explainability as a technical endpoint to treating it as an ongoing, situated exchange between system and user — one that is prone to miscommunication and confirmation bias.

This trajectory reflects how my scholarly identity has evolved throughout the PhD: from being primarily concerned with technical systems and evaluation metrics to engaging more with interpretability, epistemology, and critical perspectives on AI. It also reflects a broader shift in this thesis — from addressing isolated technical problems to interrogating the underlying assumptions and values embedded in music AI systems.

The origin of paper [1] was to find a way to compare music corpora as well as items in corpora, with the focus on symbolic music. This is somewhat addressed by Cross-Domain Analytic Evaluation Methodology for Style Imitation (CAEMSI) [57], a domain-independent framework for comparing corpora via "typicality"-based statistical tests for difference and equivalence. However, while this method enables high-level comparison of music corpora, it does not allow tracing significance back to individual pieces, as its use of permutation testing to estimate statistical significance prevents assessing items separately. Our paper [1] provides an alternative: by estimating the distributions of distances between pairs of melodies, we represent both entire music corpora and individual pieces relative to a reference set. In the context of Irish Traditional Music (ITM), we then evaluate which pairwise distance metric and distribution difference metric yield statistically significant results. Our approach provides more nuanced and interpretable results than CAEMSI, allowing us to find interesting items in corpora, such as outliers or typical elements.

However, none of the distance measures we tested are necessarily musically informed. This is what let us to explore alternative representations for symbolic music, and eventually to cosine contours [58]. Our paper [2] explores the use of cosine contours (or Discrete Cosine Transform (DCT)) to analyse ITM. We find some shortcomings of the method, notably that it fails to be "efficient" as the length and complexity of the analysed melodies increase. Yet, the premise of the method still sounded promising. Using basis functions to represent melodies in symbolic music could not only make for good compression of information to use in downstream tasks, but could also reveal patterns in corpora. This is also suggested in [58], as the first DCT component describes the "ascendingness/descendingness" of a song.

Finally, our journal article [3] began with our observation of the rise in popu-

larity of AI music generation systems Suno and Udio,[22] and realising that there is a need for detecting AI-generated music [5, 20, 31, 47, 48]. The first system we analysed was Boomy, which is featured in [5]. Subsequently, with the rapid rise and widespread adoption of text-prompt-based systems, we expanded the research by collecting a large dataset of song metadata and audio from these platforms. From May 2024 to September 2025, the dataset has grown to include over 460,000 audio recordings and more than 1.8 TB of data (including mp3 and JSON files). Our ongoing data collection effort — aiming for one million AI songs to match the Million Song Dataset (MSD) [60] — supports computational analysis of AI-generated music and aims to document the current state of AI music generation as it unfolds.

Using the collected data from Boomy, Suno and Udio, we first wrote the paper "AI Music: A New Frontier for Music Information Retrieval" and submitted it to ISMIR 2024. Our submission presented two case studies. The first explored whether K-means clustering could recover Boomy's proprietary style taxonomy using MusiCNN embeddings, achieving 64% accuracy and comparing this to human listening test results. The second case study trained Support Vector Machine and Random Forest classifiers to distinguish between AI-generated music from Boomy and Suno, and human-made music, achieving F1-scores above 0.90. However, the reviewers rejected the submission because they felt that even if it was a "relevant topic that needs to be present at ISMIR, that would clearly trigger interesting discussions", "its scientific quality is not sufficient for ISMIR". They noted limitations such as reliance on relatively basic Machine Learning (ML) (no deep learning), potential dataset biases, unclear motivation for some analyses, and lack of rigorous evaluation, particularly in the listening tests.

This rejection led us to expand the work into a comprehensive journal article that examined three case studies: composer identification (distinguishing AI-generated from human music), structural analysis (comparing musical structures of AI and human music creations), and prompt analysis (investigating the textual prompts used to generate AI music). We then decided to focus on Suno and Udio, as those systems were more similar in the way they were used, and by then had gained a lot more popularity and had more engagement from users (something that Boomy did not really have anymore). However, as we developed this expanded version, we realized that each case study deserved deeper treatment than a single journal article could provide. The scope had grown too broad again, and we faced the same problem that had affected our ISMIR submission — trying to cover too much ground without sufficient depth.

We thus made the decision to split our work into three separate journal articles, each focusing on one aspect with the thoroughness it warranted. The detection aspect became the foundation for [3]. The two other articles arising

[22]The lawsuit filed by the RIAA et al. against Suno on June 24, 2024, states that the platform has "over 10,000,000" users [59].

from this line of research are currently under review for the Transactions of the International Society for Music Information Retrieval (TISMIR): the first presents the first systematic analysis of the users of AI music generation platforms through their prompts and lyrics, while the second examines musical form and harmonic structure in human- and AI-generated music.

During the review process of [3], another important dimension of our research surfaced: the legal and ethical implications of collecting and using large-scale datasets from commercial AI music platforms. One reviewer expressed concern that systematically scraping music and metadata from Suno and Udio might violate their Terms of Service, and questioned whether the release of embedding features and anonymized metadata would constitute a breach. There were also broader concerns about whether such AI-generated content might contain copyrighted material, and whether its redistribution in any form could have legal consequences.

To address these concerns, we consulted legal and research data experts at KTH. The core issue revolves around whether scraping publicly available, user-shared content constitutes a violation of platform terms, and whether releasing abstracted, non-invertible representations of that content — such as audio embeddings and metadata stripped of any user-identifiable information — was permissible under current EU law. Some of the guidance we received was that, as both services explicitly prohibit scraping in their Terms of Service (with Udio extending this restriction even to manual collection), we had to stop our research and seek written permission from Suno and Udio to use the dataset we collected, and to possibly start collecting the dataset from scratch once the permission was obtained. Other guidance was that our work is protected under several legal provisions. The EU Directive 2019/790 on copyright in the Digital Single Market [35] allows exceptions for text and data mining (TDM) for research purposes, even when copyrighted content is involved. Our non-commercial, transformative use of the data falls under recognized research exceptions, similar to the U.S. fair use doctrine [36]. Furthermore, the anonymization and minimization of personal data ensures compliance with data protection regulations, such as GDPR.

We believe that seeking such permission can risk undermining our independence as public researchers. We are not building products or services that compete with these companies, and we are not interested in partnering with them in ways that could compromise our ability to critique their systems. Moreover, the dataset we collected does not include personal or sensitive user information, and the data representations we intend to release cannot be reverse-engineered to recover the original audio. We believe that this work falls squarely within the public interest, and that the potential effect of corporate control over research access must be resisted. Even so, we sent letters to both Suno and Udio asking for permission, explaining our non-commercial intent, commitment to anonymization, and the transformative nature of our research focused on AI

creativity, bias, and detection challenges. At the time of submission, we have not received any responses and thus only release IDs one can use to find the tunes on their websites rather than the extracted embeddings.

The legal and ethical uncertainties surrounding this project are not peripheral but central to the questions this thesis raises. As generative systems become more embedded in cultural production, artists, as well as researchers, will increasingly face conflicts between public inquiry and corporate gatekeeping. Our work provides a concrete example of how such tensions can play out in practice — and highlights the urgent need for clearer legal frameworks and institutional support for independent research on generative AI.

1.3 Research Questions and Overview

This thesis explores how AI-generated music can be **evaluated**, **detected**, and **explained**, through four interrelated studies. The work developed in response to four core challenges encountered during the research:

1. **How can we compare music collections, both across corpora and within them, while maintaining item-level traceability?**

   [1] critiques and extends the CAEMSI method [57] for evaluating symbolic music corpora. By estimating distributions of pairwise distances between pieces, our method supports both corpora-level comparison and item-level analysis.

2. **How can we represent melodic structure in symbolic music using transparent and efficient methods that support both analysis and interpretability?**

   [2] investigates the use of the DCT for encoding melodic contour. Using ITM as a case study, the method is evaluated for its interpretability, efficiency, and suitability for capturing structural features such as phrase direction and overall contour.

3. **How can we reliably detect AI-generated music across platforms, and what are the limits of current detection models?**

   [3] introduces a dataset of $20,000$ tracks from Suno and Udio and evaluates detection models based on audio embeddings (e.g., Contrastive Language-Audio Pretraining (CLAP)[61]), comparing the AI-generated audio recordings to $10,000$ human-made recordings from the MSD [60]. The study reveals significant limitations in generalization across platforms and highlights vulnerabilities to audio transformations such as resampling and filtering. It also examines ethical concerns, including the risk of false positives, as well as legal considerations when collecting the dataset we created.

4. **How do current explainability methods for AI music systems misalign with human expectations, and how can we reconceptualize explainability as communication?**

   [4] reframes explainability as a communication process between model and user, drawing from communication theory to analyze failures in understanding. Presenting case studies in music and healthcare, it emphasizes the need for common ground, iterative feedback, and attention to user assumptions and possible biases.

Together, these works address core technical challenges in music AI: how to represent and evaluate musical structure, how to detect AI-generated music reliably, and how to analyse the behaviour of explainability methods. While each paper focuses on a different aspect, they collectively contribute methods and analyses that support more systematic, interpretable, and robust approaches to working with AI in music.

1.4 Disclosure of AI use for writing this thesis

In the writing of this thesis, I occasionally used AI writing tools (mostly Claude) to help improve clarity and concision of the text. The process of using such tools was to draft the text myself, and when sections felt overly wordy or awkwardly phrased, I sometimes used such tools to suggest alternative wording. I then revised, rewrote, and integrated those suggestions into my own writing. All the narrative, arguments, and the final phrasing are my own. Extensive feedback and suggestions were also provided by my supervisor Bob Sturm, with whom we had back and forth editing sessions through many versions of this thesis.

Background

2 Background

This chapter surveys the current landscape of AI-generated music, highlighting key computational approaches, evaluation challenges, detection issues, and challenges in Explainable Artificial Intelligence (XAI). It also outlines broader cultural, legal, and ethical considerations relevant to this thesis.

2.1 Computational Systems for AI Music

The use of electronic computers to generate music dates back to the mid-20th century, with early algorithmic compositions such as the Illiac Suite [62] for string quartet. Its four movements explore different generative methods, ranging from random number sequences filtered by music compositional rules to Markov chains. Rule-based systems like CHORAL [63] model specific styles, such as Bach chorales, through handcrafted rules. These foundational efforts evolved into statistical models, including Recursive Neural Networks, e.g. [64], and Markov models, e.g. [65], which provide probabilistic frameworks for music generation.

More recently, deep learning has been applied to music generation [66]. Symbolic music generation models such as MusicVAE [67] and MuseGAN [68] focus on latent space representations and multi-track generation. In the audio domain, convolutional approaches started to be explored around 2016 [69]. Transformer-based architectures, exemplified by Jukebox [70] and MusicLM [31], leverage large-scale pretraining and text conditioning for more versatile music synthesis. Recent diffusion-based models including MusicGen [32], Diff-Sound [71], Moûsai [72], and SDMuse [73] further improve audio quality and structure, while alternative approaches such as Musika! [74] demonstrate the potential of GAN-based methods for fast, real-time music generation. Music audio generation is now of sufficient quality that it is being commercialized.

Commercial platforms such as Boomy, Suno and Udio package generative models into end-user services that combine editing, distribution and sometimes even monetization, making it easy for non-experts to create and distribute tracks [5]. The inner-workings of such systems are not publicly available, although it has been speculated that Boomy uses a template-based approach with sample/VST rendering [5], while Suno is likely based on transformer text-to-audio models, like its open-source predecessor Bark[1]. Udio possibly follows a similar approach.

There are multiple ways to taxonomize the commercial AI-music landscape. Moore and Acharya [75], for example, identify three target audiences (consumer, prosumer and professional) and six core use cases (streaming, AI covers, royalty-free generation, music-generation tools, stem splitters and editing tools). Other companies are categorised as providing foundation models, and they also highlight that many products overlap these axes. The commercial landscape itself reveals a diversity of platforms, which can be loosely grouped by their primary function and targeted user-base. For instance, platforms like Suno, Udio, and Boomy prioritize accessibility for casual creators, offering simple text-to-music interfaces, social sharing features, and one-click generation for personal or social media use. In contrast, tools such as AIVA, Amadeus Code, and Beatoven.ai are geared more towards professional content creation, providing advanced controls, commercial licensing, and high-quality export options. Another segment of the market focuses on enterprise and B2B integration, with companies like Soundful and Loudly offering API access and customizable workflows for use in streaming services, wellness apps, or retail. Finally, a distinct category of interactive systems, including Splash and Aimi, enables real-time, responsive music generation for live performances, interactive installations, and adaptive game environments. Complementing such taxonomies, Wilson et al. [29] provide a systematic survey of 27 contemporary GenAI music tools and models, examining not only their creative inputs and outputs but also their transparency, fairness, and ethical implications within the broader context of responsible AI music generation.

Research shows these tools often shift users toward the curation and edition of tracks proposed by the system rather than developing every idea themselves. Suh et al. [76] found that AI serves as "social glue" in collaborative music composition — a neutral intermediary that binds collaborators together by providing shared creative material they can collectively build upon. The AI acts as a mediator that helps mitigate interpersonal stalling and friction while providing a psychological safety net in creative risk-taking. Drott [19] examines the economic implications of AI music tools, analyzing how they disrupt traditional copyright, compensation, and commons structures in the music industry. Morreale [21] raises broader ethical and political concerns about responsibility

[1]See the Bark model on GitHub: `https://github.com/suno-ai/bark`, last accessed Aug. 19, 2025.

and accountability in AI-assisted music creation. Together, these studies highlight concrete concerns about attribution, remuneration and deskilling in music labour. The broad study of AI music is advocated in [18] under the name of "AI music studies".

2.2 Evaluation and Representation Challenges

This subsection discusses two important aspects related to analyzing AI-generated music: first, how to evaluate the output of Style Imitation (SI) systems, and second, how to represent melodic data effectively for analysis. Both are central to understanding and improving the outputs of SI systems and to supporting tasks in Music Information Retrieval (MIR). I use the term SI systems here as it reflects the terminological conventions found in the relevant literature in [1], which focuses on the comparison and evaluation of musical style across corpora and individual items.

2.2.1 Evaluation of SI Systems

Evaluating music generated by SI systems involves determining how well the output reflects the stylistic features of a target corpus. Quantitative approaches compare distributions of low-level musical features using metrics such as Kullback-Leibler divergence applied to pitch distributions, as demonstrated by Yang and Lerch [77] for generative models in music, or the Normalized Compression Distance (NCD), as implemented by Ens and Pasquier [57], which offers another quantitative measure of similarity. These approaches are particularly useful for analyzing pitch and rhythm statistics, though their relevance for comparing higher-level musical structures is not clear.

Beyond statistical comparisons, analytical methods grounded in music theory and practice offer deeper stylistic assessment. Some approaches focus on detecting excessive replication of source material to evaluate originality [78], while others develop cross-domain frameworks for ranking stylistic similarity [79]. Other work emphasizes the importance of evaluation methodologies that connect directly to musical practice [80, 81], examining how well generated outputs adhere to stylistic conventions.

Listening tests can offer perceptually grounded assessments and are widely used to validate stylistic fidelity [82, 83]. Despite their value, they are costly in terms of time and effort and are difficult to scale for evaluating the extensive output typical of modern SI systems. The 'AI Music Generation Challenges' (2020–2023) [84–87] provide a useful counterpoint to purely statistical tests. For example, the 2020 challenge [84] focused on Irish double jigs, 2021 [85] on Swedish slängpolskas, and 2022 [86] on Irish reels (with added sub-challenges for automated judging and title generation). Submissions were rendered as ABC/MIDI/score/audio, and were judged by expert practitioners for low-level correctness and plagiarism as well as higher-level melody, structure, memora-

bility and playability on traditional instruments [88]. Given the rich human expert evaluations from these challenges, we found Irish Traditional Music (ITM) to be a compelling case study for analysing systems like folk-rnn [89], and using the expert judges' assessments as human evaluation.

2.2.2 Musical Representations for Melodic Analysis

Melodic representation has relied on symbolic encodings and geometric methods that model melodies as shapes in multidimensional spaces. Some algorithms identify repeated patterns by comparing pitch–onset vectors [90, 91], while others represent melodic structure as graphs or point sets [92, 93]. Research on folk music variation shows that stylistic identity often hinges on short melodic segments rather than global form [94–97], and listeners tend to recognize melodies based on recurring motifs and phrases [98].

This focus on small musical units is particularly relevant in ITM. Doherty's manual analysis of double jigs in O'Neill's collection [99] finds four standard melodic structures and argues that nearly half of all musical motifs are repetitions or variations. These patterns of repetition suggest that such melodies can be represented efficiently using compact mathematical forms that capture shared features across tunes. David Huron's work on melodic shape [100] emphasizes that the direction of melodies can matter more than exact pitches, identifying nine common melodic patterns based on whether phrases go up, down, stay flat, or form arch-like shapes. This focus on overall shape provides a link between symbolic music notation and how people percieve melodies.

Basis function representations of melodic contours offer a systematic way to capture these shape patterns and repetitive structures. A melodic contour is a sequence of pitches ordered in time [100]. By expressing melodic contours as combinations of mathematical building blocks, basis functions move an analysis from discrete musical notes to continuous mathematical spaces where patterns and variations become easier to analyse. This approach can work particularly well for ITM, with its structured melodic patterns and well-defined phrasing — particularly in forms like double jigs.

Cornelissen et al. [58] propose decomposing melodic contours over a cosine basis for tasks such as visualization, classification and summarization. They apply this approach to melodies from four traditions (Gregorian chant, German, Chinese and Sioux folk songs) and argue that they can be used for comparing contours across traditions due to its data-independent, culturally neutral nature. However, we find that melodies in ITM are less efficiently represented in a cosine basis [2] due to their "undulating" shape.

2.3 Detecting AI-Generated Music

The detection of AI-generated music is an emerging challenge, with implications for copyright enforcement, content moderation, and recommendation sys-

tems. Recent work explores deep learning models for AI music detection, primarily using Convolutional Neural Networks (CNNs) or transformer-based architectures. For instance, Rahman et al. [101] propose the Spectro-Temporal Tokens Transformer (SpecTTTra), a transformer-based model that processes spectrograms along time and frequency axes to capture long-range dependencies. They show their method achieves a state-of-the-art 97% F1-score on a structured dataset of 97k tracks they generate using Suno and Udio. Li et al. [102] survey detection methods across audio deepfakes and AI music, noting that current systems struggle with generalization across different generative models. Similarly, Afchar et al. [48] investigate deepfake detection in music by training classifiers to identify artifacts from neural vocoders (e.g., Encodec, DAC). While their approach achieves high accuracy (>99%) on reconstructed audio, performance degrades under common audio manipulations (e.g., pitch shifting, re-encoding), highlighting vulnerabilities in the detectors. In subsequent work [103], they provide theoretical grounding for these observations by analyzing spectral artifacts in generative CNNs. Their analysis shows that deconvolution layers inherently produce peaking artifacts in output spectra, regardless of training data or model weights. This fundamental limitation suggests that certain detection cues may persist across architectures, though nonlinear activations and skip connections could potentially mitigate these effects.

Commercial efforts to detect AI-generated music are rapidly emerging across the industry. IRCAM Amplify [104] has deployed an AI music detector claiming 97.6% accuracy, designed for integration into streaming platforms and distribution chains to automatically tag AI-generated tracks. Deezer launched its own detection tool in early 2025 [105], and now uses it to exclude such content from music recommendations. Music distributor Believe has developed "AI Radar" [106], with CEO Denis Ladegaillerie positioning this as essential protection against the flood of AI content. These initiatives reflect broader industry concerns about content authenticity and revenue protection, with platforms increasingly adopting detection technologies similar to YouTube's Content ID [107] system for managing copyrighted content. Established content identification companies like Audible Magic are also expanding their services to address AI-generated music detection alongside traditional copyright enforcement, and have secured partnerships with leading AI music startups [108, 109]. However, this partnership seems to focus solely in copyright-protected audio that users might input into the platforms rather than the fingerprinting of the outputs.[2]

AI music detection can be seen as a kind of "composer identification" task [110–112], which aims to determine the creator of a musical piece. Another relevant task is music auto-tagging [113–115], which demonstrate how learned embeddings (e.g., from CLAP [61]) can capture high-level musical attributes that may

[2]This was made clear by representatives of Udio attending ISMIR 2025 at Daejeon, South Korea, when I directly questioned them about their partnership with Audible Magic. Further discussions with Audible Magic representatives present at the conference confirmed it.

differ between human and AI creators. However, as generative models become better at mimicking human music, detectors must evolve to identify subtler artifacts [31, 32].

Ethical and practical considerations of AI music detection remain understudied. Deploying detection systems at scale risks false positives that can disproportionately affect "legitimate" creators, echoing biases observed in AI-text detection [116]. Moreover, the definition of "AI-generated music" itself can be contested [3] — whether it encompasses human-AI collaborations, edited outputs, or fully autonomous works.

2.4 Explainable Artificial Intelligence

The field of XAI has gained substantial traction following the introduction of the European Union's General Data Protection Regulation (GDPR) [117], which formalized the right of citizens to receive explanations for algorithmic decisions. This was propelled not only by technological advances but also by ethical and regulatory imperatives [118–122]. In principle, explainability can serve multiple functions: it can facilitate debugging and development, expose algorithmic biases, provide avenues for recourse when decisions negatively impact individuals, support trust calibration in decision-making scenarios, and help vet systems prior to real-world deployment [118, 120].

Goal	Definition
Trustworthiness	Whether a user accepts a model's predictions and behaviour enough to take some action based on it.
Causality	The relationship between cause and effect among data variables.
Transferability	The ability of a user to use the knowledge gained while solving one problem in another problem.
Informativeness	Information given by the system on its internal functioning.
Confidence	How much the user has faith or relies on the model.
Fairness	Relating to the impartiality of a model's decisions without favouritism or discrimination.
Accessibility	The quality of being able to understand a system.
Interactivity	The ability to influence the model.
Privacy awareness	Gain insight into how much of the training data has been captured and stored by the model, potentially compromising the privacy of the data.

Table 2.4.1: Explainability goals and their definitions. This table is an adaptation from [118].

Audience	Description
Domain experts	People who know and understand the essential aspects of a specific field of inquiry.
Users	People who use the model.
Regulatory entities	Agencies that are responsible for enforcing the laws, rules, or regulations.
Managers and executives	Those with the power to control or administer the project.
Developers	Those who devise and create the models.

Table 2.4.2: Main target audience and their definitions. The content of this table is adapted from [118].

Type	Description
Simplification	Building a new simplified model that is optimized to function similarly to the model to explain.
Local explanation	Partitioning the solution space and giving explanations to less complex solution subspaces.
Text explanation	Generating text explanations that help explain the results from the model.
Visual explanation	Generating human-understandable visualizations of the model's behaviour.
Feature relevance	Computing how much influence a variable has upon the output of the model.
Explanation by example	Extracting samples following the inner relationships found by the model.
Transparent models	Models that are understandable by themselves.

Table 2.4.3: Explanation types and their definitions. The content of this table is adapted from [118].

Category	Approach
Application-grounded	Real humans assessing the quality of the explanation in the true context of the explanation's end task.
Human-grounded	Real humans assessing the quality of the explanation in simplified tasks rather than the real end task.
Functionally-grounded	No human subjects involved. Proxy tasks are used to prove some formal definition of interpretability. This includes complexity and fidelity. Complexity relates to how understandable the explanations are. Fidelity relates to how informative the explanations are, whether the original system's behaviour can be replicated with only the explanations.

Table 2.4.4: Evaluations of explanations and their definitions. The content of this table is adapted from [120].

XAI can be analysed through a taxonomy of four key dimensions: explainability goals, target audiences, explanation types, and evaluation strategies. Those are explained in more detail in [118, 120], and are described in Tables 2.4.1 to 2.4.4. The primary goals of XAI, detailed in Table 2.4.1, vary widely depending on stakeholder needs. For instance, causality aims to identify feature relationships that drive model predictions, such as tracing musical emotion recognition to specific spectral features [123]. Trustworthiness, critical for user adoption, requires explanations that align with human intuitions [55]. The target audience for an explanation directly shapes its form and content, as each group has distinct priorities and expertise [124]. The primary audiences and their concerns are detailed in Table 2.4.2. A developer requires technical details for debugging, a regulator needs to verify compliance and fairness, and an end-user needs a simple rationale to support trust. The types of explanations generated, categorized in Table 2.4.3, fall into two broad categories: post-hoc techniques for interpreting black-box models and transparent-by-design architectures. Post-hoc methods, such as LIME [125] or SHAP [126], approximate model behavior locally or globally through simplification. Transparent models, like decision trees or rule-based systems, expose their logic inherently but may sacrifice predictive power. Both approaches struggle with miscommunication: the misalignment between a system's explanation and a user's interpretation. For instance, saliency maps may highlight irrelevant features due to dataset biases, misleading users about a model's true decision criteria [127]. This misalignment underscores the need for domain-informed evaluation to ensure explanations reflect true model behaviour rather than artifacts of flawed training

data. Evaluating explanations remains contentious due to the lack of universal standards. Table 2.4.4 categorizes evaluation strategies. Application-grounded methods assess real-world user utility, while functionally-grounded (i.e., without user studies) measure fidelity or complexity [120].

The practice of XAI often overlooks critical dimensions, such as the target audience for an explanation or its specific goal [128]. Most research and practice in XAI seems to use the researchers' intuitions of what constitutes a "good" explanation rather basing the approach on the human explanation process [4, 55, 56]. More specifically, XAI lacks definitions, assumptions are implicit, and evaluation is ill-defined [120]. These challenges are further complicated by what is described as a "semantic gap": the disconnect between high-level human concepts and low-level machine representations [129]. For example, while text prompting strategies have gained popularity in generative AI, linguistic inputs may not align neatly with core musical properties like timbre, structure, or expression, limiting their reliability in steering model outputs [130].

In response, some researchers have proposed interaction-based approaches to XAI, where users explore model behaviour through real-time feedback and embodied input [131, 132]. Systems that link user gestures, sketches, or movement to model outputs can help establish intuitive mappings within a shared control space [133–135]. These systems frame explainability not as a static output but as an iterative process embedded in creative exploration.

However, such approaches introduce new technical and design challenges. Mapping high-dimensional sensor data onto generative model spaces remains difficult [136, 137], and the interpretability of these mappings varies across users and contexts [138, 139]. Recent work in the XAIxArts community [140] underscores the need for explainability frameworks that address not just technical transparency but also user agency, aesthetics, and situated interaction. These approaches shift the emphasis from explanation as static output to explanation as an exploratory process embedded in artistic practice.

Comparisons and Representations of Music

3 Comparisons and Representations of Music

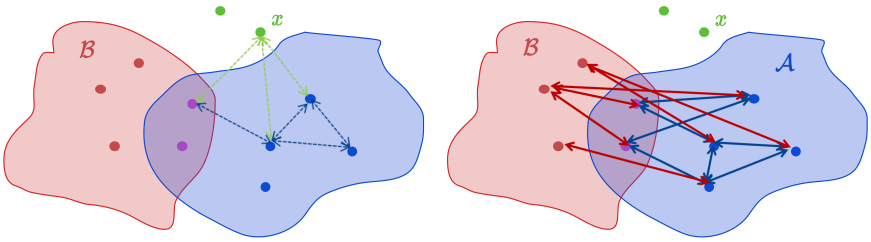


Figure 3.0.1: Visual representation of the core problems addressed in this chapter. **(Left):** Comparing an item x (green) to a corpus $\mathcal{A}$ (blue). **(Right):** Comparing two corpora $\mathcal{A}$ (blue) and $\mathcal{B}$ (red).

This chapter reviews the key findings in [1] and [2], defining the problems they address, critiquing the work, and suggesting future developments. Consider a corpus of music, denoted $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$, consisting of individual musical items. Let x denote a new item of interest, and $\mathcal{B} = \{b_1, b_2, \dots, b_m\}$ another corpus. Two central problems then arise. Item-to-corpus comparison asks whether x belongs to $\mathcal{A}$, i.e., to evaluate the likelihood that $x \in \mathcal{A}$. Corpus-to-corpus comparison asks whether two corpora $\mathcal{A}$ and $\mathcal{B}$ are stylistically similar. These tasks are visualized in Figure 3.0.1. In [1] we aim to provide domain-independent (applicable beyond music to other sequential data), scalable (computationally efficient for large datasets), and statistical (data-driven and probabilistic) methods to answer those questions.

Section 3.1 summarizes my framework for distance-based comparison of musical items and corpora, relating it to existing methods like CAESMI [57] and

StyleRank [79]. Section 3.2 summarizes our work in [2] where we use the Discrete Cosine Transform (DCT) for representing and analyzing melodic contour. Throughout, I use case studies from ITM and AI-generated melodies [89] to illustrate the methods and provide a critical discussion of their contributions, limitations, and future potential.

3.1 Comparing Items and Corpora

To address the problem of comparing items and corpora in a domain-independent, scalable, and statistical manner, we develop a framework based on distance matrices and distribution analysis. This approach allows us to move beyond corpus-level comparisons and provide a method for assessing individual items against a reference corpus, a capability not offered by reviewed domain-independent techniques.

3.1.1 Problem Description

	Domain Independent	**Domain Dependent**
Corpus-to-corpus	CAEMSI [57], and **Our contribution** [1]	Yang & Lerch [77]
Item-to-corpus	**Our contribution** [1]	StyleRank [79]

Table 3.1.1: Classification of methods for comparing items and corpora.

Table 3.1.1 relates four methods. We categorize methods for comparing melodies along two dimensions: their reliance on domain knowledge and their scope of comparison. Our work [1] addresses the gap in domain-independent methods for comparing individual items to corpora, while maintaining the ability to do corpus-to-corpus comparisons.

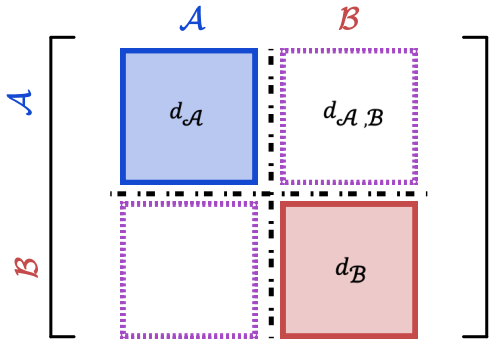


Figure 3.1.2: The distance matrix constructed from pairwise comparisons, which forms the basis for the statistical analysis. The colours align with the ones in Figure 3.0.1: distances within corpus $\mathcal{A}$ in blue and distances within corpus $\mathcal{B}$ in red

CAEMSI [57] provides a domain-independent framework for corpus-to-corpus comparison, using statistical tests based on "typicality" — the degree to which generated artifacts represent their source corpus. It employs two complementary tests: a test for difference, for which the null hypothesis is $H_0 : \mathcal{A} = \mathcal{B}$, and a test for equivalence, for which the null hypothesis is $H_0 : \mathcal{A} \neq \mathcal{B}$. To do this, it first constructs a distance matrix $\mathbf{D}$, an $(n \times n)$ symmetric matrix where $n = |\mathcal{A}| + |\mathcal{B}|$ and each element d_{ij} represents the computed distance between items i and j. This matrix, shown in Figure 3.1.2, is then partitioned into three sub-matrices: $\mathbf{d}_{\mathcal{A}} = d(i,j) \mid \forall i,j \in \mathcal{A}, i \neq j$, representing all within-corpus distances for $\mathcal{A}$; $\mathbf{d}_{\mathcal{B}}$, defined analogously for $\mathcal{B}$; and $\mathbf{d}_{\mathcal{AB}} = d(i,j) \mid \forall i \in \mathcal{A}, \forall j \in \mathcal{B}$, representing all between-corpus distances. The statistical significance of an observed test statistic T_{obs}, which we refer the reader to the publication [57] for further details, is evaluated using permutation testing. This non-parametric procedure repeatedly simulates the null hypothesis by randomly shuffling the rows and columns of the distance matrix, and recalculating the statistic T_{perm} for each permutation. This generates an empirical null distribution against which T_{obs} is compared. This method's principal strength is its distribution-free nature, requiring no assumptions about the data. Its primary limitation, however, is that it provides only a holistic assessment of significance, offering no insight into which specific items or individual distances within the matrix contribute to a rejection of H_0. Consequently, CAEMSI is designed for corpus-level comparison and offers limited utility for assessing individual items against a corpus.

StyleRank [79] measures how well a symbolic music item x matches a style represented by a "background" corpus $\mathcal{A}$. They do so by first converting pieces into many music theory informed feature representations, then training random-forest classifiers to discriminate the items in the target corpus from the chosen

background or rank set. Each forest maps an item to a sparse leaf-indicator vector (an embedding), and StyleRank computes a similarity score for x by averaging cosine similarities between x's embedding(s) and the embeddings of corpus items across the feature forests. Because the output depends on which symbolic features are chosen and which background corpus is used, it requires domain-specific feature engineering and careful corpus curation. Its effectiveness therefore depends on whether the chosen features capture the perceptual cues that define the style and the representativeness of the background corpus — limitations the authors note even as they report strong agreement with human judgments in their experiments.

Yang and Lerch [77] propose a corpus-level framework that extracts musically informed features from symbolic music representations and computes pairwise distances to form intra-set (within target) and inter-set (between target and generated) distance distributions per feature. They estimate those distributions with kernel density estimation and summarize their difference using the Kullback–Leibler divergence and the area of overlap to quantify how closely a generated corpus matches the target. The approach yields interpretable, per-feature diagnostics that are reproducible and useful for determining which musical dimensions a generator captures or fails to capture. However, it is strictly corpus-level and its conclusions depend on the chosen features, distance metric, KDE parameters, and dataset representativeness; it cannot assess individual items or features outside the ones encoded.

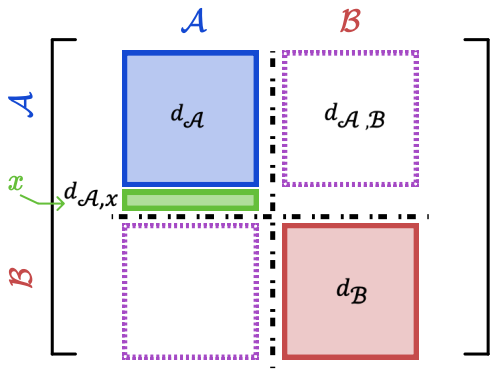


Figure 3.1.3: The distance matrix constructed from pairwise comparisons, with the addition of the item x compared to corpus $\mathcal{A}$. The colours again align with the ones in Figure 3.0.1: distances within corpus $\mathcal{A}$ in blue, distances within corpus $\mathcal{B}$ in red, and distances between item x (green) and corpus $\mathcal{A}$ in green.

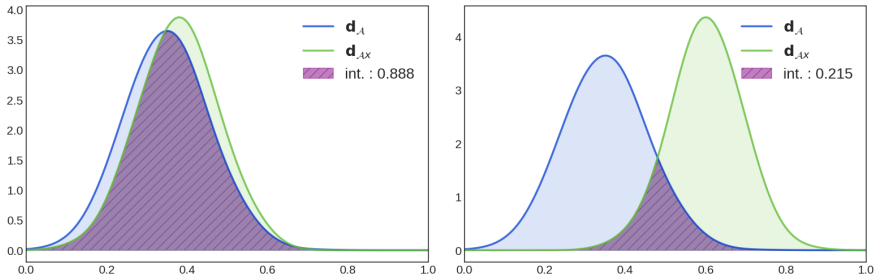


Figure 3.1.4: Kernel density estimation (KDE) intersection for item-to-corpus comparison. The x axis represents the normalized distances, while the y axis represents the density. The probability that item x belongs to corpus $\mathcal{A}$ is proportional to the overlapping area between the distribution of pairwise distances within corpus $\mathcal{A}$ (denoted $\mathbf{d}_{\mathcal{A}}$, shown in blue) and the distribution of distances between item x and all items in $\mathcal{A}$ (denoted $\mathbf{d}_{\mathcal{A}x}$, shown in green). Left panel shows a high similarity scenario where the intersection area is 0.888 and thus over 0.5, indicating that item x likely belongs to corpus $\mathcal{A}$. Right panel demonstrates a low similarity scenario where the intersection area is 0.215 and thus below 0.5, suggesting that item x is unlikely to belong to corpus $\mathcal{A}$.

Our work [1] aims to contribute a domain-independent item-to-corpus comparison method, filling the void seen in Table 3.1.1, i.e. domain independent item-to-corpus comparison, as well as domain independent corpus-to-corpus comparison. As illustrated in Figure 3.0.1, our approach handles two distinct scenarios. In the first scenario (left panel in Figure 3.0.1), in order to determine whether a specific item x (shown in green) belongs to a corpus $\mathcal{A}$ (blue region), we use kernel density estimation (KDE) to model two distributions: the distribution of pairwise distances within the corpus $\mathcal{A}$ (denoted $\mathbf{d}_{\mathcal{A}}$), and the distribution of distances between the item x and all items in $\mathcal{A}$ (denoted $\mathbf{d}_{\mathcal{A}x}$). As seen in Figure 3.1.3, this can be seen as adding a row to the distance matrix. We compute the probability that x belongs to $\mathcal{A}$ by measuring the overlapping area (intersection) between these two KDE curves, as illustrated in Figure 3.1.4. In the second scenario (right panel in Figure 3.0.1), when comparing two corpora $\mathcal{A}$ and $\mathcal{B}$, we analyse three sets of pairwise distances: $\mathbf{d}_{\mathcal{A}}$, which contains all distances between items within $\mathcal{A}$; $\mathbf{d}_{\mathcal{B}}$, containing distances within $\mathcal{B}$; and $\mathbf{d}_{\mathcal{A}\mathcal{B}}$, which includes distances between every item in $\mathcal{A}$ and $\mathcal{B}$. These distance sets are treated as samples, and statistical tests like the Mann-Whitney U (MW-test) or Kolmogorov-Smirnov (KS-test) evaluate whether $\mathbf{d}_{\mathcal{A}}$ and $\mathbf{d}_{\mathcal{B}}$ share the same underlying distribution, with the null hypothesis that the distribution underlying $\mathbf{d}_{\mathcal{A}}$ is stochastically less than the distribution underlying $\mathbf{d}_{\mathcal{A}\mathcal{B}}$.

3.1.2 Key Findings

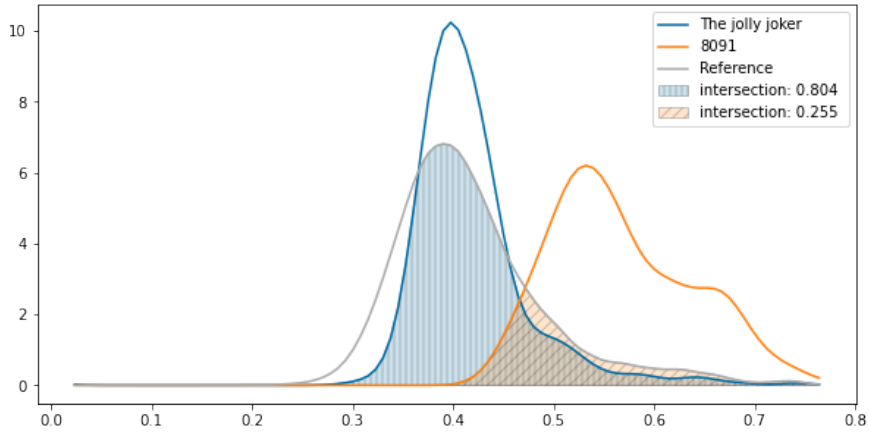


Figure 3.1.5: Distribution comparison showing the overlap between "The Jolly Joker" (O'Neill's No. 229) and folk-rnn tune No. 8091 with the reference corpus of O'Neill's using normalized Levenshtein distance.

In [1] we evaluate how our method performs using Irish Traditional Music (ITM). We test three distance measures — Normalized Compression Distance (NCD), Levenshtein distance, and cosine distance from neural embeddings — on O'Neill's 365 double jigs [141] against 365 tunes generated by folk-rnn [89], which were selected by an artificial critic [88]. We find that our approach can distinguish between human-composed and AI-generated corpora at the corpus level, with normalized Levenshtein distance proving most effective. Statistical tests (Kruskal-Wallis and Kolmogorov-Smirnov) consistently show smaller, more uniform distances within O'Neill's corpus compared to cross-corpus distances, indicating tighter stylistic cohesion. For the item to corpus analysis, our method correctly shows that O'Neill's tunes have a higher probability of belonging to their corpus, as measured by a larger KDE intersection area, just as folk-rnn tunes do for theirs.

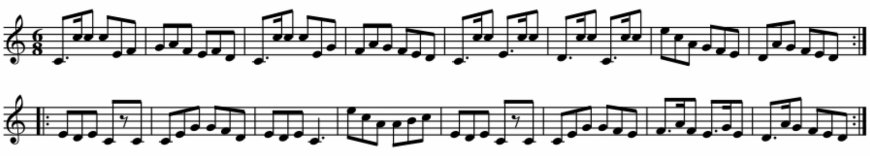


Figure 3.1.6: "The Jolly Joker," double jig No. 229 in O'Neill's, transposed to Cmaj from Dmaj. Expert evaluation deemed this tune musically deficient, though it is in a recognized collection, a quality our statistical method failed to capture.

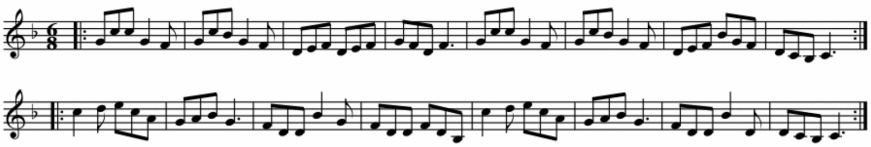


Figure 3.1.7: Folk-rnn tune No. 8091, the winner of the 2020 AI Song Contest challenge [84]. This tune occupies a central position in both Figures 3.1.8 and 3.1.9, suggesting structural similarity to many other tunes in the datasets.

As shown in Figure 3.1.5, the tune "The Jolly Joker" (O'Neill's No. 229) has a high probability of belonging to the reference corpus, in this case O'Neill's, with an intersection of 0.804. This tune (Figure 3.1.6) serves as an interesting case study, as it is considered by expert judges as not being very representative of O'Neill's. Conversely, the folk-rnn tune No. 8091 (Figure 3.1.7) has a low intersection value of 0.255, correctly suggesting that it does not belong to the reference corpus of O'Neill's. However, this particular tune won The AI Music Generation Challenge 2020 [84], representing what experts considered an AI-generated tune that aligns with ITM. These two contrasting examples highlight the complex relationship between our computational similarity measures and musicological assessment, which we expand on in Section 3.1.3.

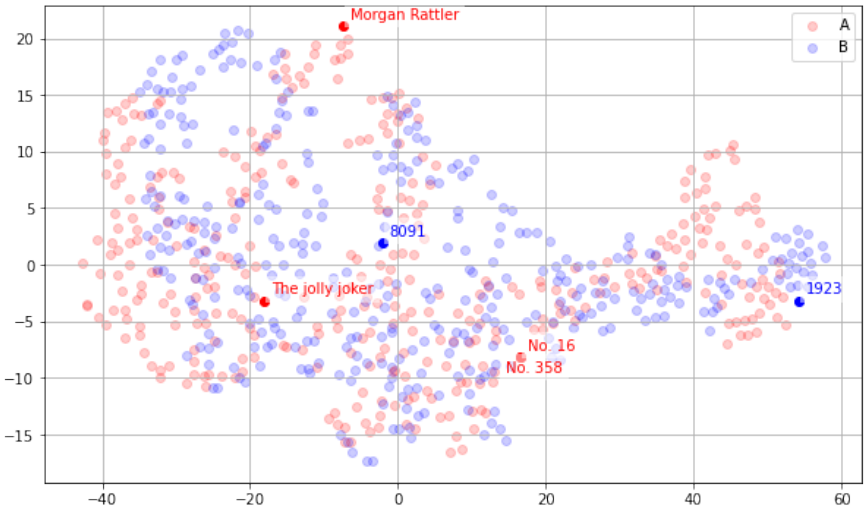


Figure 3.1.8: t-SNE visualization of the O'Neill's (red) and folk-rnn (blue) corpora based on pairwise Levenshtein distances between symbolic sequences. Several points of interest are annotated: folk-rnn tune No. 8091 (winner of the 2020 challenge) appears near the center of both datasets, suggesting broad similarity across corpora; "Morgan Rattler" and folk-rnn No. 1923 appear as isolated outliers, reflecting their unusual characteristics (exceptional length and missing repeat signs, respectively); and O'Neill's tunes No. 16 and No. 358, known duplicates, cluster closely together.

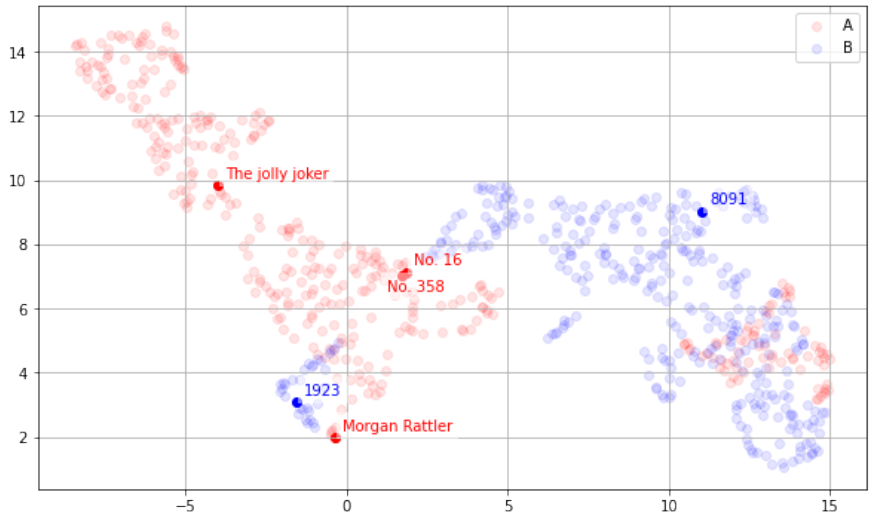


Figure 3.1.9: UMAP visualization of the O'Neill's (red) and folk-rnn (blue) corpora. Compared with the t-SNE projection, this representation places folk-rnn tune No. 8091 away from O'Neill's cluster, although it is still somewhat at the center of the folk-rnn cluster. "Morgan Rattler" and folk-rnn No. 1923 again appear as isolated points within their respective corpora, while the duplicate tunes No. 16 and No. 358 from O'Neill's remain tightly clustered.

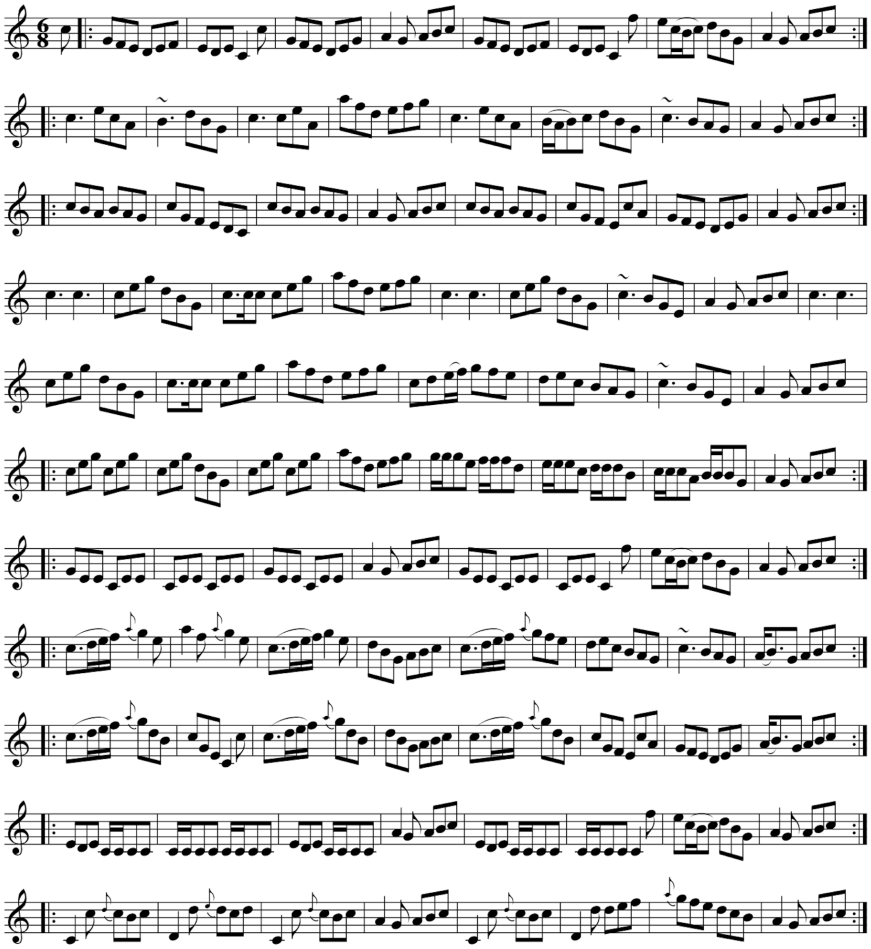


Figure 3.1.10: "Morgan Rattler," double jig No. 257 in O'Neill's 1001, transposed to Cmaj from Dmaj. Its unusual length causes it to appear as an outlier in both Figures 3.1.8 and 3.1.9, highlighting the sensitivity of the distance metric to sequence length.

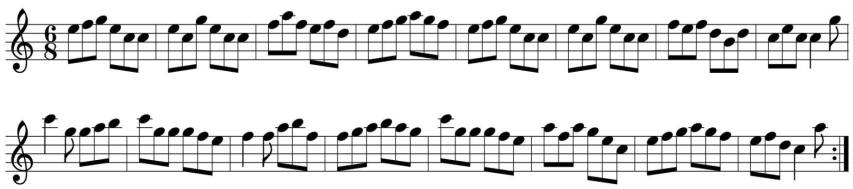


Figure 3.1.11: Folk-rnn tune No. 1923, which is missing repeat signs. This irregularity contributes to its isolated position in both Figures 3.1.8 and 3.1.9, showing how sequence-level structural anomalies are accentuated by the Levenshtein distance.

We visualize the relationships among all items in $\mathcal{A}$ and $\mathcal{B}$ by applying dimensionality reduction via t-distributed Stochastic Neighbor Embedding (t-SNE) [142] and Uniform Manifold Approximation and Projection (UMAP) [143]. Although the scope of this analysis is preliminary, we can see some interesting results. Firstly, O'Neill's tunes No. 16, titled "When sick is tea you want?", and 358, titled "Go to the Devil and shake yourself", are known duplicates, which both visualizations in Figures 3.1.8 and 3.1.9 show by clustering closely together. If we look at the t-SNE visualization of the corpora in Figure 3.1.8, we can see how folk-rnn tune No. 8091 — the 2020 challenge [84] winner (see Figure 3.1.7) — is positioned of at the centre of both corpora. However, in the UMAP visualization (Figure 3.1.9), No. 8091 is at the centre of the folk-rnn cluster, but it is not overlapping with the O'Neills cluster. This divergence illustrates how different dimensionality reduction methods can greatly affect the relationships one can infer from such visualizations, making this approach questionable to say the least. Other examples provide additional insights into the behavior of the specific distance metric used (in this case Levenshtein distance). The tune "Morgan Rattler", which is unusually long (see Figure 3.1.10), and folk-rnn No. 1923, which is missing repeat signs (see Figure 3.1.11), appear as isolated points in both Figures 3.1.8 and 3.1.9. This seems intuitive at first, as they are indeed odd: one stands out because of its odd length, and the other because of its structural irregularity. However, it is unclear to what extent these characteristics translate into genuine musical atypicality within their respective corpora. This sensitivity is most likely due to the nature of the Levenshtein distance. This distance captures the minimal sequence of edit operations (insertions, deletions, substitutions) required to transform one symbolic token sequence into another. After normalizing for length, it therefore becomes particularly sensitive to sequence-level differences such as extra or missing measures, omitted repeat signs and unusually long phrases.

3.1.3 Critical Limitations and the Musicological Gap

Despite promising statistical results, our approach suffers from a fundamental disconnect between computational and musicological evaluation. A telling example is "The Jolly Joker" (see Figure 3.1.6), which expert judges at The AI

Music Generation Challenge 2020 [84] identified as "not a good tune" from a corpus that "has plenty of poor and dull tunes in it". Despite being in the corpus, the judges deemed it "neither memorable nor interesting," which explains why no one plays it today. In Figure 3.1.5 we see how our item to corpus comparison approach finds this tune to most likely belong to the O'Neill's corpus, suggesting that whatever makes it musically deficient is not captured by surface-level string similarity measures. Conversely, folk-rnn tune No. 8091 — the 2020 challenge [84] winner — shows to be most likely not from the reference corpus of O'Neill's. This result seems paradoxical: if the tune lacked the stylistic qualities of ITM, it would hardly have impressed the challenge judges enough to earn the first place.

This failure exposes a core limitation of domain-independent approaches: they can detect patterns but miss the deeper structural and aesthetic qualities that contribute to musical style. The Levenshtein distance treats all symbol changes equally, making no distinction between musically meaningful variations (such as ornamentations) and structural deviations. Similarly, while NCD captures shared compression patterns, it cannot differentiate between coincidental repetition and intentional musical structure.

The visualization methods, while effective at clustering similar items, provide little interpretive insight into *why* certain tunes cluster together or what musical features drive the observed groupings. Because of the domain-independent nature of our approach, we have an interpretability gap that limits the method's use for musicological analysis — we can identify outliers but cannot explain their musical characteristics or significance. Moreover, different dimensionality reduction techniques provide dramatically different representations, so the usability of such is unclear.

3.1.4 Future work

The work here [1] suggests multiple directions. One immediate goal is to move beyond generic, sequence-level comparisons toward domain-sensitive measures that respect the formal properties of tunes in ITM: parts, repeats, first/second endings, and anacrusis all alter symbolic sequences in musically irrelevant ways yet can distort distances computed by metrics such as Levenshtein distance. Future work should therefore investigate domain-dependent distance measures that encode part structure and inter-part relationships, and should study how different choices in distribution comparison (KDE, intersection areas, etc.) affect conclusions.

Another promising direction lies in combining our statistical framework with approaches like Doherty's work on melodic contour analysis [99], which reveals how melodic shape patterns function in ITM. Where our method identifies statistical outliers, Doherty's contour-based analysis could explain *what* makes them musically distinctive.

For instance, cosine-based decomposition of melodic contours [58] — rather than raw symbol sequences — could capture the phrase-level patterns that appear in ITM. This approach would maintain computational scalability while incorporating meaningful musical features like stepwise motion patterns, phrase boundaries, and motivic development that our current string-based methods overlook.

Future work should investigate multi-resolution approaches that integrate several complementary methods. Surface-level statistical comparisons, as developed in this study, can provide efficient large-scale filtering. Contour-based analysis can capture melodic shape characteristics, while structural analysis can reveal patterns in phrases and overall part forms. Additionally, probabilistic modelling can be applied to identify features that define stylistic tendencies.

The integration of these approaches promises to address both the scalability requirements of analysing thousands of AI-generated tunes and the interpretability needs of musicological analysis — ultimately enabling an effective "artificial critic" capable of identifying not just statistical outliers, but musically meaningful patterns in traditional and generated repertoires [144].

3.2 Representing Melodies with Basis Functions

We now summarize the work in [2]. This work is motivated by a problem similar to that of Cornelissen et al. [58]: finding an effective, interpretable representation for melodic contour. Representing a melody in a new basis can help to better learn its global shape versus local details and reveal patterns that are obscured in the original representation. This section summarizes our work on applying the Discrete Cosine Transform (DCT) to melodic contours in ITM, exploring its potential for providing an interpretable and efficient description of melodic style.

3.2.1 Problem Description

Cornelissen et al. [58] present a domain-independent method for melody representation and comparison based on *cosine contours*. Their work is motivated by the need for a robust, interpretable, and efficient representation of melodic contour that is applicable across diverse musical traditions. However, while they validated their cosine contour representation on a diverse set of corpora of notated melodies, we wanted to investigate how well it specifically works ITM, and whether it can reveal patterns distinct from those in other corpora. Their method involves two main steps. First, a melodic contour is extracted from a symbolic melody by converting it into a step function: they extract note onsets (in quarter notes) and pitches (in MIDI semitones), interpolate a step function through these points, and then sample N = 100 equally spaced points from this function. The sequence is expressed as a vector $x = (x_0, x_1, \ldots, x_{N-1})$, which

is the melodic contour. This representation makes several key assumptions: it ignores rests, normalizes the total duration of all contours to 100 samples (retaining relative internal durations) and assumes an Euclidean distance space for comparability. This contour vector is then transformed using the Discrete Cosine Transform (DCT). The DCT coefficients serve as a new feature space for the melody.

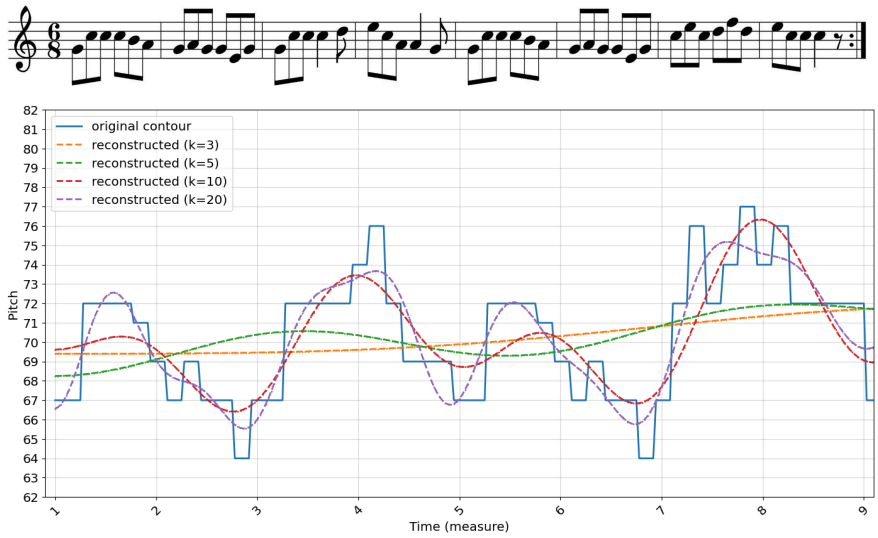


Figure 3.2.12: First 8-bar part of "When sick is tea you want?", tune No. 16 in O'Neill's 1001 [141]. Score (top) with its cosine representations (below), with the original melodic contour (blue), as well as reconstructions using 3 (orange), 5 (green), 10 (red) or 20 (purple) cosine coefficients.

The DCT represents melodic contours as sums of cosine functions with varying frequencies. For a sequence of pitch values $x = (x_0, x_1, \ldots, x_{N-1})$, the Type-II DCT computes coefficients X_k via:

$$X_k = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right], \quad k = 0, 1, \ldots, N-1.$$

Lower-order coefficients ($k \ll N$) encode global trends, while higher-order terms capture local fluctuations. Figure 3.2.12 illustrates this decomposition for the firs 8-bar part of the double jig No. 16 in O'Neills *When sick is it tea you want?*. The original contour is shown in blue, while dashed reconstructions illustrate the effect of including progressively more coefficients. With only three coefficients (orange), the reconstruction captures little more than the overall ascending trajectory of the melody, while five coefficients (green) starts suggesting a two-part structure. Using ten coefficients (red) begins to reproduce

the repeating two-bar motif structure that shapes the tune, and twenty coefficients (purple) add still finer detail, closely following the undulating contour of the original melody. For this particular melody, the first 52 coefficients of the DCT are required to explain 95% of the variance, which is relatively inefficient compared to PCA for longer segments [58].

A central claim in [58] is that the fixed cosine basis of the DCT closely approximates the optimal, data-driven basis found by Principal Component Analysis (PCA). They provide empirical validation for this by comparing the reconstruction error and explained variance of both methods. Their results show that the reconstruction error of DCT-based reconstructions closely approximates that of PCA across different contour lengths (motifs, phrases, and tunes of a Gregorian chant corpus). They report that the effective dimensionality — the number of dimensions needed to explain 95% of the variance — is very low for the cosine representation: just 1 dimension for motifs, 9 for phrases, and 61 for tunes. This demonstrates that low-dimensional DCT coefficients provide a compact and highly effective summary of a melody's global shape (e.g., its overall ascending or descending trend, or its archedness). Based on these findings, they argue that the DCT offers the performance of PCA but with the significant advantages of a fixed, interpretable, and computationally efficient basis.

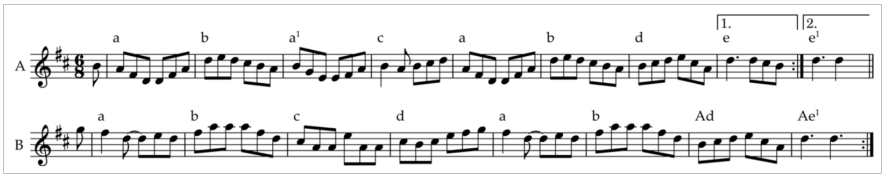


Figure 3.2.13: "Shandon Bells" (O'Neill's No. 1), as annotated by Seán Doherty [99], showing its hierarchical structure with repeated bars labeled alphabetically (e.g., "a", "b") and variations marked with superscripts ("a^1").

Our work in [2] seeks to explore how this representation and its properties apply specifically to the structured patterns of ITM. In ITM, melodies like double jigs exhibit hierarchical repetition of motifs across scales — from 2-note ornaments ($\sim$2 bars) to full 8-bar parts [99]. Figure 3.2.13 illustrates this structure in "Shandon Bells" (O'Neill's No. 1), where Doherty [99] labels bars according to recurring patterns. Such regularity suggests that a small set of basis functions could capture shared contour features across the corpus.

3.2.2 Highlights and Key Findings

Our results show that the DCT is indeed useful for identifying patterns in ITM melodies, but its effectiveness is highly dependent on the length of the contour segment. Figure 3.2.14 shows that for 16-bar segments, 52 DCT coefficients are required to explain 95% of variance, whereas PCA achieves the same threshold with only 22 components. This reduced efficiency suggests that the fixed

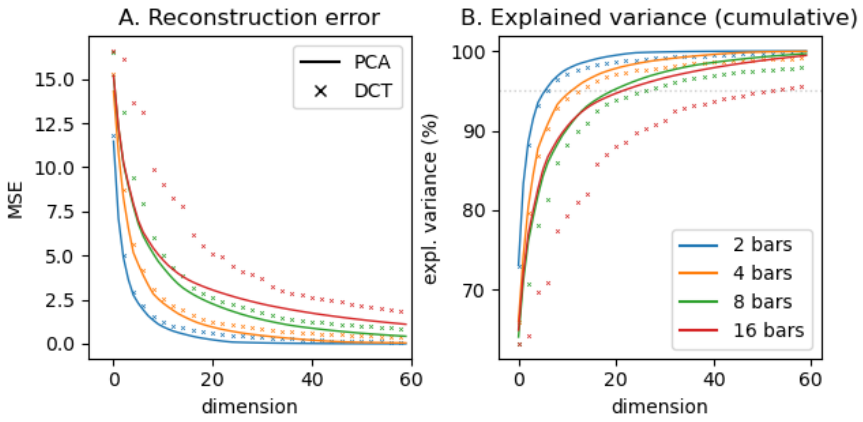


Figure 3.2.14: Comparison of the reconstruction error (Mean Squared Error) and explained variance (cumulative) of PCA and DCT for segments of different lengths from O'Neill's tunes: 2-bar segments (blue), 4-bar segments (orange), 8-bar segments (green), and 16-bar segments (red). The key observation is that while the DCT approximates PCA well for short segments (2-4 bars), its efficiency decreases for longer segments; for 16-bar segments, the DCT requires more than twice as many coefficients as PCA to explain 95% of the variance in the data.

cosine basis struggles to model the longer-range dependencies, asymmetries, and complex structural patterns in ITM. We find that for short segments (2-4 bars), the fixed cosine basis perform comparably to the data-driven PCA. This was not surprising, as the covariance structure of these shorter parts results in a Toeplitz-like matrix [2], for which the DCT is known to be a good approximation for the optimal Karhunen–Loève (KL) basis. However, this alignment between DCT and PCA breaks down as the length of the contour segment increases (8–16 bars).

Nevertheless, cosine contours have some advantages. Their fixed basis functions provide intuitive visualizations and enable interpretable feature extraction (e.g., "archedness" or "undulatingness"). This is illustrated in Figure 3.2.15, which shows melodic contours from German, Chinese, Sioux, and Irish tunes represented in two dimensions. While the overall shape of Sioux tunes is clearly descending (subplot D in Figure 3.2.15), other corpora lack a consistent pattern. Cornelissen et al. [58] propose measuring the concavity or convexity of a contour as its "archedness," but for ITM (subplot E in Figure 3.2.15), there is no typical archedness. By examining the magnitudes of cosine coefficients across all double jigs in O'Neill's corpus, we find that the fourth DCT coefficient (after the mean) is consistently the largest. We define this coefficient as representing the "undulatingness" of a contour, capturing the characteristic rise-and-fall phrasing of Irish double jigs and distinguishing them from other corpora (subplot A.2 in Figure 3.2.15). Overall, this demonstrates that even when DCT is

less efficient for long-range reconstruction, its coefficients remain a powerful tool for interpretable analysis of melodic structure.

3.2.3 Critical Limitations and Scope

The core critique arising from our evaluation is the fundamental trade-off between the fixed cosine basis's interpretability and its computational efficiency for representing longer musical structures. The DCT's strength is its fixed, mathematically defined basis, which allows for intuitive labels like "undulatingness." However, this same fixed nature is its weakness when applied to longer segments of ITM tunes. The data-driven PCA basis, by contrast, adapts to the specific covariance structure of the dataset, achieving far greater efficiency (fewer coefficients needed) for representing 16-bar segments. This indicates that the fixed cosine basis is a sub-optimal, one-size-fits-all solution for this specific musical domain. While it can effectively describe global trends and some mid-level features, it is not the most compact or efficient representation for capturing the full complexity of ITM tunes.

3.2.4 Future work

Our exploration of melodic contour representation suggests moving beyond single, global DCT decompositions toward multi-resolution, time-localized approaches. By applying windowed DCTs at different phrase lengths (4, 8, and 16 bars), we could reveal motif- and phrase-level patterns while addressing the efficiency limitations observed for longer segments. Investigating alternative basis function methods such as wavelets, dictionary learning, or Non-negative Matrix Factorization would allow us to better explore which frameworks best offer interpretability for music analysis with computational efficiency for ITM.

The development of style-specific basis functions could be another promising direction. Rather than relying on fixed mathematical transforms, future work could explore learning adaptive representations directly from symbolic music data through neural network latent spaces or statistical models like hidden Markov models. Such learned representations could capture the structural patterns that characterize ITM more effectively than domain-independent methods.

Finally, Doherty's musicological study on the same corpus [99] could be used as benchmark to compare the melodic structures found in the corpus. We could then evaluate how well different bases capture the hierarchical patterns and phrase structures that define the tradition, as identified by Doherty.

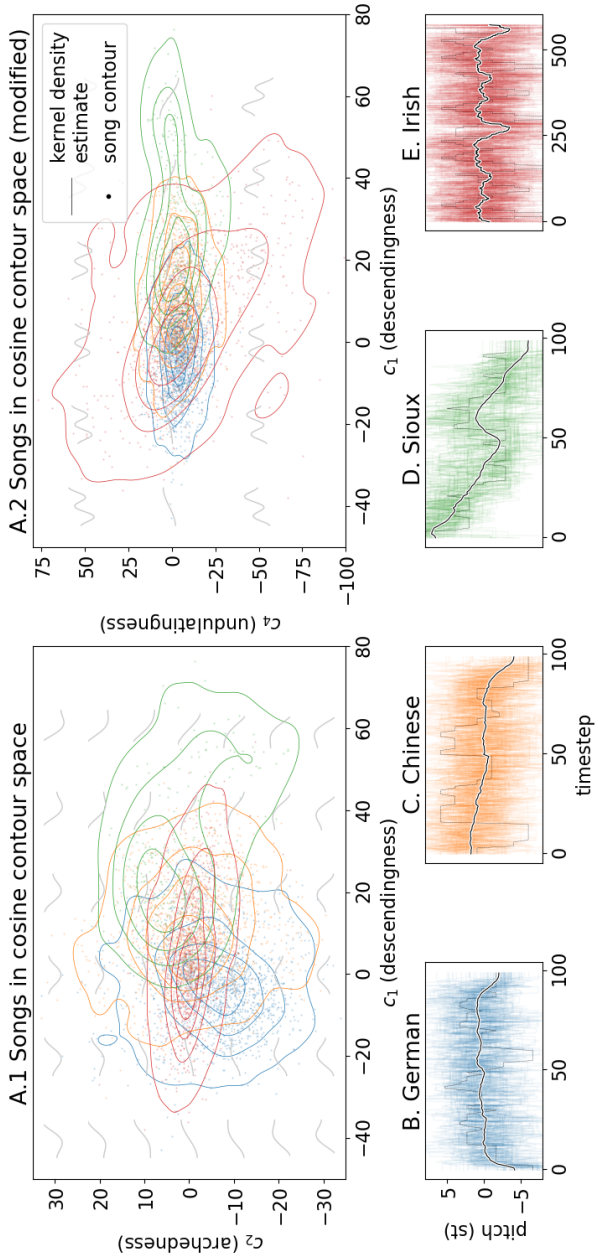


Figure 3.2.15: (A) Different corpora represented in the cosine contour space in terms of "descendingness" and "archedness" and "undulatingness". The average of all melodic contours in a corpus (B-E) show the overall shape of melodic contours in that corpus.

Detection of current AI music generation systems

4 Detection of current AI music generation systems

This chapter discusses the key findings in [3], defining the problems it addresses, critiquing the current limitations, and suggesting future developments. In our article, we focus on detection techniques, using a dataset that combines human-made tracks from the Million Song Dataset [60] with music generated by users of Suno and Udio. We evaluate their robustness across different generative systems, and explore the ethical and practical implications of applying them at scale.

Our work contributes to an emerging body of research addressing AI music detection. Key studies in this area include the SONICS dataset [101], which provides over 97,000 songs from Suno and Udio generated with systematic prompt variations, as well as a detector model trained on such dataset. Other recent work by [48] uses autoencoders to detect compression artifacts in AI-generated audio, while follow-up work [103] links such artifacts to inherent frequency patterns in generative models, providing a theoretical basis for detection. Commercial systems like IRCAM Amplify's detector claim high accuracy (98.5%) [145] but remain largely opaque in their methodology. Collectively, these efforts highlight both the promise of AI music detection and the fundamental challenges of building robust, transparent, and generalizable systems in the face of rapidly advancing generation technologies.

4.1 Defining the Problem

AI-generated music is no longer a novelty. In April 2025, Deezer announced that 18% of the songs uploaded to its platform each day were created entirely by AI [16]. Just a few months earlier, that number was half as high, and as of

September 2025 that number has raised to 28% [17]. AI platforms like Suno and Udio are enabling the mass-scale creation of music without much human effort. This rapid proliferation raises some urgent questions: How can we detect AI music? Should we? And what do we mean by "AI music" in the first place?

Current detection systems face a fundamental tension between surface-level detection — which can be accurate but fragile — and content-based detection, which is more conceptually aligned with music but technically harder to achieve. Many systems rely heavily on artifacts introduced during rendering or encoding [3, 48, 103]. For instance, Suno's tracks exhibit high-frequency artifacts detectable by simple classifiers — likely a technical artifact produced by an upsampling process [3, 103]. Moreover, the landscape of audio artifacts is rapidly evolving as streaming services increasingly adopt neural compression methods like EnCodec and SoundStream, which introduce their own distinctive artifacts that may overlap with those produced by generative AI systems [146].[1] This convergence further undermines the reliability of artifact-based detection pipelines. If AI systems eventually remove such artifacts, these methods will become obsolete. Conversely, if detection systems rely too much on stylistic analysis, they risk misclassifying unconventional or low-fidelity human music, particularly from genres or regions underrepresented in training data [20].

4.2 Key Findings

To date, most academic detection systems for AI-generated music rely either on low- or mid-level audio descriptors (e.g. mel-spectrograms) or learned audio-language embeddings, such as those produced by models like CLAP [61]. While these methods have shown promising results in constrained settings, they often fall short in real-world scenarios involving post-processing or adversarial intent.

Our work contributes to this landscape by evaluating how well different detection approaches perform when applied to music audio generated by popular platforms such as Suno and Udio. In our study, low- and mid-level features like spectral centroid (measuring brightness), MFCCs (capturing timbre), and Beats Per Minute (BPM) (rhythm) reveal subtle differences between AI and human music. Tracks generated by Suno and Udio exhibit measurable differences from human music. In particular, we find that AI-generated music tends to have less high-frequency energy and more uniform rhythmic structures. Many of the AI-generated tracks lack expressive timing variation, and their dynamics are often compressed or flat. These patterns, while subtle, are consistent enough to allow Machine Learning (ML) models to classify them with over 90% accuracy in test conditions. At least for now.

[1] Thanks to Nicolas Jonason for his observation of neural compression methods — it's a great point about how streaming platforms are inadvertently making AI detection even harder!

Models like CLAP create embeddings that map audio to semantic meanings (e.g., "jazz" or "upbeat"). These embeddings can capture global patterns, such as repetitive structures common in AI music. Our article [2] shows that CLAP-based detectors achieve over 90% accuracy on held-out data. However, these systems often rely on technical artifacts, introduced by the AI system's pipeline, rather than musical quality. For instance, we find that halving the sampling rate of the AI-generated audios makes some content detectors flag them as non-AI.

4.2.1 Our Dataset

Like most ML approaches, detecting AI-generated music reliably depends on training on the right data. But building useful datasets is difficult. Some existing collections, like [101], are based on AI music created in controlled environments, which doesn't reflect how people actually use platforms like Suno or Udio.

To address these problems, we built our own dataset using real music from Suno and Udio. Since these platforms don't offer public APIs, we used web scraping techniques between May and October 2024 to collect audio and metadata. While this scraping may go against the platforms' terms of service, our work is protected under EU law (Directive 2019/790) as non-commercial academic research [35].

From Suno, we collected songs from the "New Songs" playlist,[2] which is updated frequently. We ran a script every two hours to download the latest 50 tracks. For Udio, we automated the use of the platform's search tool to find the most popular songs each day, based on play counts. For both platforms, we saved the audio, along with any prompts, genres, and lyrics we could extract. We only kept songs longer than 60 seconds to ensure that we were collecting complete tracks.

As of September 2025, we have collected $281,674$ tracks from Suno and $186,067$ tracks from Udio. However, in our paper [3] we only use $10,000$ of each source. This was done to ensure we had a balanced dataset, as at the time we started working on the article we had a bit less than $20,000$ audio recordings from Udio. To compare with human music, we downloaded $10,000$ audio recordings from the MSD [60] using links to YouTube provided on Last.fm. However, it is important to note that MSD focuses on popular Western music, which may already appear in the training data of Suno and Udio [59, 147]. This could affect how well we can measure originality or detect bias.

To visualize how the sources of music in our dataset differ, we project the CLAP audio embeddings of the recordings into a two-dimensional space using UMAP

[2] `https://suno.com/playlist/cc14084a-2622-4c4b-8258-1f6b4b4f54b3`, last accessed Aug. 13, 2025

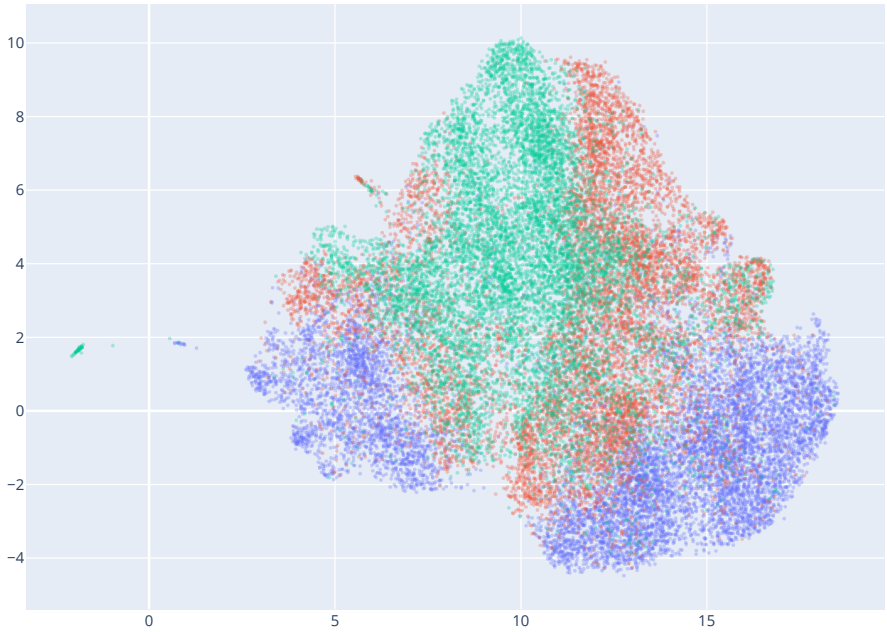


Figure 4.2.1: 2D UMAP projection of CLAP audio embeddings from our dataset, with colour denoting source: Suno (blue), Udio (red), and MSD (green).

[143]. As shown in Figure 4.2.1, the Udio and MSD clusters appear closer together, while Suno samples tend to cluster farther away. We also provide interactive 2D and 3D plots for further exploration.[3]

4.2.2 The detection task

Based on the audio features of our dataset, we find measurable differences between the sources. Suno audio exhibits a lower spectral centroid and pitch salience. This suggests that Suno's audio recordings have reduced high-frequency and harmonic content compared to Udio and MSD. Classifiers can distinguish between the three sources using these features, with Suno being the most identifiable.

For the detection task itself, we implement a hierarchical classifier based on the local classifier per node (LCN) approach [148]. This method first distinguishes between AI and non-AI music at a parent level before further classifying the AI music into its specific source (Suno or Udio) at a child level. We build three versions of detectors, with SVM, Random Forest, and 5-NN at each node within a class tree, as is done in LCN. We train these on CLAP embeddings and evaluate

[3] `https://is-this-ai.github.io/visualizations/umap_plots.html`, last accessed Aug. 13, 2025

their performance using standard metrics, comparing them against the open-source SpecTTTra model [101] and the commercial IRCAM Amplify AI Music Detector.

For AI detection (a binary AI vs. non-AI task), we find that our hierarchical classifiers using CLAP embeddings perform well. The SVM model achieves an F1-score of 0.969 for AI music and 0.935 for non-AI music on our test set. However, we see that performance is highly asymmetric; the commercial IRCAM Amplify detector, while achieving near-perfect scores (F1=0.988 for AI, 0.976 for non-AI), still misclassifies 4.7% of human-made music as AI — a significant error rate with serious ethical implications if scaled. The open-source SpecTTTra model performs poorly, especially on Udio music, which it misclassifies as human 75.3% of the time.

Crucially, we find that all detectors fail to generalize. When tested on music from an out-of-sample platform, Boomy, detection rates are very low (IRCAM: 6%, our best model: 24%). Cross-platform tests show that models trained on Udio generalize well to Suno, but models trained on Suno perform poorly on Udio, indicating that detection relies on platform-specific artifacts rather than a general signature of "AI-ness."

Finally, we find AI detection is highly sensitive to simple audio processing. Our experiments show that simply downsampling AI-generated audio from 48 kHz to 22.05 kHz can cause detectors to misclassify it as human-made. This fragility suggests that models may be exploiting spurious correlations or confounds, a phenomenon well-documented in shortcut learning literature [149, 150]. We provide audio examples illustrating the impact of such transformations, including filtering and resampling, on both baseline detectors as well as the ones we trained with our collected dataset.[4]

4.3 Critical Limitations

Training reliable detection systems requires diverse, balanced datasets. However, the most common human music datasets — such as the MSD — are dominated by Western, commercially produced music. These datasets are not only unrepresentative of global musical styles but may also be present in the training corpora of generative models, introducing subtle circularity. When a detector identifies music as "AI" because it does not match the distribution of MSD, it risks flagging human music from underrepresented genres as AI. Conversely, AI-generated tracks trained on MSD-like corpora may appear more "human," reducing detection accuracy. This bias mirrors concerns from other domains of generative detection, such as AI images misclassifying minority demographics due to underrepresentation in training data [151]. In music, such biases are

[4] `https://is-this-ai.github.io/visualizations/audio_analysis_results.html`, last accessed Aug. 13, 2025

harder to quantify because the criteria for "real" music are themselves culturally and historically contingent [20, 152].

About a month after we submitted our article to TISMIR, IRCAM Amplify announced that their detector can now detect Boomy.[5] A month after that they announced that their detector can now detect Riffusion.[6] And a few days after, they announced it could detect Sonauto.[7] Moreover, platforms like Suno and Udio release new, improved versions frequently — Suno is currently at version v4.5[8] and Udio is at version Allegro v1.5.[9] This rapid-fire updating is the clearest sign of an escalating technological conflict: an "arms race" between generative AI systems and the detectors built to catch them.

This arms race is fundamentally asymmetric. Generative model developers, driven by market competition and user demand, are primarily incentivized to improve the quality, diversity, and fidelity of their output. Removing the very technical artifacts that current detectors rely upon is a natural consequence of this quality improvement. A new version of Suno could remove the upsampling artifact identified by [3, 48], and an entire class of detectors trained on that artifact would instantly become obsolete. Detectors are always playing catch-up. They can only be trained on the outputs of past versions of these systems. Each new model update from Suno, Udio, or a new competitor like Sonauto, forces detection efforts to revise or start from scratch: new data must be collected and models retrained. This is a slow, reactive, and costly process.

This creates a fragile detection ecosystem. As our tests show [3], even simple post-processing edits — like downsampling an audio recording — can easily fool a detector that is looking for specific, dataset-dependent artifacts. A bad actor needs only a basic digital audio workstation (DAW) to strip away the superficial identifiers that detection systems depend on, mirroring evasion techniques seen in voice spoofing detectors [153] and video deepfake detection [154]. This is especially worrying because humans are not good at detecting AI music either. A study by [155] finds listeners only identify AI music 51% of the time — about as good as guessing.

Efforts to detect AI music confront an immediate ambiguity: what exactly are we trying to detect? A track might be generated entirely by an AI model, or it might involve human composition alongside AI mastering or arrangement. Hu-

[5] `https://www.ircamamplify.io/blog/ai-music-detector-now-detects-boomy`, last accessed Aug. 25, 2025

[6] `https://www.ircamamplify.io/blog/ai-music-detector-now-detects-riffusion`, last accessed Aug. 25, 2025

[7] `https://www.ircamamplify.io/blog/ai-music-detector-now-detects-sonauto`, last accessed Aug. 25, 2025

[8] `https://suno.com/blog/introducing-v4-5`, launched May 1, 2025, last accessed Sept. 2, 2025

[9] `https://help.udio.com/en/articles/10748731-changelog-what-s-new-with-udio`, launched March 18, 2025, last accessed Sept. 2, 2025

man music often includes AI tools (e.g., auto-tune), while AI music can sometimes feel very human. Overreliance on detection could unfairly suppress human artists, especially those in styles underrepresented in training data [20]. Finally, some artists use AI to generate stems, lyrics, or chord progressions, which are later refined or reinterpreted. Others prompt systems like Suno or Udio to create fully formed tracks with minimal input. Given this spectrum, drawing a strict line between "AI" and "human" music is inherently reductive and thus problematic.

4.4 Future Work

The commercial pressure to market AI detectors as a solved problem — illustrated by high but context-dependent accuracy claims — obscures these fundamental limitations and creates a false sense of security. This is made worse by a lack of transparency. Proprietary detectors like IRCAM Amplify hinder independent verification of them. Open benchmarks and shared datasets (e.g., the FMA corpus [156]) could accelerate progress while ensuring fairness.

In essence, this reactive cycle is a losing battle. The relentless pace of generative AI development suggests that a sustainable solution will require moving beyond technical detection alone, towards systems with built-in transparency, such as robust watermarking or provenance standards, infused into generated music audio from the start. The field must prioritize interpretability (to explain why a track is flagged) and fairness (to avoid penalizing human artists). Only through collaboration between technologists, ethicists, and musicians can detection serve its purpose: to inform, not control, the future of creativity.

Explanation as Communication

5 Explanation as Communication

This chapter highlights our findings in [4], defining the problems it addresses and the framework it proposes, including some experimental examples that, while informative, did not make it into the final published version. We also provide a critique of our work, as well as future work.

5.1 Problem description

Within the domain of AI there is a growing need for understanding how AI systems are operating and making particular decisions. The EU General Data Protection Regulation (put into effect in 2018) [117] introduces the right of EU citizens to receive an explanation for decisions made by such algorithmic systems. The discipline of XAI aims to address this demand by engineering a variety of methods for interrogating such systems and their components. Specifically, XAI aims to make AI systems provide justifications for their output or procedures in a way that is intelligible to humans [157]. Arrieta et al. [118] and Langer et al. [124] are two recent surveys of XAI, and give an overview of the state of the art in the domain and the many challenges that remain.

Much work in XAI can overlook key aspects, such as who the explanation is crafted for (target audience), and the goal of explanation [128]. What is ultimately desired in an explanation of an AI system is a clear and interpretable link between its input and output with reference to the environment in which the system is applied or even built. What constitutes a "good" explanation in XAI can seem defined more by the intuition of a researcher rather than considerations relevant to evaluating an explanation crafted by a human [55, 56]. A human might select a subset of the causes for an event as an explanation, but this can be influenced by their own cognitive biases. Doshi-Velez and Kim [120] appeal to the social nature of explanations to categorize evaluation approaches

for interpretability, taking into account the target audience as well as the goal of explanation and how that changes the evaluation of the explanation.

Furthermore, there is a need to consider the social dimensions in which these systems and such explanations exist, since they are interpreted by people. Miller et al. [55, 56] propose looking at XAI as a whole from the perspective of social science. They show how, for humans, explanations are contrastive, selected, and social/conversational. Moreover, Been [158] argues that finding a language to communicate effectively with AI is crucial to bridge the gap between human understanding and AI capabilities. Dialogue and collaboration are key to constructing machines aligned with our goals.

While the process of XAI is essentially one of communication – the act of explaining something is the exchange of knowledge between two people as a social interaction or conversation – Miller et al. [55, 56] do not explicitly consider the process of communication. Thus in [4] we ask: How might an explicit consideration of communication models in XAI formalize and extend the conception of the communication process in XAI, and thus motivate ways of identifying and overcoming miscommunication?

5.2 Highlights of our contributions

In [4] we propose that the methods in XAI are fundamentally an act of communication — a social process where a sender seeks to transfer knowledge or explain its behaviour to a receiver to achieve a shared understanding. We build from a taxonomy that combines concepts defined in detailed surveys of XAI [118–122], which helps us identify the components of the XAI pipeline and how they can be interpreted in the communication process. This motivates a particular model of communication showing the importance of establishing a set of shared knowledge, beliefs, and assumptions of the participants in a conversation, which is called *common ground* [52]. When common ground is absent, explanations may appear convincing while misrepresenting model reasoning. Saliency maps, for instance, may highlight spurious correlations [127], and users may accept misleading explanations aligned with their biases (e.g., the "Kiki-Bouba classifier" misattributing musical structure to irrelevant statistics [159]).

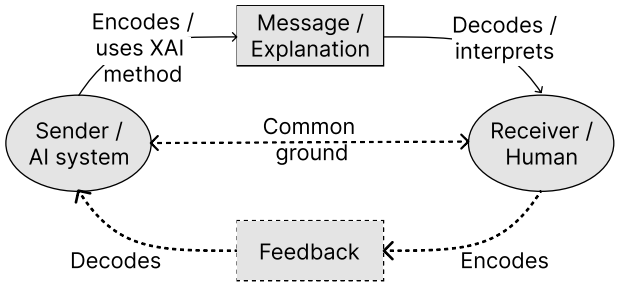


Figure 5.2.1: Our simplified cyclic communication model with the addition of common ground [52] between sender and receiver. The sender encodes a message and sends it to the receiver, who then decodes it and interprets it. The receiver then provides feedback on their processing of the message in order to update the common ground.

Thus, we cast XAI in the framework of a communication model. We propose a simplified cyclical model, illustrated in Figure 5.2.1, which explicitly includes feedback and common ground. The steps of our simplified cyclic communication process between an AI system (sender) and a human (receiver) are the following:

1. The AI system encodes an explanation of its behaviour through XAI methods: local approximations (e.g., LIME [125]), visualizations (e.g., saliency maps [160]), or others detailed in Section 2.4.
2. The receiver interprets (decodes) that explanation.
3. The receiver encodes their version of the explanation (based on their interpretation) and then checks if the AI system's behaviour fits that alternative description, using positive evidence of understanding as feedback.

Building on the definition of XAI as a human-agent interaction problem by Miller [56], we use this model to describe the XAI process. Through this, we find that what is largely missing from XAI research is the need to formally consider miscommunication. This motivates developing a mechanism for *feedback* in order to detect miscommunication and to establish a common ground.

We now demonstrate our framework using several practical examples that highlight the encoding, decoding, and feedback process.

Example I: Kiki-Bouba challenge

As a way of demonstration of the process we consider first the "Kiki-Bouba Challenge" (KBC). Sturm et al. [159] make use of algorithmic music composition for generating music audio belonging to either one of two categories, named Kiki

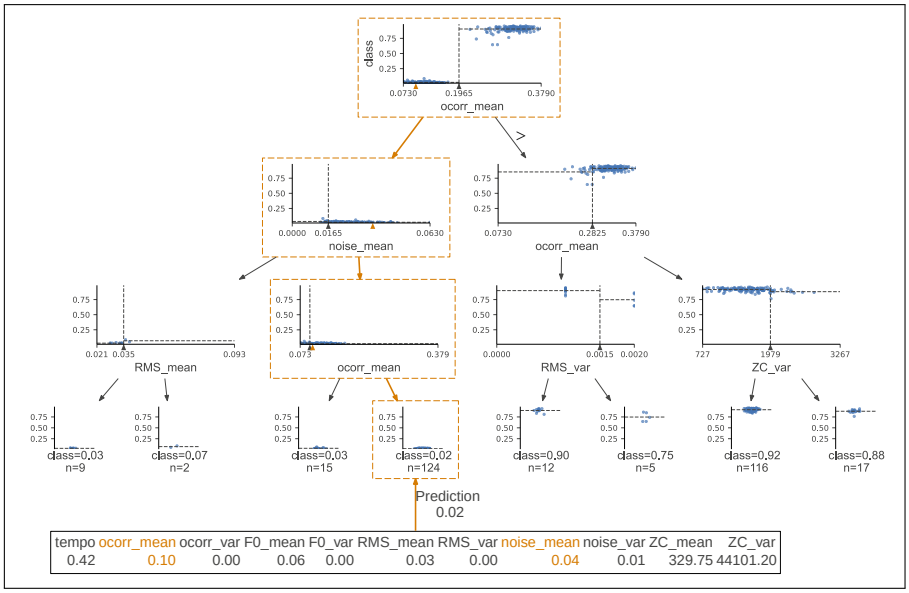


Figure 5.2.2: Explanation of an instance being predicted as Kiki. The most important feature here is the onset autocorrelation mean across the frames of the instance (`ocorr_mean`). The second most important feature is the spectral flatness mean (`noise_mean`).

and Bouba. These two categories are well-differentiated with respect to some high- and low-level musical properties.

Our explainability goal for this classifier is confidence. Our initial hypothesis is that the model uses high-level (content) criteria for classification. An explanation that reinforces this hypothesis would boost our confidence in the model, while one showing reliance on low-level features would decrease it. The target audience is the model's developers, so the explanation must be technically interpretable. We incorporate a feedback loop to assess both the quality of the encoding "does the explanation match the model's behaviour?" and the decoding "was our interpretation correct?". We use the KBC's "unacceptable solution" [159] as our black-box model. This model was deemed unacceptable precisely because it relies on low-level, musically irrelevant features, despite achieving a perfect classification score.

Taking into account the target audience (developers) and the explainability goal (confidence) we devise a way of encoding an explanation. We build a new simplified model, a linear regression tree, and optimize it to replicate the black-box model's predictions. We compute a set of low-level (e.g., onset autocorrelation, spectral flatness) and mid-level (e.g., tempo) audio features from the music samples. The tree is optimized to approximate the black-box model's continuous output, which ranges from 0.00 (Kiki) to 1.00 (Bouba). We then visualize

the tree and interpret it as a way of decoding the message. In Figure 5.2.2 we can see how an explanation of an instance being predicted as Kiki looks like. The decoding of the message is that the features at the highest nodes are the most discriminative. For an instance predicted as Kiki, the explanation indicates that the mean onset autocorrelation and mean spectral flatness are the most important features influencing the decision.

We test the encoding by verifying that the simple model's accuracy matches the black-box model's (which it does). We then test the decoding by synthesizing new audio samples that manipulate the top features identified by the explanation (e.g., altering the onset autocorrelation via time-stretching). If our interpretation is correct, these manipulations should change the black-box model's predictions. However, we find that while the simple model's predictions change, the black-box model's output remains unchanged, it still classifies the time-stretched audio as Kiki. This feedback provides decisive evidence that the explanation, while technically accurate in the way it was encoded, does not correctly represent the black-box model's true decision-making behaviour.

The experiment reveals several critical shortcomings in the explanation process. The explanation derived from the simplified model incorrectly identifies mid-level features as primary decision factors, while the black-box model's true decision logic relies heavily on low-level features such as zero crossings. The feedback loop provides decisive evidence that naively trusting an explanation's surface-level features can lead to incorrect conclusions about model behaviour.

Example II: Tempo estimation using LIME (naive approach)

We now "blindly" apply LIME [49] to `librosa`'s [161] tempo estimation algorithm, treating it as a black-box model. To encode a message, we use the explanation proposed in SLIME [162], a version of LIME used for audio signals. We apply a STFT to an audio signal and then treat this time-frequency representation as an image to partition it into regions to be perturbed, otherwise called super-pixels [49]. This partition is done using k-means clustering segmentation (naive approach). To perturb the signals, we either keep the super-pixels the same (activated) or we replace them by the minimum value of the original STFT (deactivated). We use a least squares linear regression model as the interpretable model and take the top-4 coefficients of the linear regression model to generate the explanation, meaning that the four super-pixels associated to those coefficients will be active while the rest will be attenuated. In Figure 5.2.3 we can see the STFT magnitude plot of the top-4 coefficients reconstructed signal, with its detected onsets. The original audio signal for this experiment is a clip of the well-known "Dance of the Sugar Plum Fairy" from the ballet "The

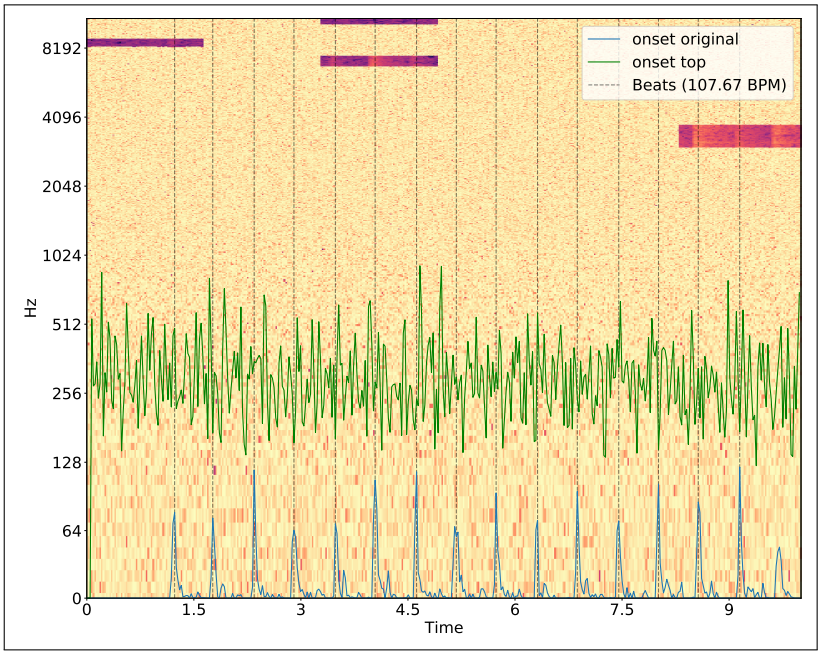


Figure 5.2.3: Comparison of the detected onsets of the original signal and the top signal with the top 4 super-pixels active (and the rest attenuated). The original signal had a tempo of 107.67 BPM, while the top signal has a tempo of 123.05 BPM.

Nutcracker" by Tchaikovsky, provided by `librosa`.[1] The tempo is estimated using the same library, by first calculating the signal's onset strength and later estimating the tempo in BPM. The interpretation of this explanation would be that in order to estimate a tempo of 107.97 BPM, the model looks mostly at frequencies above $2048Hz$.

Looking at the explanation it is clear that the encoding was not adequate. The resulting explanation does not highlight any interpretable aspects of the spectrogram that gives us insight about anything related to tempo. The estimated tempo of the synthetic signal with only the most important super-pixels active is 123.05 BPM, proving that the model does not agree with the explanation. Consequently, this encoding error made an accurate decoding by the human receiver impossible.

This experiment shows how a well-studied algorithm such as SLIME [162] cannot be used naively. There are many aspects of it that have to be taken into account: the type of the proxy model used (linear models, decision trees, or falling rule lists), the number of segments (or super-pixels) to use, image segmentation technique (how to shape the super-pixels), the number of perturbations to generate, the kernel width (to measure distances between perturbations and original), number of top features to use for each explanation. It is also difficult to extract general explanations of the model's behaviour from local explanations. These challenges have also been identified in later work by Mishra et al. [163].

Example III: Health Diagnostics

This example illustrates XAI as a communication process in the context of AI-driven treatment recommendations for diabetic patients. While AI holds promise in healthcare, especially in diagnostics and decision support, recent studies highlight significant challenges. Ullah et al. [164] point to issues such as model opacity and limited clinical understanding, often due to a lack of shared context. Goh et al. [165] emphasize that diagnostic performance is sensitive to how clinicians interact with and interpret AI prompts, reinforcing the need for clear communication and trust-building mechanisms.

Consider an AI system recommending a treatment plan based on factors like blood glucose levels, BMI, and patient history. It reports that elevated blood glucose is the primary reason for increasing insulin dosage. We map this interaction onto our communication model (Figure 5.2.1):

1. Encoding: The AI system explains, e.g., "Due to elevated blood glucose levels, insulin dosage is increased by X units," possibly supported by feature relevance indicators.

[1] `https://librosa.org/data/audio/Kevin_MacLeod_-_P_I_Tchaikovsky_Dance_of_the_Sugar_Plum_Fairy.ogg`, last accessed Aug. 13, 2025

2. Decoding: The clinician interprets this as a causal link — high glucose leads to a dosage increase — and expects changes in glucose to affect the recommendation.
3. Feedback: The clinician tests this by modifying glucose values in the system's input. If the AI responds as expected, the explanation is validated; if not, it indicates miscommunication or a flawed explanation.

Feedback mechanisms like this can help clinicians verify their understanding of AI outputs, directly addressing concerns raised by Ullah et al. [164] and Goh et al. [165], and supporting clearer, more trustworthy AI-human collaboration.

5.3 Critique of our contribution

Our work in [4] frames XAI through the lens of established communication models, emphasizing the cyclical nature of explanation and the critical role of common ground and feedback. This perspective helps to formalize and extend the conception of the communication process in XAI, moving beyond one-way explanation and emphasizing the undesired outcome of miscommunication. However, applying human communication models to human-agent interaction presents inherent challenges. In the context of XAI, the sender (model explaining its behaviour) has no intention of being understood. Therefore, the interpretation of the explanation is highly influenced by the receiver's expectations on what the explanation should contain, as well as their knowledge on the model's mechanics. Furthermore, the ways a machine "conceptualises" a problem fundamentally differs to how humans do [158]. Another aspect of AI systems is that they are designed and optimized to solve a specific task, which mostly does not involve explainability. Therefore the system lacks intention towards being understood, which can result in the receiver projecting their own intention into the conversation, influenced by their own cognitive biases.

Our practical examples, while illustrative, also highlight the limitations and difficulties in implementing this framework. As shown in the Kiki-Bouba and LIME examples, XAI techniques can produce explanations that are misaligned with the model's true behaviour, and designing an effective feedback loop is non-trivial. The health diagnostics example points towards a promising direction but remains a conceptual model that requires empirical validation in real-world settings.

A problem with the aforementioned models of communication is the limited analysis of miscommunication they provide [166]. Shannon and Weaver's model [167], being one of the most referenced, only captures the communication problem of mishearing due to factors like background noise. However, this does not consider communication problems arising from differences in interpretation during conversation. Our model attempts to address this through feedback,

but fully capturing the nuances of miscommunication in XAI remains a complex challenge.

5.4 Future work

If we want to provide explanations on how an AI system works to people, we need to understand how people will unpack the explanation before attempting to craft one. There are a lot of resources that we can employ from research in social sciences on how explaining something works between humans, as done by Miller [56]. One evident problem with current XAI approaches is how they constrict what constitutes a "good" explanation to the researchers' intuition, without taking into account who is demanding an explanation and whether the explanation was misunderstood.

By looking at communication models we identify that when humans want to communicate they rely on grounding, constantly providing positive (or negative) evidence of understanding. This suggests XAI frameworks should implement iterative feedback mechanisms where explanations can be validated through controlled experiments, similar to the feedback loops demonstrated in the practical examples. Without this, explanations risk becoming misinterpreted by the receiver of the explanation. Future work should focus on developing and rigorously testing these feedback mechanisms. Researchers should focus on formalizing methods for users to provide positive evidence of understanding [168] within XAI interfaces. Another key area is exploring how to build and update a common ground between the AI system and the human user over the course of an interaction. Furthermore, it is necessary to conduct user studies to evaluate whether a cyclical communication model leads to more accurate user understanding of AI systems compared to one-way explanation methods. Another important direction involves investigating how to tailor the communication process and the form of feedback to different target audiences (e.g., developers vs. end-users) and different explainability goals (e.g., trust, causality, confidence). Ultimately, addressing the challenge of miscommunication is essential for developing XAI methods that are truly effective, trustworthy, and aligned with human understanding.

Concluding Remarks

6 Concluding Remarks

The rapid proliferation of AI-generated music — from Suno and Udio enabling millions of users to create songs with a single prompt, to Deezer reporting that 28% of daily uploads are now AI-generated [17] — has arguably begun to transform the creative landscape. Yet as this technology scales, it exposes profound tensions: between accessibility and deskilling, between innovation and exploitation, between technical capability and ethical concerns. By focusing on the interconnected problems of evaluation, detection, and explanation, this thesis provides a multi-faceted foundation for critically engaging with music created by AI.

The first research question of this thesis (Chapter 3) examines how to compare music collections with item-level traceability. A distance-based statistical framework shows that domain-independent methods can distinguish collections, identify outliers, and highlight typical items. Yet, cases like "The Jolly Joker" and others reveal that statistical similarity does not equal musical similarity, exposing the gap between computational metrics and musicological understanding. The second question (Chapter 3) explores melodic representation with basis functions, focusing on the Discrete Cosine Transform. While cosine contours are an efficient representation for short melodic segments and features like "undulatingness" in Irish traditional music, their effectiveness declines with longer excerpts. This highlights unresolved trade-offs between efficiency, interpretability, and fidelity. The third question (Chapter 4) addresses detecting AI-generated music across platforms. Detection models perform well on the surface but fail with cross-platform tests or simple audio transformations. These weaknesses are further exacerbated by an arms race dynamic, where detection hinges on technical artifacts rather than musical content, making current methods fragile and unsustainable. The fourth question (Chapter 5) reframes explainability in AI as a communication problem. Current XAI meth-

ods often mislead without shared context. A communication model emphasizing feedback loops and common ground offers a framework for more effective, situated explanation systems.

The development of the work in this thesis, particularly the legal and ethical challenges encountered in building a dataset for AI music detection research, underscores a broader, more urgent crisis in AI research. We stand at a precipice where the most popular AI systems are developed by a handful of private companies behind closed doors. The "terms of service" of such platforms conflict with the objectives of academic research, their rapid version updates make detectors obsolete, and their narratives of "democratization" often obscure the realities of deskilling, cultural homogenization, and the exploitation of training data.

Academic research cannot compete with large technology companies on scale, data, or computational resources. Nor should this be its primary objective. The unique value and imperative of academia lie elsewhere: in serving as a critical counterweight. The role of public-interest scholarship is not necessarily to build the most powerful model, but to develop the necessary checks and balances. This entails moving beyond mere performance metrics (e.g., higher accuracy, better fidelity) towards rich, multi-faceted evaluation that assesses cultural bias, stylistic diversity, economic impact, and ecological validity. It calls for the development of "artificial critics" capable of meaningful musicological analysis. Moreover, the struggle to collect and study data from commercial platforms is not a peripheral issue; it is a central research challenge. There is a pressing need to defend legal rights for text and data mining for academic research within the EU and fair use elsewhere. Ceding control of research access to corporate gatekeepers would be detrimental to open science and public-interest scholarship. Therefore, technical research must be coupled with critical analysis. Music informatics and AI research would benefit from deeper engagement with ethics, law, media studies, and musicology to fully understand the societal impact of the systems we build and study [18].

The rise of AI-generated music is not merely a technical shift; it is a cultural and economic event. This thesis shows how research practices can adapt accordingly. There is a fundamental responsibility for academia to foster research that asks the uncomfortable questions, probes the failures, documents the biases, and builds frameworks for accountability. While academic institutions may not be able to out-train the large models built by big tech companies, they are uniquely positioned to critically assess them—to help ensure that the future of music, and of AI, remains diverse, equitable, and profoundly human. This is not a niche concern; it is a central challenge for scholarship in the 21st century.

Bibliography

[1] Laura Cros Vila and Bob L. T. Sturm. "Statistical evaluation of abc-formatted music at the levels of items and corpora". In: *Proceedings of the International Conference on AI and Musical Creativity (AIMC)*. 2023. URL: https://aimc2023.pubpub.org/pub/tv38rj87/release/1.

[2] Laura Cros Vila and Bob L. T. Sturm. "Cosine Contours: A Case Study with Melodies from Irish Traditional Dance Music". In: *Late Breaking Demo in the 24th International Society for Music Information Retrieval Conference (ISMIR)*. 2023. URL: https://ismir2023program.ismir.net/lbd_318.html.

[3] Laura Cros Vila, Bob L. T. Sturm, Luca Casini, et al. "The AI Music Arms Race: On the Detection of AI-Generated Music". In: *Transactions of the International Society for Music Information Retrieval* 8.1 (2025), pp. 179–194. DOI: 10.5334/tismir.254.

[4] Laura Cros Vila and Bob L. T. Sturm. "(Mis)Communicating with our AI Systems". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI. Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713771.

[5] Bob L. T. Sturm, Ken Déguernel, Rujing Stacy Huang, et al. "AI Music Studies: Preparing for the Coming Flood". In: *Proceedings of the International Conference on AI and Musical Creativity (AIMC)*. 2024. URL: https://aimc2024.pubpub.org/pub/ej9b5mv1/release/2.

[6] Bob L. T. Sturm, Marco Amerotti, David Dalmazzo, et al. "Stochastic Pirate Radio (KSPR): Generative AI applied to simulate commercial radio". In: *Proceedings of the International Conference on AI and Musical Creativity (AIMC)*. 2024. URL: https://aimc2024.pubpub.org/pub/rppff8mc/release/2.

[7] OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. https://openai.com/blog/chatgpt (last accessed Aug. 6, 2025). 2023.

[8] UBS. *ChatGPT sets record for fastest-growing consumer app in history*. https://www.ubs.com/global/en/wealth-management/insights/chief-investment-office/house-view/daily/2023/latest-25052023.html (last accessed Aug. 18, 2025). 2023.

[9] Yujie Sun, Dongfang Sheng, Zihan Zhou, et al. "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content". In: *Humanities and Social Sciences Communications* 11.1 (2024), p. 1278. DOI: 10.1057/s41599-024-03811-x.

[10] Hussam Alkaissi and Samy I. McFarlane. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing". In: *Cureus* 15.2 (2023), e35179. DOI: 10.7759/cureus.35179.

[11] M. Sallam. "ChatGPT Utility in Healthcare Education, Research, and Practice". In: *Healthcare* 11.6 (2023), p. 887. DOI: 10.3390/healthcare11060887.

[12] Jerome Goddard. "Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers". In: *The American Journal of Medicine* 136.11 (2023), pp. 1059–1060. DOI: 10.1016/j.amjmed.2023.06.012.

[13] Fereshteh Hasanzadeh, Colin B. Josephson, Gabriella Waters, et al. "Bias recognition and mitigation strategies in artificial intelligence healthcare applications". In: *npj Digital Medicine* 8.1 (2025), p. 154. DOI: 10.1038/s41746-025-01503-7.

[14] Eleanor Drage and Kerry Mackereth. "Does AI Debias Recruitment? Race, Gender, and AI's "Eradication of Difference"". In: *Philosophy & Technology* 35.4 (2022), p. 89. DOI: 10.1007/s13347-022-00543-1.

[15] Will Simpson. *The first ever AI-generated track to hit the charts has landed in Germany, and people aren't happy*. https://www.musicradar.com/news/first-ever-ai-chart-germany (last accessed Aug. 18, 2025). Aug. 2024.

[16] Kristin Robinson. *Deezer Says Number of Fully AI-Generated Songs Delivered Daily Has Doubled Since January*. https://www.billboard.com/pro/deezer-fully-ai-generated-songs-uploaded-daily/ (last accessed Aug. 18, 2025). Apr. 2025.

[17] Deezer. *Deezer: 28% of all music delivered to streaming is now fully AI-generated*. https://newsroom-deezer.com/2025/09/28-fully-ai-generated-music/ (last accessed Sep. 22, 2025). Sept. 2025.

[18] Bob L. T. Sturm, Ken Déguernel, Rujing Huang, et al. *MusAIcology: AI Music and the Need for a New Kind of Music Studies*. Preprint, OSF. 2024. DOI: 10.31235/osf.io/9pz4x.

[19] Eric Drott. "Copyright, compensation, and commons in the music AI industry". In: *Creative Industries Journal* 14.2 (2021), pp. 190–207.

[20] Andre Holzapfel, Bob L. T. Sturm, and Mark Coeckelbergh. "Ethical Dimensions of Music Information Retrieval Technology". In: *Transactions of the International Society for Music Information Retrieval* (2018). DOI: 10.5334/tismir.13.

[21] Fabio 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). DOI: 10.5334/tismir.86.

[22] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. "Homogenization Effects of Large Language Models on Human Creative Ideation". In: *Pro-*

ceedings of the 16th Conference on Creativity & Cognition. Association for Computing Machinery, 2024, pp. 413–425. ISBN: 9798400704857. DOI: 10.1145/3635636.3656204.

[23] S. Tan. "Are We All Musicians Now? Authenticity, Musicianship, and AI Music Generator Suno". In: *SocArXiv* (2024). DOI: 10.31235/osf.io/4nt8z.

[24] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. "DeepBach: a Steerable Model for Bach Chorales Generation". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 1362–1371.

[25] Atharva Mehta, Shivam Chauhan, and Monojit Choudhury. "Missing Melodies: AI Music Generation and its "Nearly" Complete Omission of the Global South". In: *arXiv preprint arXiv:2412.04100* (2024).

[26] Atharva Mehta, Shivam Chauhan, Amirbek Djanibekov, et al. "Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models". In: *arXiv preprint arXiv:2502.07328* (2025).

[27] C. Willman. *Music Streaming Hits Major Milestone as 100,000 Songs are Uploaded Daily to Spotify and Other DSPs*. https://variety.com/2022/music/news/new-songs-100000-being-released-every-day-dsps-1235395788/ (last accessed Aug. 18, 2025). 2022.

[28] Julia Barnett. "The ethical implications of generative audio models: A systematic literature review". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 146–161.

[29] Elizabeth Wilson, Anna Wszeborowska, and Nick Bryan-Kinns. "A Short Review of Responsible AI Music Generation". In: *Proceedings of the 6th Conference on AI Music Creativity (AIMC 2025)*. Brussels, Belgium, Sept. 2025, pp. 1–10.

[30] Fabio Morreale, Megha Sharma, and I-Chieh Wei. "Data Collection in Music Generation Training Sets: A Critical Analysis". In: *Proceedings of the 24th International Society for Music Information Retrieval Conference* (Milan, Italy). ISMIR, 2023, pp. 37–46. DOI: 10.5281/zenodo.10265217.

[31] Andrea Agostinelli, Timo I Denk, Zalán Borsos, et al. "Musiclm: Generating music from text". In: *arXiv preprint arXiv:2301.11325* (2023).

[32] Jade Copet, Felix Kreuk, Itai Gat, et al. "Simple and Controllable Music Generation". In: *Proceedings of the Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 47704–47720.

[33] Kyle Wiggers. *Google created an AI that can generate music from text descriptions, but won't release it*. https://techcrunch.com/2023/01/27/google-created-an-ai-that-can-generate-music-from-text-descriptions-but-wont-release-it/ (last accessed Aug. 21, 2025). 2023.

[34] US District Court for the Northern District of California. "X Corp. v. Bright Data Ltd. 3:23-cv-03698". More information available at https://www.courtlistener.com/docket/67637345/x-corp-v-bright-data-ltd/ (last accessed Aug. 18, 2025). 2024.

[35] European Union. *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital*

Single Market. Official Journal of the European Union, L 130, 17 May 2019, pp. 92–125. 2019. URL: http://data.europa.eu/eli/dir/2019/790/oj.

[36] United States Congress. *17 U.S. Code § 107 - Limitations on exclusive rights: Fair use*. https://www.law.cornell.edu/uscode/text/17/107 (last accessed Aug. 18, 2025). 1976.

[37] Future of Life Institute. *Pause Giant AI Experiments: An Open Letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed Aug. 19, 2025). 2023.

[38] Mark Gottlieb. *The Books3 Dataset: Controversy Surrounds Unauthorized Use in AI Training*. https://literaryagentmarkgottlieb.com/blog/the-books3-dataset-controversy-surrounds-unauthorized-use-in-ai-training (last accessed Aug. 18, 2025). 2023.

[39] Rizwan Choudhury. *Anti-piracy group shuts down Books3, a popular dataset for AI models*. https://interestingengineering.com/innovation/anti-piracy-group-shuts-down-books3-a-popular-dataset-for-ai-models (last accessed Aug. 18, 2025). 2023.

[40] Chat GPT Is Eating the World. *Infographic on which AI companies face most copyright lawsuits*. last accessed Aug. 18, 2025. Mar. 2025. URL: https://chatgptiseatingtheworld.com/2025/03/30/infographic-on-which-ai-companies-face-most-copyright-lawsuits/.

[41] VentureBeat. *Suno raises $125M to become the ChatGPT of music creation*. https://venturebeat.com/ai/suno-raises-125m-to-become-the-chatgpt-of-music-creation/ (last accessed Aug. 18, 2025). 2024.

[42] James Steinhoff. "The Social Reconfiguration of Artificial Intelligence: Utility and Feasibility". In: University of Westminster Press, 2021, pp. 123–143. ISBN: 9781914386169. DOI: 10.16997/book55.h.

[43] John Roberts. "Art After Deskilling". In: *Historical Materialism* 18.2 (2010), pp. 77–96. DOI: 10.1163/156920610X512444.

[44] Mehmet Uçar, Hüseyin Çapuk, and Muhammet Faruk Yiğit. "The relationship between artificial intelligence anxiety and unemployment anxiety among university students". In: *WORK* 80.2 (2024), pp. 701–710. DOI: 10.1177/10519815241290648. URL: https://doi.org/10.1177/10519815241290648.

[45] 20VC with Harry Stebbings. *Mikey Shulman, CEO @Suno: The Future of Music, What is Gonna Happen? | E1244*. https://www.youtube.com/watch?v=E0YL83U5VWk (accessed Aug. 19, 2025), timestamp: 23:30. 2025.

[46] Liz Pelly. *Mood Machine: The Rise of Spotify and the Costs of the Perfect Playlist*. Hodder & Stoughton, 2025.

[47] Goldmedia. *AI and Music: Market Development and Impact on Music Authors and Creators in Germany and France*. Tech. rep. GEMA, SACEM and Gold Media, 2024.

[48] Darius Afchar, Gabriel Meseguer-Brocal, and Romain Hennequin. "AI-Generated Music Detection and its Challenges". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process-*

ing (ICASSP). 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10890655.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you? Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[50] Stewart L. Tubbs. *Human communication : principles and contexts*. New York: McGraw-Hill New York, 2012. ISBN: 9780078036781.

[51] Denis McQuail and Sven Windahl. *Communication models for the study of mass communications*. Routledge, 2015. ISBN: 9781317900689.

[52] Herbert H Clark and Susan E Brennan. "Grounding in communication". In: *Perspectives on socially shared cognition*. American Psychological Association, 1991, pp. 127–149. DOI: 10.1037/10096-006.

[53] Simon Garrod and Martin J Pickering. "Joint action, interactive alignment, and dialog". In: *Topics in Cognitive Science* 1.2 (2009), pp. 292–304. DOI: 10.1111/j.1756-8765.2009.01020.x.

[54] Simon Garrod and Martin J Pickering. "Why is conversation so easy?" In: *Trends in cognitive sciences* 8.1 (2004), pp. 8–11. ISSN: 1364-6613. DOI: 10.1016/j.tics.2003.10.016.

[55] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences". In: *arXiv preprint arXiv:1712.00547* (2017). DOI: 10.48550/arXiv.1712.00547.

[56] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[57] Jeffrey Ens and Philippe Pasquier. "CAEMSI: A Cross-Domain Analytic Evaluation Methodology for Style Imitation." In: *Proceedings of the ICCC*. 2018, pp. 64–71.

[58] Bas Cornelissen, Willem Zuidema, and John Ashley Burgoyne. "Cosine Contours: A Multipurpose Representation for Melodies". In: *Proceedings of ISMIR*. 2021.

[59] US District Court for the District of Massachusetts. "UMG Recordings, Inc. et al. v. Suno, Inc. et al 1:24-cv-11611". More information available at https://www.courtlistener.com/docket/68878608/umg-recordings-inc-v-suno-inc/ (last accessed Aug. 18, 2025). 2024.

[60] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, et al. "The million song dataset". In: *Proc. of the 12th Int. Society for Music Information Retrieval Conf.* Miami, Florida, 2011. DOI: 10.7916/D8NZ8J07.

[61] Yusong Wu, Ke Chen, Tianyu Zhang, et al. "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation". In: *Proc. IEEE ICASSP 2023*. 2023.

[62] Lejaren Hiller and Leonard Isaacson. *Experimental Music: Composition with an Electronic Computer*. New York, USA: McGraw-Hill Book Company, 1959.

[63] Kemal Ebcioğlu. "An expert system for harmonizing four-part chorales". In: *Computer Music Journal* 12.3 (1988), pp. 43–51.

[64] Douglas Eck and Jürgen Schmidhuber. "Finding temporal structure in music: blues improvisation with LSTM recurrent networks". In: *Proceedings of*

the 12th IEEE Workshop on Neural Networks for Signal Processing. 2002, pp. 747–756.

[65] François Pachet. "The Continuator: Musical Interaction With Style". In: *Journal of New Music Research* 32.3 (2003), pp. 333–341.

[66] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. *Deep Learning Techniques for Music Generation*. Springer, 2019.

[67] Adam Roberts, Jesse Engel, Colin Raffel, et al. "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 4364–4373.

[68] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, et al. "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[69] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, et al. "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499* 12 (2016).

[70] Prafulla Dhariwal, Heewoo Jun, Christine Payne, et al. "Jukebox: A generative model for music". In: *arXiv preprint arXiv:2005.00341* (2020).

[71] Dongchao Yang, Jianwei Yu, Helin Wang, et al. "Diffsound: Discrete Diffusion Model for Text-to-Sound Generation". In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 31 (Apr. 2023), pp. 1720–1733.

[72] Flavio Schneider, Ojasv Kamal, Zhijing Jin, et al. *Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion*. arXiv preprint arXiv:2301.11757. 2023.

[73] Chen Zhang, Yi Ren, Kejun Zhang, et al. *SDMuse: Stochastic Differential Music Editing and Generation via Hybrid Representation*. arXiv preprint arXiv:2211.00222. 2022.

[74] Marco Pasini and Jan Schlüter. *Musika! Fast Infinite Waveform Music Generation*. arXiv preprint arXiv:2208.08706. 2022.

[75] Justine Moore and Anish Acharya. *The Future of Music: How Generative AI Is Transforming the Music Industry*. https://a16z.com/the-future-of-music-how-generative-ai-is-transforming-the-music-industry/ (last accessed Aug. 26, 2025). 2023.

[76] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, et al. "AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445219.

[77] Li-Chia Yang and Alexander Lerch. "On the evaluation of generative models in music". In: *Neural Computing and Applications* (2018).

[78] Zongyu Yin, Federico Reuben, Susan Stepney, et al. ""A Good Algorithm Does Not Steal–It Imitates": The Originality Report as a Means of Measuring When a Music Generation Algorithm Copies Too Much". In: *Proceedings of the Artificial Intelligence in Music, Sound, Art and Design. EvoMUSART*. Springer. 2021, pp. 360–375.

[79] Jeffrey Ens and Philippe Pasquier. “Quantifying Musical Style: Ranking Symbolic Music Based on Similarity to a Style”. In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 2019.

[80] Bob L. T. Sturm and Oded Ben-Tal. “Taking the models back to music practice: Evaluating generative transcription models built using deep learning”. In: *Journal of Creative Music Systems* 2 (2017), pp. 32–60.

[81] Bob L. T. Sturm, Oded Ben-Tal, Úna Monaghan, et al. “Machine learning research that matters for music creation: A case study”. In: *Journal of New Music Research* 48.1 (2018), pp. 36–55. DOI: 10.1080/09298215.2018.1515233.

[82] Christopher Ariza. “The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems”. In: *Computer Music Journal* 33.2 (2009), pp. 48–70. ISSN: 01489267, 15315169. URL: http://www.jstor.org/stable/40301027 (visited on 07/29/2025).

[83] Jeffrey Ens and Philippe Pasquier. “Quantifying Musical Style: Ranking Symbolic Music based on Similarity to a Style”. In: *arXiv preprint arXiv:2003.06226* (2020).

[84] Bob L.T. Sturm and Hugo Maruri-Aguilar. “The Ai Music Generation Challenge 2020: Double Jigs in the Style of O’Neill’s 1001”. In: *Journal of Creative Music Systems* 5.1 (2021). DOI: 10.5920/jcms.950.

[85] Bob L.T. Sturm. “The AI Music Generation Challenge 2021: Summary and results”. In: *Proceedings of the 3rd Conference on AI Music Creativity*. Digitala Vetenskapliga Arkivet, 2022, pp. 1–10. DOI: 10.5281/zenodo.7088406.

[86] Bob L.T. Sturm. “The AI Music Generation Challenge 2022: Summary and results”. In: *AIMC 2023*. 2023. URL: https://aimc2023.pubpub.org/pub/ynxz9uu7.

[87] Bob L.T. Sturm. *The AI Music Generation Challenge 2023*. Github. Retrieved August 20, 2025. Aug. 2023. URL: https://github.com/boblsturm/aimusicgenerationchallenge2023/blob/main/README.md.

[88] Bob L. T. Sturm. “An artificial critic of Irish double jigs”. In: *Proceedings of AI Music Creativity*. 2021, p. 10.

[89] Bob L. T- Sturm, João Felipe Santos, Oded Ben-Tal, et al. “Music Transcription Modelling and Composition Using Deep Learning”. In: *Proc. Conf. Computer Simulation of Musical Creativity*. Huddersfield, UK, 2016.

[90] David Meredith, Kjell Lemström, and Geraint A. Wiggins. “Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music”. In: *Journal of New Music Research* 31.4 (2002), pp. 321–345. DOI: 10.1076/jnmr.31.4.321.14162.

[91] Tom Collins, Jeremy Thurlow, Robin Laney, et al. “A Comparative Evaluation of Algorithms for Discovering Translational Patterns in Baroque Keyboard Works”. In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 2010.

[92] Chantal Buteau and John Vipperman. “Melodic Clustering within Motivic Spaces: Visualization in OpenMusic and Application to Schumann’s Träumerei”. In: *Proceedings of Mathematics and Computation in Music*. Ed. by Timour Klouche and Thomas Noll. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 59–66. ISBN: 978-3-642-04579-0.

[93] Wai Man Szeto and Man Hon Wong. “A graph-theoretical approach for pattern matching in post-tonal music analysis”. In: *Journal of New Music Research* 35.4 (2006), pp. 307–321. DOI: 10.1080/09298210701535749.

[94] Zoltán Juhász. “A systematic comparison of different European folk music traditions using self-organizing maps”. In: *Journal of New Music Research* 35.2 (2006), pp. 95–112. DOI: 10.1080/09298210600834912.

[95] Darrell Conklin and Christina Anagnostopoulou. “Comparative Pattern Analysis of Cretan Folk Songs”. In: *Journal of New Music Research* 40.2 (2011), pp. 119–125. DOI: 10.1080/09298215.2011.573562.

[96] Bertrand Harris Bronson. “Some Observations about Melodic Variation in British-American Folk Tunes”. In: *Journal of the American Musicological Society* 3.2 (1950), pp. 120–134. ISSN: 00030139, 15473848.

[97] Merwin Olthof, Berit Janssen, and Henkjan Honing. “The Role of Absolute Pitch Memory in the Oral Transmission of Folksongs”. In: *Empirical Musicology Review* 10.3 (2015), pp. 161–174. DOI: 10.18061/emr.v10i3.4435.

[98] Anja Volk and Peter van Kranenburg. “Melodic similarity among folk songs: An annotation study on similarity-based categorization in music”. In: *Musicae Scientiae* 16.3 (2012), pp. 317–339. DOI: 10.1177/1029864912448329.

[99] Seán Doherty. “Melodic Structures in the Double Jigs of O’Neill’s The Dance Music of Ireland: 1001 Gems (1907)”. In: *J. Soc. Musicology in Ireland* (2022), pp. 19–45.

[100] David Huron. “The Melodic Arch in Western Folksongs”. In: *Computing in Musicology* (1996).

[101] Md Awsafur Rahman, Zaber Ibn Abdul Hakim, Najibul Haque Sarker, et al. “SONICS: Synthetic Or Not - Identifying Counterfeit Songs”. In: *The Thirteenth International Conference on Learning Representations*. last accessed Aug. 18, 2025. 2025. URL: https://openreview.net/forum?id=PY7KSh29Z8.

[102] Yupei Li, Manuel Milling, Lucia Specia, et al. “From Audio Deepfake Detection to AI-Generated Music Detection–A Pathway and Overview”. In: *arXiv preprint arXiv:2412.00571* (2024).

[103] Darius Afchar, Gabriel Meseguer-Brocal, Kamil Akesbi, et al. “A Fourier Explanation of AI-music Artifacts”. In: *Proceedings of the 26th International Society for Music Information Retrieval Conference*. 2025.

[104] IRCAM Amplify. *AI Generated Music Detector*. https://www.ircamamplify.io/product/ai-generated-music-detector (last accessed Apr. 23, 2025). 2025.

[105] Deezer. *Deezer deploys cutting-edge AI detection tool for music streaming*. https://newsroom-deezer.com/2025/01/deezer-deploys-cutting-edge-ai-detection-tool-for-music-streaming/ (last accessed Apr. 23, 2025). Deezer Newsroom, Jan. 2025.

[106] Music Business Worldwide. *Believe has developed its own AI-made music detector with 98% accuracy. What might this mean for the future?* https://www.musicbusinessworldwide.com/believe-has-developed-its-own-ai-made-music-detector-with-98-accuracy-what-might-this-mean-for-the-future1/ (last accessed Apr. 23, 2025). 2025.

[107] YouTube Help. *How Content ID works*. https://support.google.com/youtube/answer/2797370 (last accessed Apr. 23, 2025). 2025.

[108] Audible Magic. *Ensuring Content Integrity: Suno Partners with Audible Magic for User Uploads*. https://www.audiblemagic.com/2024/10/18/ensuring-content-integrity-suno-partners-with-audible-magic-for-user-uploads/ (last accessed Sep. 4, 2025). 2024.

[109] Audible Magic. *Udio Partners with Audible Magic to Fingerprint AI-Generated Tracks and to Check for Infringements*. https://www.audiblemagic.com/2025/04/30/udio-partners-with-audible-magic-to-fingerprint-ai-generated-tracks-and-to-check-for-infringements/ (last accessed Sep. 4, 2025). 2025.

[110] Bill Manaris, Juan Romero, Penousal MacHado, et al. "Zipf's Law, Music Classification, and Aesthetics". In: *Computer Music Journal* 29.1 (2005), pp. 55–69. ISSN: 01489267, 15315169.

[111] Jacek Wołkowicz, Zbigniew Kulka, and Vlado Kešelj. "N-gram-based approach to composer recognition". In: *Archives of Acoustics* 33.1 (2008), pp. 43–55.

[112] Maximos A Kaliakatsos-Papakostas, Michael G Epitropakis, and Michael N Vrahatis. "Weighted Markov Chain Model for Musical Composer Identification". eng. In: *Proceedings of Applications of Evolutionary Computation*. Vol. 6625. 2. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 334–343. ISBN: 9783642205194.

[113] Thierry Bertin-Mahieux, Douglas Eck, and Michael Mandel. "Machine Audition: Principles, Algorithms and Systems". In: ed. by W. Wang. IGI Publishing, 2010. Chap. Automatic tagging of audio: The state-of-the-art.

[114] Keunwoo Choi, Gyorgy Fazekas, Mark Sandler, et al. "Convolutional recurrent neural networks for music classification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[115] Jordi Pons and Xavier Serra. "musicnn: Pre-trained convolutional neural networks for music audio tagging". In: *arXiv preprint arXiv:1909.06654* (2019). DOI: 10.48550/arXiv.1909.06654.

[116] Weixin Liang, Mert Yuksekgonul, Yining Mao, et al. "GPT detectors are biased against non-native English writers". In: *Patterns* 4.7 (2023). DOI: 10.1016/j.patter.2023.100779.

[117] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI magazine* 38.3 (2017), pp. 50–57.

[118] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[119] Arun Das and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey". In: *arXiv preprint arXiv:2006.11371* (2020). DOI: 10.48550/arXiv.2006.11371.

[120] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[121] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey”. In: *Proceedings of the 41st International convention on information and communication technology, electronics and microelectronics*. IEEE. 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040.

[122] Michael Chromik and Andreas Butz. “Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces”. In: *Proceedings of the 18th IFIP TC 13 International Conference on Human-Computer Interaction*. Springer International Publishing, 2021, pp. 619–640. ISBN: 978-3-030-85616-8.

[123] Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, et al. “Towards Explainable Music Emotion Recognition: The Route via Mid-level Features”. In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*. 2019, pp. 237–243.

[124] Markus Langer, Daniel Oster, Timo Speith, et al. “What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research”. In: *Artificial Intelligence* 296 (2021), p. 103473.

[125] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Model-agnostic interpretability of machine learning”. In: *arXiv preprint arXiv:1606.05386* (2016).

[126] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[127] Julius Adebayo, Justin Gilmer, Michael Muelly, et al. “Sanity Checks for Saliency Maps”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Vol. 31. Montréal, Canada: Curran Associates, Inc., 2018.

[128] Zachary C Lipton. “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue* 16.3 (2018), pp. 31–57.

[129] Òscar Celma and Xavier Serra. “FOAFing the music: Bridging the semantic gap in music recommendation”. In: *Journal of Web Semantics* 6.4 (2008), pp. 250–256. ISSN: 1570-8268. DOI: 10.1016/j.websem.2008.09.004.

[130] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. “Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering”. In: *International Journal of Human–Computer Interaction* (2023), pp. 1–23. DOI: 10.1080/10447318.2024.2431761.

[131] Rebecca Fiebrink, Dan Trueman, and Perry R. Cook. “A Meta-Instrument for Interactive, On-the-Fly Machine Learning”. In: *Proceedings of the International Conference on New Interfaces for Musical Expression*. Zenodo, 2009, pp. 280–285. DOI: 10.5281/zenodo.1177513.

[132] Nick Bryan-Kinns, Berker Banar, Corey Ford, et al. “Exploring XAI for the Arts: Explaining Latent Space in Generative Music”. In: *Workshop on eXplainable AI approaches for debugging and diagnosis, NeurIPS*. 2021.

[133] Sebastian Lobbers and George Fazekas. “SketchSynth: a browser-based sketching interface for sound control”. In: *Proceedings of the International Conference on New Interfaces for Musical Expression*. Zenodo, 2023, pp. 637–641. DOI: 10.5281/zenodo.11189331.

[134] Hugo Scurto and Ludmila Postel. “Soundwalking Deep Latent Spaces”. In: *Proceedings of the 23rd International Conference on New Interfaces for Musical Expression (NIME'23)*. 2023.

[135] Nick Bryan-Kinns, Corey Ford, Alan Chamberlain, et al. “Explainable AI for the Arts: XAIxArts”. In: *Proceedings of the 15th Conference on Creativity and Cognition*. C&C '23. Association for Computing Machinery, 2023, pp. 1–7. ISBN: 9798400701801. DOI: 10.1145/3591196.3593517.

[136] Michael Zbyszynski, Atau Tanaka, and Federico Visi. “Interactive machine learning: Strategies for live performance using electromyography”. In: *Open Source Biomedical Engineering* (2020).

[137] Tim Murray-Browne and Panagiotis Tigas. “Latent Mappings: Generating Open-Ended Expressive Mappings Using Variational Autoencoders”. In: *Proceedings of NIME*. 2021.

[138] Michael Clemens. “Explaining the arts: Toward a framework for matching creative tasks with appropriate explanation mediums”. In: *arXiv preprint arXiv:2308.09586* (2023).

[139] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, et al. “Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle”. In: *Proceedings of the Designing Interactive Systems Conference*. DIS '21. ACM, 2021, pp. 1591–1602. DOI: 10.1145/3461778.3462131.

[140] Nick Bryan-Kinns. “Reflections on Explainable AI for the Arts (XAIxArts)”. In: *Interactions* 31.1 (2024), pp. 43–47. ISSN: 1072-5520. DOI: 10.1145/3636457.

[141] Francis O'Neill. *The Dance Music of Ireland: 1001 Gems. Double Jigs, Single Jigs, Hop or Slip Jigs, Reels, Hornpipes, Long Dances, Set Dances, etc.* Ed. by Francis O'Neill. Collected and edited by Captain Francis O'Neill; arranged by Sergeant James O'Neill. Chicago, IL: Lyon & Healy, 1907, p. 172.

[142] Laurens van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[143] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[144] Bob L. T. Sturm, Joao Felipe Santos, Oded Ben-Tal, et al. “Music transcription modelling and composition using deep learning”. In: *1st Conf. Computer Simulation of Musical Creativity*. Huddersfield, UK, July 2016.

[145] IRCAM Amplify. *Why Our AI Music Detector Stays Ahead*. https://www.ircamamplify.io/blog/why-our-ai-music-detector-stays-ahead (last accessed Apr. 23, 2025). 2025.

[146] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, et al. “High fidelity neural audio compression”. In: *arXiv preprint arXiv:2210.13438* (2022).

[147] US District Court for the District of Massachusetts. “UMG Recordings, Inc. et al. v. Uncharted Labs, Inc. et al 1:24-cv-04777”. More information available at https://www.courtlistener.com/docket/68878697/umg-

recordings-inc-v-uncharted-labs-inc-dba-udiocom/ (last accessed Aug. 18, 2025). 2024.

[148] Fábio M. Miranda, Niklas Köhnecke, and Bernhard Y. Renard. "HiClass: a Python Library for Local Hierarchical Classification Compatible with Scikit-learn". In: *Journal of Machine Learning Research* 24.29 (2023), pp. 1–17. URL: http://jmlr.org/papers/v24/21-1518.html.

[149] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673. DOI: 10.1038/s42256-020-00257-z.

[150] Bob L. T. Sturm. "A simple method to determine if a music information retrieval system is a "horse"". In: *IEEE Transactions on Multimedia* 16.6 (2014), pp. 1636–1644. DOI: 10.1109/TMM.2014.2330697.

[151] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91.

[152] Georgina Born. "Diversifying MIR: Knowledge and Real-World Challenges, and New Interdisciplinary Futures". In: *Transactions of the International Society for Music Information Retrieval* 3.1 (2020), pp. 193–204. DOI: 10.5334/tismir.58.

[153] Run Wang, Felix Juefei-Xu, Yihao Huang, et al. "DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Association for Computing Machinery, 2020, pp. 1207–1216. ISBN: 9781450379885. DOI: 10.1145/3394171.3413716.

[154] Yisroel Mirsky and Wenke Lee. "The creation and detection of deepfakes: A survey". In: *ACM computing surveys (CSUR)* 54.1 (2021), pp. 1–41.

[155] Di Cooke, Abigail Edwards, Sophia Barkoff, et al. *As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli*. 2025. arXiv: 2403.16760.

[156] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, et al. "FMA: A Dataset For Music Analysis". In: *18th International Society for Music Information Retrieval Conference*. 2017.

[157] David Gunning, Mark Stefik, Jaesik Choi, et al. "XAI—Explainable artificial intelligence". In: *Science Robotics* 4.37 (2019). DOI: 10.1126/scirobotics.aay7120.

[158] Been Kim. *Beyond interpretability: developing a language to shape our relationships with AI*. https://medium.com/@beenkim/beyond-interpretability-4bf03bbd9394 (last accessed Aug. 18, 2025). 2022.

[159] Bob L. T. Sturm and Nick Collins. "The kiki-bouba challenge: Algorithmic composition for content-based mir research & development". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan: ISMIR, 2014, pp. 21–26. DOI: 10.5281/zenodo.1416364.

[160] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[161] Brian McFee, Colin Raffel, Dawen Liang, et al. “librosa: Audio and music signal analysis in python”. In: *Proceedings of the 14th python in science conference*. Vol. 8. Citeseer. 2015, pp. 18–25.

[162] Saumitra Mishra, Bob L. T. Sturm, and Simon Dixon. “Local Interpretable Model-Agnostic Explanations for Music Content Analysis”. In: *Proceedings of the 18th International Society for Music Information Retrieval Conference*. Suzhou, China: ISMIR, 2017, pp. 537–543. DOI: 10.5281/zenodo.1417387.

[163] Saumitra Mishra, Emmanouil Benetos, Bob L. T. Sturm, et al. “Reliable Local Explanations for Machine Listening”. In: *Proceedings of the International Joint Conference on Neural Networks*. Glasgow, United Kingdom: IEEE, 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9207444.

[164] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, et al. “Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review”. In: *Diagnostic pathology* 19.1 (2024), pp. 43–9. DOI: 10.1186/s13000-024-01464-7.

[165] Ethan Goh, Robert Gallo, Jason Hom, et al. “Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial”. In: *JAMA network open* 7.10 (2024), e2440969. DOI: 10.1001/jamanetworkopen.2024.40969.

[166] Patrick G. T. Healey, Jan P. de Ruiter, and Gregory J. Mills. “Editors’ Introduction: Miscommunication”. In: *Topics in Cognitive Science* 10.2 (2018), pp. 264–278. DOI: 10.1111/tops.12340.

[167] CE Shannon and W Weaver. “The mathematical theory of communication. Univ. of Illinois Press, Urbana.” In: *The mathematical theory of communication. Univ. of Illinois Press, Urbana.* (1949).

[168] Herbert H Clark and Edward F Schaefer. “Contributing to discourse”. In: *Cognitive science* 13.2 (1989), pp. 259–294. ISSN: 0364-0213. DOI: 10.1016/0364-0213(89)90008-6.