Leveraging Game Mechanics for Dynamic Music Co-Creation

Mário Escarce Junior, BSc (Hons), MRes
School of Computing and Communications
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

April, 2025

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words, including appendices and footnotes, but excluding the bibliography. A rough estimate of the word count is 37588.

Mário Escarce Junior

Leveraging Game Mechanics for Dynamic Music Co-Creation

Mário Escarce Junior, BSc (Hons), MRes.
School of Computing and Communications, Lancaster University
A thesis submitted for the degree of *Doctor of Philosophy*. April, 2025

Abstract

This thesis explores game mechanics for creating coherent artworks, with a specific focus on music composition. The investigation spans three interconnected areas of partially-autonomous and autonomous systems: implicit cooperation, meta-interactivity, and autonomous generation. The main research question is: How do different degrees of human-machine collaboration, ranging from implicit cooperation to meta-interactivity and machine autonomy, impact the quality, user experience, and aesthetic attributes of music created within virtual environments?

The thesis begins by exploring co-creativity within games through an algorithm fostering unwitting cooperation between humans and gameplay mechanics, termed implicit cooperation. This approach enables cooperative music emergence, fostering engaging artistic collaborations and pleasant musical experiences.

Next, it introduces meta-interactivity for music creation, empowering novices to achieve unexpected outcomes in composition and practice. Using imagetic elements in a virtual environment, this approach converts ludic interactions into music. A user study involving experts and novices highlights its potential to unlock creativity in individuals with limited musical training, while also prompting questions about the role of human sentiment and expressivity in dynamic artistic creation.

Lastly, this thesis presents an autonomous system for dynamically generating immersive soundscapes for games and artistic installations. This system simultaneously produces music and images, preserving human intent and coherence. An algorithm for audiovisual instance generation demonstrates its effectiveness compared to alternatives.

Through these explorations, this thesis sheds light on the evolving landscape of music co-creation, proposing novel interactive experiences based on game mechanics. It aims to contribute to the ongoing debate on the collaborative potential between humans and autonomous systems, with a specific emphasis on their transformative influence within the domains of music and games. By examining the impact of partially and fully autonomous systems on human sentiment, this research offers insights into the evolving relationship between humans and technology, as well as the intricate interplay between music and imagery in audiovisual works, presenting promising avenues for future research and innovation in this domain.

Publications

Contributing publications

- **M. Escarce Junior**, G. R. Martins, L. S. Marcolino, E. Rubegni. "The Aesthetics of Disharmony: Harnessing Sounds and Images for Dynamic Soundscapes Generation". Proceedings. ACM Hum.-Comput. Interact., CHI PLAY (September 2023), 33 pages.

- **M. Escarce Junior**, G. R. Martins, L. S. Marcolino, E. Rubegni. 2021. "A Meta-interactive Compositional Approach that Fosters Musical Emergence through Ludic Expressivity". Proceedings. ACM Hum.-Comput. Interact. 5, CHI PLAY, Article 262 (September 2021), 32 pages.

- **M. Escarce Junior**, G. R. Martins, L. S. Marcolino, Y. T. dos Passos. "Emerging Sounds Through Implicit Cooperation: A Novel Model for Dynamic Music Generation". In Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2017), Utah, USA, October 2017, 7 pages.

Additional publications

- **M. Escarce Junior**. "Meta-interactivity and Playful Approaches for Musical Composition'. In CHI - Conference on Human Factors in Computing Systems. Extended Abstracts (pp. 1-4). CHI Doctoral Consortium, New Orleans, USA, 2022.

To my family —

for your unwavering love, strength, and support,
in every step of this journey.

Acknowledgements

First and foremost, I thank Georgia for her support, shared goals, and the many long hours spent side by side co-developing the games and interactive experiences presented in this work. I also want to thank Ivy, Kiwi, and Totoro – my beloved birds – for the joy, company, and gentle reminders to take breaks when I needed them most.

I'm deeply grateful to Leandro, my advisor and friend, for his encouragement, guidance, and for always pushing me to explore new ideas.

To my family – my mothers, my brothers and sister, and my father – thank you for your resilience, strength, and love.

I'm also thankful to the professors and colleagues who supported this journey with insights, feedback, and encouragement. Special thanks to Jalver and João Victor. Warm thanks as well to all my colleagues at CoLab for the enriching weekly meetings throughout the program – sharing ideas and inspiration. I also extend my gratitude to Magy and my colleagues from the GUII Lab at UCSC, where I spent part of my final year engaged in amazing project discussions – a valuable experience that helped pave the way for my next steps.

To all the staff at Lancaster University who contributed directly or indirectly throughout the program, especially during the challenging times of the pandemic – thank you for the shared learning and support.

Lastly, to my friends and all those who walk with me: your presence made all the difference. Thank you!

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence.

AIVA Artificial Intelligence Virtual Artist.

AR Augmented Reality.

BGM Background Music.

FSMC Functional Scaffolding for Musical Composition.

HCI Human Computer Interaction.

IFMG Federal Institute of Minas Gerais.

IVMI Immersive Virtual Musical Instrument.

MIDI Musical Instrument Digital Interface.

PCG Procedural Content Generation.

RQ Research Question.

UFMG Federal University of Minas Gerais.

FUMEC Minas Gerais Foundation of Education and Culture.

VR Virtual Reality.

VRMI Virtual Reality Musical Instrument.

Chapter 1

Introduction

The integration of novel artificial intelligence (AI) approaches into creative processes has fostered many possibilities for producing artworks that challenge the distinction between human and machine authorship [102, 93, 106]. These advancements have not only established new forms through which we create, perceive, and engage with art (especially in digital mediums) but have also ignited discussions regarding the essence of artistic expression in light of the automation of traditional artistic practices. The enthusiasm and curiosity surrounding the emerging AI techniques that are reshaping our world and the arts have far-reaching effects across various research domains, including Human-Computer Interaction (HCI) and Computational Creativity. As a result, amidst this excitement, numerous questions, considerations, and concerns emerged [61]. Debates arise regarding the classification of algorithmically created art as authentic artistic expression [39, 40] and the consequent impact on traditional artists, leading to heated discussions, for instance, regarding ownership disputes surrounding AI-generated works [99]. In this way, this fusion raises critical questions about authorship, creative freedom, and the role of human emotion in the creation of contemporary digital artworks. These questions pave the way for examining co-creativity, where human and machine collaboration can lead to innovative artistic expressions. The complex interplay between human creativity and algorithmic precision in art constitutes the core of this thesis, guiding our exploration into co-creative processes and their capacity to generate emotionally resonant and aesthetically coherent music. Moreover, concerns regarding the emotional depth and intrinsic value of content created without explicit human involvement present a hurdle to the further exploration of partially-autonomous and autonomous approaches for art generation [10].

The rise of AI-generated art has catalyzed a reevaluation of creative authorship and the interplay between human intuition and machine efficiency. While tools like DALL-E [102], MidJourney [93], and NightCafe [106] have demonstrated AI's capacity

to produce compelling artworks, they also prompt critical inquiry into the nature of creativity and the potential for genuine collaboration between artists and algorithms. The discourse extends into how these technologies challenge traditional boundaries and prompt new theoretical and practical explorations [61, 39, 40].

Given these considerations, the primary objective of this thesis is to explore approaches for dynamically generating coherent music. Coherent music, within the context of this thesis, is defined as compositions that are logically structured, emotionally resonant, and aligned with human aesthetic sensibilities [118]. Such music resonates with the human ability to recognize and appreciate patterns of sound that convey emotion and meaning. In computational creativity, the collaboration between humans and machines, known as co-creativity, also plays a pivotal role in this thesis. Jordanous [65] defines co-creativity in computational research as environments where at least one of the participants is computational, often leading to innovative outcomes. Davis [25] further elaborates on co-creativity, describing it as a collaborative environment where humans and computers co-improvise in real time to generate a creative product, emphasizing that creativity emerges through the interaction of both the human and the computer, with each party's contributions being mutually influential. This thesis will examine how such collaborative processes can lead to the generation of interesting and unpredictable artistic outcomes, emphasizing the synergy between human intuition and computational algorithms.

This investigation delves into how humans perceive and value artistic works produced through different levels of human-machine collaboration, examining scenarios that range from significant human involvement to those where machines possess greater autonomy. Throughout this thesis, "human-machine" collaboration denotes the interactive and cooperative process between human beings and interactive systems in the creation of art. By exploring these dynamics, the thesis seeks to uncover the implications and transformative potential of art created through both autonomous and partially-autonomous methods, emphasizing the balance between technological innovation and human creative input. The structure of this thesis will comprise several sections that analyze human-machine collaboration in artistic endeavors (partially-autonomous approaches), explore machine-produced art (autonomous approaches), and discuss their respective implications for creativity, ethics, and the evolving role of humans in the creative process.

Partially-autonomous approaches introduce a complex challenge in harmonizing human expressiveness with the limitations imposed by the system. The goal is to ensure outcomes that are coherent and consistent while respecting the core of human creative intent. Additionally, the challenge lies in crafting scenarios where human evaluators perceive these collaborative creations as meaningful. Striking this delicate balance is essential to blend human expressive capabilities with system constraints. The ultimate aim is to produce outputs that possess both structural cohesion and

overall quality standards, enabling human evaluators to identify and appreciate the significance of these combined creations. This investigation not only seeks to integrate collaborative approaches across diverse fields but also to explore its potential to empower individuals within their artistic pursuits. Collaborative techniques can assist individuals in their initial steps, alleviating the often overwhelming learning processes of, for example, mastering a musical instrument. Such techniques hold the potential to unlock an individual's creativity and inspire them to further develop their expertise.

On the other hand, autonomous approaches present challenges in establishing a meaningful dialogue with artists, as the "human touch" becomes more discreet, operating behind the curtains. This challenge involves generating novel aesthetic patterns that invite personal interpretations, enabling individuals to shape their own unique experiences. The objective here is to create outputs that are expressive yet consistent, resulting in scenarios where humans perceive the creations as both meaningful and coherent. Achieving this balance requires the system's autonomous decision-making to align with the preferences and style guides set by human artists. It is also worth noting that while recent advancements in Machine Learning (ML) enabled new Computational Creative capabilities, these developments often prioritize AI-centric approaches [87]. This trend can limit access for human creators who lack a deep understanding of AI models, hindering effective collaboration between human expertise and machine capabilities. In this way, the goal is to yield experiences that are not only creatively innovative but also coherent and comprehensible. These attributes would enable human evaluators to discern the value and importance of partially-autonomous creations, potentially integrating them into their own projects.

Recent endeavors in autonomous and partially-autonomous art generation have explored the translation of human-created images into music and vice versa, adapting to narrative shifts or desired emotional tones in interactive mediums like games [101, 123]. However, simultaneously generating music and images while retaining their inherent coherence presents challenges. The absence of a "translation guideline" or an initial stimulus, such as an image that directs the creation of corresponding music, complicates establishing coherence between the auditory and visual elements. This simultaneous generation increases the risk of producing arbitrary audio and video outcomes, which could dilute their meaning and undermine the overall experiential quality due to the divergences in their creative intents. Thus, a critical challenge is preserving the intended meaning and coherence between audio and visual elements during simultaneous generation. Developing innovative strategies to synchronize the generation process is imperative, aiming for seamless alignment between audio and visual elements to enhance the overall experience and achieve a unified fusion of these components.

Therefore, this thesis will explore partially-autonomous and autonomous approaches for the real-time generation of coherent audiovisual instances. It will be

presented three distinct systems, each focusing on a different aspect and proposing its unique approach. The first two systems investigate different techniques for partially-autonomous approaches, one involving "indirect" user intent for music creation, and the other based on aware or "lucid" music creation. Both approaches aim to foster a balance between preserving the artistic intent and expressivity of humans while ensuring a coherent generation of outcomes, such as a musical corpus. The aim is to avoid imposing excessive constraints on human creativity. The third system, on the other hand, explores a more autonomous approach to audiovisual generation, although still preserving human intent.

The first system, which we named Implicit Cooperation, is designed to enable players to unintentionally create a coherent musical corpus through their interaction with game worlds. The second system explores an approach we call Meta-interactivity, where players engage in ludic interactions, such as bursting bubbles in a virtual environment, to create music through a system that resembles a kind of ludic musical instrument. This system guides novices in musical theory and practice, facilitating the creation of interesting music through a color-to-tone translation approach. The third system addresses the challenge of simultaneous music and video generation while preserving coherence. Through this system, developers can create and synchronize music, landscapes, or both, ensuring a cohesive and harmonious audiovisual experience. This system empowers artists and developers to maintain control and shape the desired outcome in terms of intent, aesthetics, and mood. It offers versatility and can be implemented across various interactive applications, including games, and can support small development teams alleviating overwhelming tasks.

The upcoming chapters will explore the theoretical foundations, methodologies, and evaluations of partially-autonomous and autonomous audiovisual generation. This research contributes to understanding algorithmic art and its integration into game development, human-computer interaction, and computational creativity. It also addresses ethical considerations in these fields, aiming to advance the responsible implementation of generative art and music. Recent AI techniques, while promising, are not deeply explored due to timing and focus. At the start of this research, methods like large language models and stable diffusion were unavailable. This work emphasizes human-machine co-creativity, prioritizing human intuition and emotional engagement over high-autonomy AI, as the complexity of cutting-edge AI could detract from using game mechanics for artistic expression and collaboration.

1.1 Confluence of Human Emotion and Machine Expression

Art has long served as a powerful medium for human expression, allowing individuals to convey their deepest emotions and thoughts. In recent times, the fusion of art and technology, particularly through digital mediums, has led to the emergence of Art Games, a genre that leverages the interactive and immersive nature of games to create rich artistic experiences [85, 38]. These art forms enable artists to explore and integrate various creative facets, such as aesthetics, visuals, music, and narrative, into experiences that resonate with specific feelings, messages, or concepts.

The advent of machine learning techniques has also significantly transformed the artistic landscape, influencing how art is created, interpreted, and valued. This technological evolution has shifted the balance from human expressivity, reflection, and intent toward a more objective, data-driven approach, often resulting in artworks that are less abstract and more figurative [93]. This trend raises questions about the delineation between art and design, as the processes of reflection and contemplation increasingly give way to pragmatic utility and application.

Within this evolving context, the collaboration between humans and machines can be visualized as a spectrum. On one end lies the pure expression of human creativity, where art is crafted solely based on individual skills, imagination, and intuition. On the opposite end, we encounter machine-generated content, predominantly driven by algorithms and machine learning, where the focus is on computational efficiency and pattern generation. Yet, it is within the intermediary space that the most intriguing possibilities arise – where human expressivity intersects with sophisticated interactive systems, yielding innovative and sometimes unexpected creative outputs.

This thesis will explore these hybrid collaborative modalities, particularly how dynamic interfaces can facilitate a balanced collaboration between humans and systems based on game-like mechanics. It investigates scenarios where artists maintain creative control, as well as those where machines operate with autonomy, never completely sidelining human intervention. The underlying premise is that autonomous systems should complement rather than supplant human creativity, thereby enhancing and expanding the artist's capabilities through synergistic collaboration.

An aspect of this research is the examination of how collaborative art creation transcends traditional forms, utilizing game-based interactive systems to offer new perspectives and enhance creative expression. This approach not only democratizes the act of creation – making it accessible to novices and those without specific artistic training – but also enriches the overall artistic landscape by integrating diverse cognitive and emotional layers into the creative process.

1.2 Exploring Co-Creativity for Music Creation

This thesis explores the dynamic generation of music through three systems that represent a spectrum of human-machine collaboration. These systems, categorized under partially-autonomous and autonomous generation, are pivotal in understanding the evolving role of interactive systems in artistic creation, impacting fields like artistic expression and game development.

The approaches discussed in this thesis – Implicit Cooperation, Meta-interactivity, and Autonomous Generation – each delineate distinct facets of game mechanics and AI integration into the creative process. These systems offer insights into the synergistic potential of human-machine collaboration for music co-creation. In the field of computational creativity, co-creativity involves at least one participant being computational [65]. They not only enhance artistic productivity and creativity but also raise fundamental questions about the nature of creativity, the role of technology in artistic endeavors, and some of the ethical implications of machine-generated art.

By exploring these concepts, this research aims to elucidate the complex dynamics of machine-assisted art creation, advancing the discourse on computational creativity and setting the stage for future discussions and innovations in the field.

1.2.1 Implicit Cooperation

Implicit Cooperation denotes a system where interaction with AI is subtle, often unbeknownst to the user, facilitating music composition as a byproduct of other activities, like gameplay. This concept, which originated in this research, emphasizes the intelligent system's role as a background enhancer of the creative experience, where the system generates aesthetically pleasing outcomes without overt user intent for music creation.

In this thesis, Implicit Cooperation refers to music generation occurring as an incidental byproduct of user interaction within interactive systems. While this concept shares commonalities with existing paradigms in interactive art and game design, our research focuses on elucidating the mechanisms through which such interactions can spontaneously foster music generation. This approach aims to bridge identified gaps in the current literature, providing a fresh perspective on user-driven creative processes.

This approach, highlighted in our paper "Emerging Sounds Through Implicit Cooperation: A Novel Model for Dynamic Music Generation" [38], challenges traditional paradigms by melding user interaction with algorithmic creativity, nurturing an environment where organic, user-influenced musical pieces emerge, enriching the gaming experience and extending the domain of computational creativity.

1.2.2 Meta-interactivity

Meta-interactivity advances the role of users in the creative process, leveraging procedural content generation (PCG) to enable active participation in music creation. This system, which we introduce in this thesis, transforms players into composers, using the game's interface as a musical instrument.

Meta-Interactivity, another concept introduced in this thesis, encapsulates the idea of users engaging consciously and purposefully with systems to steer the creative output, particularly in the realm of music generation. While instances of interactivity influencing creativity have been explored in digital media studies, our work seeks to enhance and refine these interactions, thereby amplifying their impact on the musical creative process and offering innovative contributions to the field of computational music.

This approach is designed to be accessible, allowing those without musical training to produce interesting compositions, thus democratizing the art of music creation. The system's underpinning philosophy and its contribution to fostering musical emergence through ludic expressivity are discussed in our paper "A Meta-interactive Compositional Approach that Fosters Musical Emergence through Ludic Expressivity" [39].

1.2.3 Autonomous Generation

Autonomous Generation refers, in the context of this thesis, to independent creative processes that generate music and visual art through a rule-based approach, relying on embedded algorithms rather than direct human input or machine learning models. While machine learning techniques are also a form of autonomous generation, this thesis focuses specifically on a rule-based framework. This distinction emphasizes transparency, interpretability, and control over the creative process, in contrast to machine learning models, which often function as black boxes with limited insight into their internal logic.

By integrating rule-based autonomous processes with interactive elements, this approach facilitates a deeper exploration of the interplay between user input and algorithmic output, providing valuable insights into the complexities of creative expression. Our paper, "The Aesthetics of Disharmony: Harnessing Sounds and Images for Dynamic Soundscapes Generation" [40], further examines these ideas, illustrating how rule-based autonomous systems can retain a sense of human influence while expanding the boundaries of artistic creation.

1.3 Research Gaps

Despite the growth in Computational Creativity and Co-Creativity studies, notable research gaps persist, particularly in understanding the spectrum of user agency in co-creative systems and its impact on producing emotionally resonant and aesthetically compelling artistic outputs. Prior research has predominantly focused on the technical aspects of human-machine interaction in the creative process. However, there is a lack of comprehensive exploration into how varying degrees of user involvement – from passive observation to active creation – affect the emotional and aesthetic dimensions of the produced art, especially in music composition.

The interplay between human emotional depth and machine computational capabilities is particularly underexplored in contexts requiring the generation of music that resonates with humans on an emotional level. Although studies by Jordanous [65] and Davis [25] have examined co-creativity mechanisms, there remains a gap in understanding how systems can be designed to offer a spectrum of user agency that augments human creativity rather than merely automating it. This spectrum, which in this thesis ranges from implicit cooperation to meta-interactive systems and autonomous generation, necessitates further investigation to discern how different levels of user control and interaction influence the creative output and user experience.

Moreover, while the role of autonomous and partially autonomous systems in the creative process has been recognized [67, 103], their potential to collaborate with humans to produce innovative artistic outcomes requires deeper examination. This involves assessing how these systems can support and enhance the creative intentions of human artists, balancing machine efficiency with human expressiveness, and how game mechanics can be leveraged to create engaging co-creative experiences. Understanding these dynamics can enhance the human creative process, making it more intuitive, enjoyable, and accessible.

Therefore, this thesis aims to address these gaps by providing insights into co-creativity mechanisms, particularly in music composition, and examining the interplay between human creativity, game mechanics, algorithms, and the spectrum of user agency. This will contribute to a better understanding of the potential and limitations of current technologies in enhancing the creative process and propose new frameworks for music co-creation.

1.3.1 Research Questions

In light of the identified research gaps, the Main Research Question (MRQ) of this thesis is:

"How do different levels of human-machine collaboration, ranging from partially-autonomous to fully-autonomous approaches (i.e. implicit cooperation, meta-interactivity,

and machine autonomy) affect the quality, user experience, and aesthetic properties of music produced in virtual environments?"

To address the nuances of this main question, the following specific research questions (RQs) are proposed:

- **RQ1:** How does implicit cooperation between humans and machines influence the emergent qualities of artistic creation, particularly in terms of musical coherence and emotional resonance?
- **RQ2:** In what ways can meta-interactivity enhance the creative agency of users in music generation, and how is this perceived in terms of artistic quality and engagement?
- **RQ3:** Is it possible for autonomous systems to generate coherent compositions of music and images simultaneously?

These questions explore the complex dynamics of human-machine collaboration in the creative process, aiming to foster innovative artistic expressions while maintaining the emotional and aesthetic integrity of the human creator.

1.4 Summary of Contributions

This thesis advances the design and development of interactive musical systems, emphasizing user-centric interfaces, interactive experiences, and the integration of game mechanics. It explores the continuum of human-machine interaction in music creation, providing insights into the dynamics of partially-autonomous and autonomous systems. This facilitates creative collaboration and enables individuals without formal musical training to produce complex compositions, democratizing the music creation process.

The research addresses the balance between human creativity and machine autonomy, contributing to the broader discourse on using technology to enhance human artistic capabilities. It presents methods for ethically and effectively incorporating game mechanics into artistic practices.

Moreover, the thesis intersects with several domains, including algorithmic music, music education, interactive media development, and the use of virtual reality in artistic installations. It investigates the user experience and aesthetic appreciation of collaboratively created music, laying the groundwork for future studies in game design and immersive musical experiences.

An important aspect of this research is the analysis of the emotional impact of machine-generated music. It examines how different computational approaches, from co-creativity to autonomous methods, affect human perception and emotional

engagement. This thesis underscores the potential of integrating technology with human creativity to forge a synergistic relationship. It sets the stage for a new paradigm where technology serves not just as a tool but as a collaborator in the creative process, enriching the human experience of art and music.

1.5 Thesis Outline

Chapter 2: Related Work
This chapter frames the research within the existing scholarly landscape by exploring related works in HCI, AI, co-creativity, music generation, computational creativity, and procedural content generation.
Chapter 3: Background
This chapter provides background information to contextualize the research, exploring the relationship between humans, machines, and expressivity in artistic endeavors. It introduces key concepts like implicit cooperation, meta-interactivity, and machine autonomy within the evolving arts and audiovisual creation paradigm.
Chapter 4: Implicit Cooperation
This chapter focuses on Microbial Art, describing its gameplay mechanics that facilitate emergent music generation through implicit cooperation. It includes a comprehensive evaluation of the system, discussing results, and limitations.
Chapter 5: Meta-interactivity
This chapter details Bubble Sounds, which enables users to engage in meta-interactivity for music generation. It describes the system's functionality, evaluates outcomes, and discusses limitations.
Chapter 6: Machine Autonomy
This chapter presents Solato, a system that generates landscape-inspired music. It provides an overview of Solato, followed by an evaluation discussing its implementation and effectiveness.
Chapter 7: Discussion
This chapter synthesizes the findings, highlighting the strengths and limitations of the systems. It discusses the broader implications of the research and identifies avenues for future exploration and improvement.
Chapter 8: Conclusions
The final chapter summarizes the key findings, emphasizing the research's implications and its impact on art, music, and game development. It outlines future research directions to further advance the collaborative potential between humans and machines in creative endeavors.

Chapter 2

Background

This chapter will provide an overview of the key factors pertinent to the experiences discussed in further detail later in this work. It will explore the musical theory underpinning the developed systems and approaches and illuminate the motivations behind the concepts developed in this research. Furthermore, it will elucidate the multifaceted approaches to music generation, encompassing rule-based systems, co-creativity techniques, and AI methodologies. Understanding these foundations is fundamental to comprehending the complex interplay between computational and creative endeavors that characterizes the innovative landscape of music generation within the context of this thesis.

2.1 Music Background

This section addresses musical definitions relevant to our system and highlights the challenges we aim to address. Drawing upon Edgard Varèse's concept of music as "masses of organized sounds that move against each other" (McAnally, 1995) [86] and David Temperley's emphasis on the listening experience as crucial for meaning emergence [118], we ground our approach in classical music theory while also discussing specific implications brought by interactive musical systems. Varèse's view allows us to explore the spatial and dynamic aspects of sound organization, informing the design of our interactive systems. Similarly, Temperley's perspective guides our understanding of how listeners perceive and derive meaning from music, which is pivotal in shaping the user experience in our system.

To address the subjective nature of musical quality, influenced by factors such as cultural background and individual preferences, we explore "figurative" elements within the generated music. Identifying and emphasizing repetitive structures enables us to anchor the abstract notion of music quality in more tangible, recognizable patterns. Our analysis focuses on fundamental musical elements like harmony and

rhythm, creating loop patterns that not only establish a musical identity but also resonate with the listener's cognitive and emotional frameworks.

Objectively evaluating music generation systems remains a considerable challenge. It requires balancing subjective listener feedback with quantifiable musical metrics. Drawing inspiration from Applebaum's proposal [6], we shift the evaluative lens from a binary classification of "is it music?" to a more nuanced query: "is it interesting?" This rephrasing reframes our assessment criteria and aligns with our aim to foster engagement and stimulate interest through music.

According to some musical theory definitions ([5, 69, 81, 104]), the main concepts that will be discussed further in this work are:

- Note: a note is a single musical sound. They can vary in pitch and duration, and it is a fundamental element of music.
- Tone: a constant sound most frequently characterized by its pitch, such as "C" or "D", which also comprises sound quality (i.e. sound texture), duration, and even intensity. In many forms of music, different tones are changed by modulation or vibrato (fluctuations in height and frequency).
- Melody: a linear succession of musical notes that can be perceived as a single entity that when combined generates variations in pitch and rhythm.
- Harmony: a simultaneous combination of musical notes that evolves across time, producing a pleasant effect among listeners. It can be perceived as a base structure of music, from which the melody comes upon. Along with melody, it is also a very important structure of Western music.
- Chord: multiple notes played simultaneously, the most common being triads (3 notes being played at the same time) and tetrads (4 notes being played at the same time).
- Inverted chords, on the other hand, is a variation of traditional chords, particularly triads. Inversions maintain the same notes as their parent triad but rearrange the order in which these notes are played. This rearrangement gives inverted chords a unique sound and function in musical compositions, allowing for smoother voice leading and harmonic variety. Inversions are designated by which note is in the bass position. For example, if the third note of a triad is in the bass position, it is referred to as the first inversion; if the fifth note is in the bass position, it is the second inversion. Inversions add depth and versatility to chord progressions, enriching the musical experience.
- Octave: an octave is the distance between a given note, like C, and the next repetition of the same note (either higher or lower) in a scale. In this way, we

have a full 12 cycle of notes (e.g. taking the C note as a reference: C, C#, D, D#, E, F, F#, G, G#, A, A#, B, and when we reach the C again). Although it is the same note (C), it is going to be one octave above (or below, in case it is an ascending scale).

- Rhythmic figure: symbols that represent the duration of notes. It is not an absolute measure, since it can vary depending on the beat value and the tempo.
- Scale: a scale is any set of musical notes ordered by a fundamental frequency or pitch. A scale ordered by an increasing pitch is an ascending scale, and a scale ordered by a decreasing pitch is a descending scale. For instance, in a C major scale, we encounter the notes C-D-E-F-G-A-B, spanning one octave. There are many types of scales, and the most common ones that will be mentioned in this thesis are the *chromatic scale*, which presents all the 12 notes in the scale, considering its accidents (C, C#, D, D#, E, F, F#, G, G#, A, A#, B), the *pentatonic scale*, that presents 5 notes per octave, and the *diatonic scale*, that is a sequence of 7 successive natural notes. It includes five whole steps (i.e. whole tones) and two half steps (i.e. semitones) in each octave, in which the two half steps are separated from each other by either two or three whole steps, according to their position in the scale (e.g., if we determine F as a fundamental, we will then have the sequence F—C—G—D—A—E—B).
- Fundamental: a reference note was chosen among all the 12 possible notes in a chromatic scale. Starting from a fundamental, it is possible to build musical scales.
- Pitch: refers to how the human ear perceives the fundamental frequency of sounds. Low frequencies are perceived as low tones and the highest as high tones. Adult humans can detect sounds in a frequency range from 20 Hz to 20 kHz.
- Accident: a quality of a musical note that increases or decreases it in semitones. It allows the same note to sound slightly different, varying from bass to treble (e.g. C#, D#, F#, etc).
- Comma: is the smallest interval the human ear can perceive. Between semitones intervals, such as C - C#, for example, we have these microtonal variations.

2.2 Art Games

The digital game art field has undergone significant transformations, giving rise to innovative forms of creative expression and collaboration between humans and

machines that transcend playful experiences. This thesis delves into two distinct yet interconnected elements that shape the foundation of this research: Art Games and Implicit Cooperation, both in light of game systems that foster the dynamic emergence of music. This exploration will take place mainly in Chapter 4 and Chapter 5 of this thesis.

In the early 2000s, the digital arts witnessed the emergence of a distinctive form of creative expression known as Art Games. These games are distinguished by their combination of innovative gameplay mechanics, thought-provoking narratives, and surreal aesthetics. Often constrained by modest budgets, Art Games are renowned for their unwavering commitment to defying expectations, challenging established paradigms, and delivering unconventional and stimulating experiences.

The term "Art Game" was initially introduced by [53], signifying games intentionally crafted to evoke a broad spectrum of reactions in their audience. While Art Games share certain similarities with mainstream entertainment-focused games, including the incorporation of audio-visual elements and interactive interfaces, they set themselves apart through their unconventional treatment of these elements.

Art Games combine inventive gameplay mechanics, immersive narratives, and surreal aesthetics, challenging traditional gaming conventions. Implicit Cooperation examines interactions among different agents with potentially divergent intents, supporting the emergence of a coherent outcome, such as a musical composition. Our Emergent Music Generation approach introduces a novel method for creating music, where agent movements in a digital environment spontaneously generate melodies.

Within the "genre" of Art Games, a myriad of captivating sub-genres have surfaced, each with its own peculiarities. For instance, *Music Video Games* centralize their gameplay around fundamental musical elements, like rhythm. Although these games share common ground with traditional puzzle games that utilize rhythmic structures to propose challenges, they occupy a distinct category. Examples include titles like *Vib-Ribbon* (Sony Interactive Entertainment, 1999) and *Patapon* (SIE Japan Studio, 2007), where players engage by pressing buttons at predetermined moments to interact with musical elements. Another noteworthy project is *ElectroPlankton* [62], wherein users interact with virtual objects that influence the movement of digital "planktons" responsible for generating music. Importantly, this genre often demands active participation from players, effectively bridging the gap between artistic expression and gameplay mechanics.

The continuously evolving landscape of Art Games has prompted a reevaluation of their classification as a distinct genre. The demarcation between Art Games and other game genres has gradually blurred, mirroring the ongoing evolution of digital art and interactive media. This dynamic invites reflection on the essence of Art Games and their place within the wider field of digital art.

The backdrop for our investigation into Implicit Cooperation, aesthetics, and

player experience is set against the backdrop of the evolving landscape of Art Games and their sub-genres within the digital arts. This evolution aligns with our study of Implicit Cooperation, where collaboration unfolds organically, transcending traditional boundaries. Just as Art Games challenge established gaming paradigms, this research seeks to reveal novel dimensions of player interaction and aesthetic perception, enhancing our comprehension of the dynamic interplay between players and digital arts in interactive experiences.

2.3 Soundscapes and Concrete Music

Chapter 6 of this thesis will explore the concept of soundscapes, drawing inspiration from the work of R. Murray Schafer [108]. The concept of a soundscape, as defined by Schafer, offers a critical lens through which we can analyze the auditory environment that shapes our daily experiences. Within this framework, a soundscape emerges as a composition based on diverse sounds that collectively define a given space. These sounds may emanate from natural or synthetic sources, or even manifest as abstract constructions, contributing to a rich range of sensory perceptions.

Furthermore, the chapter is influenced by the concept of Concrete Music, a technique in musical composition discussed by the French composer Pierre Schaeffer [107]. Concrete Music elevates the ordinary to the extraordinary by employing recorded sounds from everyday objects as the raw materials for musical creation. This approach challenges traditional notions of musical composition, opening new avenues for sonic exploration and artistic expression.

As we explore the interplay of soundscapes and Concrete Music in Chapter 6, the thesis aims to unravel the intricate connections between auditory perception, environmental influence, and the boundaries of musical creativity. How does the dialogue between these two concepts contribute to a deeper understanding of our auditory surroundings and redefine the possibilities within musical composition? In exploring this synthesis, we embark on a journey that transcends the conventional boundaries of sound, inviting a reevaluation of our auditory landscape and the transformative power of everyday sounds in the realm of artistic expression.

Hence, in Chapter 6, this thesis explores the dynamic relationship between soundscapes and Concrete Music, seeking to uncover the intricate links between auditory perception and the creative limits of musical expression. The goal is to motivate readers to step outside the usual limits of sound, urging novel perspectives on our everyday soundscapes and how ordinary sounds can influence artistic expression.

2.4 Rule-Based Systems in Music Generation

Rule-based systems in music generation utilize predefined sets of rules and structures derived from music theory to automate the creation of music. These systems often mimic traditional compositional processes, applying established principles of harmony, melody, and rhythm to generate music that adheres to specific stylistic guidelines [21].

In computer programs, music is typically represented through digital scores or MIDI data, which detail the pitch, duration, velocity, and timbre of each note. Rule-based systems leverage this data to construct musical pieces by following logical sequences and patterns, ensuring that the generated music maintains a coherent structure and aesthetic quality [32].

2.4.1 Algorithmic Composition

Algorithmic composition within rule-based systems employs mathematical models and computational algorithms to create music. Key approaches include:

- **Grammar-based systems**: These systems utilize formal grammar akin to those in language processing to define the syntax and structure of music. By establishing a set of rules that dictate the progression and combination of musical elements, grammar-based systems can generate compositions with logical coherence and stylistic consistency [58].
- **Transition networks**: These involve the use of state machines to manage musical phrases and their transitions. By mapping possible paths through a network of musical states, transition networks can produce dynamic and varied musical sequences, often based on probabilistic choices to add unpredictability and creativity to the composition process [4].

2.5 Co-Creativity Techniques in Music

Co-creativity in music is a synergistic process that entails collaboration between human musicians and computational systems to produce innovative musical works. This collaboration merges the intuitive, emotional, and creative facets of human musicianship with the computational efficiency and data-processing prowess of machines. Interactive composition stands at the core of this co-creative process, where human-machine improvisation and generative systems play pivotal roles [8].

Human-machine improvisation features real-time interaction between musicians and AI systems, where each responds to the other's musical inputs. This dynamic interplay can lead to performances that surpass the capabilities of either participant

alone, achieving a unique blend of human expressiveness and machine precision [72]. Similarly, generative systems act as creative partners, providing musical suggestions, variations, or accompaniments based on human input. Through iterative feedback loops with these systems, musicians can refine and develop their ideas, enhancing the compositional workflow and creative output [20].

This co-creative methodology not only fosters a novel approach to music composition but also amplifies the expressive and innovative potential of musical works, embodying the fusion of human creativity and technological advancement.

Chapter 3

Related Work

This thesis explores the intersection of several fields related to the dynamic generation of music and art in interactive environments, including Human-Computer Interaction (HCI), Artificial Intelligence (AI), Co-Creation, Procedural Content Generation, and Computational Creativity. This chapter provides an overview of research within these areas, aiming to establish a theoretical and methodological foundation for this thesis. It highlights the contributions of each field to the understanding and advancement of music and art generation processes.

Through these foundations, this thesis aspires to contribute to the discourse in computational creativity and music generation. By situating our methodologies and findings within the broader range of existing works, we aim to lay the groundwork for future explorations in interactive music generation systems, thereby advancing the field and opening new avenues for further research.

3.1 Human-Computer Interaction

Interactive design approaches strive to integrate daily life experiences into the virtual environment. SoundSelf: A Technodelic [37] offers a guided meditation experience, where users can explore a sound-based virtual world through voice modulation, fostering a sense of presence in an abstract environment. Liu et al. [73] propose a VR self-transcending flying interface that enhances users' confidence and well-being. Moreover, novel mechanic designs in music are explored, such as by Koray et al. [116], who utilize "cultural probes" – a technique for gathering inspirational life insights – as inputs to digital musical instruments. This encourages the emergence of new musical practices and meditative experiences that align with the contemplative nature of our research. Additionally, the Drift Table [45], an electronic coffee table, uses weight distribution to control slow-moving aerial photography, promoting reflection on ludic design and technology's role in supporting playful activities. In the field of

interactive music interfaces, Electronauts [114] provides a VR platform where users can engage in DJ-like activities, remixing tracks and collaborating with others in a virtual music space. These examples illustrate the evolving landscape of interactive design, contributing to the broader context of our research on meditative and oniric experiences in music generation.

Holland et al. [52] explored Music and Interactivity, highlighting its multifaceted role within HCI research. They note that musical activities have expanded to include not just performance and composition but also collaborative music-making and improvisation between humans and machines. They raise questions about accessibility for newcomers, insights gained from immersed individuals "in the groove", and the engagement of novices with musical theory concerning system mechanics and interfaces. In line with exploring dynamic interactions, the approaches we will present in this thesis engage users in co-creating meaningful experiences, contributing to the field of music interactivity by enabling personalized engagement in music generation processes.

HCI research has increasingly focused on understanding how music influences player experience in gaming contexts. This area aligns with our investigation into the impact of music generation methods on perceptual outcomes and human experiences. Rogers et al. [105] conducted a study assessing the role of music in enhancing player experience within VR settings, discovering that music alters the perception of time duration during gameplay. This finding is particularly important considering the often subordinate role of audio to visuals in gaming environments. Today, with the prevalent use of headsets for communication in online games, the experiential impact of music can be diminished. Altmeyer et al. [3] explored the effect of sound in gamified tasks, noting that sound effects, despite their ubiquity, may not drastically alter task-oriented user experiences. This observation paves the way for further explorations into how sound can enhance interactive experiences differently. Lucero et al. [77] explore the potential of playful interactive models to bolster engagement, underscoring the often-overlooked importance of playfulness in enhancing the quality of both entertainment and utility-driven interactive systems. Our research contributes to this discourse by exploring how music and images, autonomously generated, can coalesce to shape user perception, thereby fostering a deeper engagement in both artistic and practical applications, such as game development and immersive experience design.

3.1.1 VRMIs and IVMIs

Works focused on the development of Virtual Reality Musical Instruments (VRMIs) and Immersive Virtual Musical Instruments (IVMIs) have also been exploring different forms in which users interact with virtual musical instruments. For instance, Mäki-Patola et al. [80] introduce and analyze four gesture-controlled musical instruments

designed to allow for rapid experimentation of new interfaces and control mappings. Cabral et al. [13] present a novel 3D virtual instrument with a particular mapping of touchable virtual spheres to notes and chords of a given musical scale. The objects are metaphors for note keys organized in multiple lines, forming a playable spatial instrument where the player can perform certain sequences of notes and chords across the scale employing specific gestures, similar to the system we will present in Chapter 5, where we propose a ludic musical instrument using an undersea theme, with bubbles serving as metaphors for musical note arrangements, enabling users to intuitively engage with music creation. Mitchusson [89] uses the audio module and the Virtual Reality capabilities from the Unity engine to generate, through a dice-rolling mechanic, sample effects and audio mixing to generate experimental music outcomes. Gibson and Polfreman [48] present a framework that supports the development and evaluation of graphical interpolated mapping for sound design. The approach is capable of comparing the functionalities of previously developed systems leading to a better understanding of the outcomes these systems are capable of generating.

Innovative devices that work like digital musical instruments were also proposed. For instance, the world-renowned reacTable [64] is a novel multi-user electro-acoustic music instrument with a tabletop tangible user interface. It proposes new engaging ways in which musicians and sound designers can perform and compose music on the fly. Unlike many audio interfaces, the application does not use regular input devices, such as a mouse or a keyboard, and presents a more flexible way for individuals can interact with the installation.

The exploration of VRMIs and IVMIs reflects a broader trend towards more immersive and interactive forms of music creation, aligning with the objectives of this thesis to investigate meta-interactivity and its potential to enhance the music generation process. Systems like the reacTable and the 3D virtual instruments mentioned above embody principles of meta-interactivity by providing users with intuitive and engaging ways to influence musical compositions through direct interaction. This approach resonates with the Bubble Sounds system developed in this research, which aims to democratize music creation by offering an interactive environment where users, regardless of their musical training, can engage in and influence the music generation process. By integrating user interaction with sophisticated computational algorithms, these systems represent a step forward to creating more accessible and expressive musical tools. Thus, examining VRMIs and IVMIs provides valuable insights into how virtual environments can facilitate creative expression, serving as a foundation for the meta-interactive experience explored in this thesis.

3.2 Artificial Intelligence

A common approach is to use Deep Learning models to support the generation of music and new composition methods. E.g., Roberts et al. [103] enable the integration of generative models through the use of the Google Magenta Interface to arouse musician's creativity. This system was tested on a live jazz performance with a piano and drums duet. Ferreira et al. [43] explore how Deep Learning models can be employed for the composition of music with pre-established feelings, which can be extended for sentiment analysis of symbolic music as well. Similarly, AIVA (Artificial Intelligence Virtual Artist) [98] generates emotional music for media such as games, movies, and other endings. Also, Huang and Raymond [59] create coherent music with melodic and harmonic structures that can be acknowledged as pieces composed by a human agent. Castro [15] trained Deep Learning models to perform coherent improvisations. Agarwala et al. [1] use generative Recurrent Neural Networks to create musical sheets without predefining compositional rules to the models. Jukebox [28] generates music with singing tracks in the raw audio domain. MusicLM [2] proposes a model for generating high-fidelity music from text descriptions, working similarly as stable diffusion approaches for the generation of images.

Biot et al. [11] performed a detailed analysis of how deep learning can generate musical content. The authors offer a comprehensive presentation of the foundations of deep learning techniques for music generation as well as a presentation of a conceptual framework used to classify and analyze various types of architecture, encoding models, generation strategies, and ways for users to control the creation process. Works also evaluated the music creation methods of Machine Learning systems in public performance environments [113].

As for the dynamic generation of images, Dall-e [102] generates images from textual descriptions as input. Deepsing [96] proposes a learning method for performing an attributed-based music-to-image translation approach. Some of these works focus on conveying a specific feeling, however, in our case, we are interested in understanding what feelings such systems can convey to a general audience.

Rocksmith [119] presents a mode where it is possible to explore the cooperation between the user and an AI system through a conventional musical instrument, such as an electric guitar. The user can jam with the system, which provides a non-adaptive background track (i.e., it is the user who must adapt to the tempo of the song).

Rodriguez et al. [42] discuss different computational techniques related to Artificial Intelligence that have been used for algorithmic composition, including grammatical representations, probabilistic methods, neural networks, symbolic rule-based systems, constraint programming, and evolutionary algorithms. In their work's survey, they aimed to be a comprehensive account of research on algorithmic composition, presenting a thorough view of the field for researchers in Artificial Intelligence.

Differently from these works, however, in this thesis, we present an algorithm that can be employed for the generation of both music and images – thus addressing a complex problem in generative art – that is, we can efficiently manipulate music and image databases for coherent outcomes to emerge.

The integration of music and image generation within AI research, particularly through Machine Learning and Deep Learning techniques, represents a relevant area of exploration [34] because it pushes the boundaries of traditional creative processes, enabling the generation of complex, multimodal artistic expressions that reflect human-like understanding and creativity. For instance, Williams et al. [122] have developed a method for generating music that aligns emotionally with mindfulness practices. These studies resonate with the discussions in Chapter 6 of this thesis, which delves into autonomous methods for generating music and visuals that maintain coherence, showcasing how AI techniques can facilitate a symbiotic relationship between auditory and visual creative outputs.

3.2.1 Music Co-creation

Miguel Civit et al. [18] presents an in-depth view of AI's role in music generation, showcasing the growing interest and technological advancements in this field. It aligns with this thesis by highlighting the need for innovative human-machine collaboration approaches, which are explored through implicit cooperation, meta-interactivity, and autonomous creation, aiming to enrich the computational music generation landscape.

Yotam Mann's AI Duet [82] is a Google Magenta-based tool that enables a collaborative musical experience by allowing users to play a piano duet with an AI system. Users initiate the interaction by playing notes on their keyboard, to which AI Duet responds with complementary notes, creating a cohesive musical piece. The system has been trained on a vast array of MIDI files to understand musical structures and timings. This tool exemplifies how AI can enhance creative processes, a concept central to my thesis. This thesis presents systems with similar collaborative dynamics, exploring how such AI-enhanced interactions can lead to innovative musical creations, aligning with my focus on implicit cooperation and meta-interactivity in music generation.

Calliope [9] is a web application for co-creative multi-track music composition in the symbolic domain, enabling users to upload, visualize, edit, and generate MIDI tracks. This system supports interactive music creation, combining human input with automated generation to facilitate creative exploration and ideation. Its integration of algorithmic processes with user-driven composition aligns with this thesis, where we explore co-creative environments that leverage technology to enhance musical creativity, aiming to enrich the artistic process and foster new musical experiences.

3.2.2 Procedural Content Generation

Many procedural-generated audiovisual works are approached in the context of partially autonomous systems, such as games, where individuals participate in the emergence of the meaning by shaping it through their agency while the system works to maintain coherence. For instance, this is the case of Proteus [67], where the player can explore a procedural world that presents different elements at each run. In its procedurally generated world, fauna and flora emit their own musical signature, whose combination generates dynamic changes in the audio output. Mezzo [12] is a computer program that procedurally writes Romantic-Era style music in real-time to accompany computer games. Lopes et al. [74] investigate how designers can control and guide the automated generation of levels and their soundscapes by authoring the intended tension of a player traversing them.

Hoover et al. [56] introduce a novel approach based on evolutionary computation called functional scaffolding for musical composition (FSMC), which helps the user explore potential accompaniments for existing musical pieces or scaffolds. Similarly, Hoover and Stanley [54] propose a model based on the idea that the multiple threads of a song are temporal patterns that are functionally related, which means that one instrument's sequence is a function of another's. This idea is implemented in a program called NEAT Drummer [57] that interactively evolves a type of artificial neural network called a compositional pattern-producing network, representing the functional relationship between the instruments and drums. Scirea et al. [110] present the MetaCompose music generator, a compositional framework for affective music composition. In the context of this thesis "affective" refers to the music generator's ability to express emotional information.

While our work shares common ground with Kajihara et al.'s "Imaginary Soundscape" [66] in generating pseudo sound environments, it differs in both approach and focus. "Imaginary Soundscape" uses a machine learning model to create soundscapes based on visual cues, whereas our system employs a rule-based mechanism to facilitate music co-creation. Our approach subtly incorporates a human-in-the-loop design, offering a more interactive and less deterministic experience than purely machine-generated soundscapes. This distinction underscores our contribution to enhancing user engagement within generative music environments.

3.3 Computational Creativity

Other works also see the creative process for the generation of artworks as a collaboration between a human and an AI system [22, 94, 91]. Davis et al. [26] propose an experience where a user takes turns with a computer AI system when drawing on the same canvas. In Jacob and Magerko [63], a human and an agent

collaborate to produce movement-based performance pieces using a co-creative agent. These works, however, happen apart from the electronic games development efforts and, therefore, are not focused on providing friendly imagetic interfaces that guide non-experts in any kind of content generation process, demanding them to be properly trained for the desired outcome to emerge.

Procedural systems show potency for extending the life span of an interactive experience since they present new elements to an audience every time they are executed. In this sense, Hoover et al. [55] propose the use of Functional Scaffolding, a method that modularizes the structure of a MIDI song and autonomously generates a harmonized follow-up. This follow-up interprets the tile assets of the early stages of Super Mario Bros [90] as if they were a musical score, fitting them into a world matrix. These systems, however, require humans to explicitly join the music generation process, also demanding some prior knowledge of music theory. Carnovalini and Rodà [14] perform a survey in an attempt to give a complete introduction to those who wish to explore Computational Creativity and Music Generation. To do so, the authors first give a glimpse of the research on the definition and the evaluation of creativity, both human and computational, needed to understand how computational means can be used to obtain creative behaviors and its importance within Artificial Intelligence studies.

Pasquier et al. [95] introduce the concept of Musical Metacreation (MuMe), a subfield of computational creativity that focuses on endowing machines with the ability to achieve creative musical tasks. It covers all dimensions of the theory and practice of computational generative music systems, ranging from purely artistic approaches to purely scientific ones, inclusive of discourses relevant to this topic from the humanities. Tatar and Pasquier [117] discuss artificial agents that tackle musical creative tasks, partially or completely. The authors examine the evaluation methodologies of musical agents and propose possible future steps while mentioning ongoing discussions in the field.

Works also focused on the analysis of affection in music. For instance, Williams et al. [121] mention that there has been a significant amount of work implementing systems for algorithmic composition to target specific emotional responses in the listener, however, a full review of this work's outcomes is not currently available. This gap creates a shared obstacle for those entering the field. Lopes et al. [76, 75] explore how an autonomous computational designer can create frames of tension that guide the procedural creation of levels and their soundscapes in a digital horror game.

3.4 Summary

This chapter has provided a comprehensive overview of the existing literature across the domains of Human-Computer Interaction (HCI), Artificial Intelligence (AI),

Creative Computing, Procedural Content Generation (PCG), and the Arts. While many of the discussed works explore a broad range of strategies, from human-computer interaction to AI-human collaboration, our research focuses on the audio-visual relationship within the context of implicit cooperation, meta-interactivity, and both partially and fully autonomous approaches.

Our study aims to deepen the understanding of how these methods can foster innovative and emotionally resonant artistic outputs, particularly in music composition. We also explore the potential for a coherent musical corpus to emerge without explicit human involvement in the interactive dialogue. Moreover, this investigation delves into how visuals impact human perception and appreciation of music, an area that has received limited attention in prior research, especially within the context of game experiences.

Chapter 4

Implicit Cooperation

"Works of art make rules; rules do not make works of art."

— Claude Debussy

This chapter explores a concept called implicit cooperation, where agents collaborate unintentionally. More specifically, it will be examined the effects of this concept in light of Art Games, where a novel algorithm will be presented. This algorithm enables a human agent to engage in emergent cooperation with a video game system, contributing to the creation of music without explicit awareness of the collaborative process. An experiment involving human subjects was conducted, where they interacted with a game that integrates piano keys within its virtual environment. Players, unaware of their cooperative role, shape the game's soundtrack through interaction with these keys, generating harmonious compositions.

4.1 Introduction

Cooperative agents can accomplish hard tasks through their joint work. However, most systems assume that agents explicitly collaborate, by having a joint goal, a utility function that fosters collaboration, or even pre-specified coordination rules. In many situations, however, we may have a system where the actions of an agent unintentionally help another. In particular, we may be able to use the actions of an agent to produce works of art in an emergent fashion, without requiring artistic knowledge from the agent, or an explicit intention to create an artistic piece.

Past works view the creative process as a collaboration between a human and an AI system [22]. Pachet et al. [94] presents a system where a human musician plays a music sample, and an AI system, after learning the basic music pattern, joins the

musician in producing music. Hence, both humans and the system "jam" together, creating a unique music that neither would construct alone. Moreira et al. [91] shows a set of agents that react to human musicians, and humans and agents cooperate in producing a live music performance. However, in all these works the user has to explicitly collaborate with the AI system in the music generation process, even requiring a musical background for the system to work well.

In this chapter, it will be presented a new algorithm where the actions of a user are used to dynamically emerge a musical piece. This system may be used in the context of Art Games, a genre that views games as artistic experiences rather than just entertainment. Our algorithm places invisible musical cells on the floor of a virtual scenario, which are arranged in a way that fosters the musical production. It was developed an Art Game based on the algorithm that is going to be presented, which also served as an artifact to evaluate the implicit cooperation approach with real human players. This study shows that real humans, without realizing the effect of their actions, effectively generate a large number of arpeggios, and classify the product of the system as "music".

The key findings of the conducted user studies are:

- The study demonstrates that Implicit Cooperation effectively merges the creative capabilities of humans and machines. By leveraging the actions of human agents within a structured environment, the system facilitates the generation of music that resembles human-composed pieces, challenging the conventional boundaries between human and machine-generated art.

- Compared to random movements or random note selection, a human agent navigating the specified grids is more likely to produce coherent musical sequences. This finding underscores the human agent's intuitive interaction with the system, leading to a more structured and melodious output than could be achieved through random processes alone.

These findings have implications for the Human-Computer Interaction (HCI) and game development communities by shedding light on how users perceive and interact with interactive systems. They offer a fresh perspective on collaboration between humans and machines and open exciting avenues for enhancing interactive experiences. In the forthcoming sections, we will explore the implications of the implicit cooperation approach within a simple exploration game, emphasizing its significance for game developers in creating mood-enhancing music that aligns seamlessly with the game experience.

4.2 Implicit Cooperation – Overview

Implicit cooperation consists of a multi-agent system where agents collaborate without the *intention* of doing so. That is, while an agent is pursuing its own objectives, its actions "end up" aiding another agent.

In this thesis, it will be studied a restricted version of implicit cooperation, where the focus will be on systems with only two agents. The games developed for studying this approach are in the context of human-computer interaction, and hence, we will consider the following two agents:

(i) A human-controlled agent integrated into a game environment (i.e., a game character), pursuing objectives directly related to the game's proposal.

(ii) A computational system (e.g., game mechanics), which uses the actions of the player to accomplish another objective.

Therefore, it will be designed systems that will accomplish their objectives with the help of a user, however without requiring from him/her an explicit intention of collaborating with the system.

That is, given an agent ϕ, which receives a reward r_a for each action a, according to the current world state. Let's assume that ϕ wants to maximize its total reward (for instance, explore the world as much as possible, or collect items in a virtual environment). Given now a system S, with an objective O (for instance, generate music with a certain characteristic). The implicit cooperation problem under study in this paper is: how can the system S induce agent ϕ to accomplish objective O?

In this chapter, we take the first step towards addressing emergent music generation. In the following section, we will introduce an algorithm that enables system S to modify the game environment's floor dynamically. This adjustment facilitates music production based on the movements of agent ϕ within the digital game scenario.

4.2.1 Emergent Music Generation

In this section, it will be presented a system where the movement of an agent produces music in an emergent fashion. First, it will be provided some concepts that help illuminate the discussion that follows.

A key element that served as a guideline for the system that will be presented is *repetition*. Repetition is a strong factor for musicalization because it "breaks" a song into pieces and seams them together forming new patterns in a way to preserve an initial structure, making it easier for our brains, an avid "devourer of patterns" [70] to easily assimilate it and recognize it as music. According to Elizabeth Hellmuth

C MAJOR SCALE						
C (1) - D (2) - E (3) - F (4) - G (5) - A (6) - B (7)						
TRIAD CHORDS OF THE C MAJOR SCALE						
C (C - E - G)	Dm (D - F - A)	Em (E - G - B)	F (F - A - C)	G (G - B - D)	Am (A - C - E)	BØ (B - D - F)
EXAMPLES OF INVERSION ON THE C MAJOR CHORD						
(C - E - G)	(C - G - E)	(E - C - G)	(E - G - C)	(G - C - E)	(G - E - C)	

Figure 4.1: The C Major Scale, all possible triads of thirds, and an example of all possible inversions for the triad CEG.

[83], if we are asked whether a particular piece is music or not, a remarkably large part of the answer appears to be: "I know it when I hear it again." She also stated that repetition serves as a "handprint" of human intent, and a phrase that might have sounded arbitrary at first may sound reasonable the second time it is heard.

Other important concepts from music theory that will be explored in this chapter are **chords** and **arpeggios**, that was previously defined in Chapter 2, subsection 1.1. Chords are any harmonic set of three or more notes that are heard resonating simultaneously [68]. Arpeggios are the successive execution of the notes of a chord (in any order) [100]. In the system that will be presented, we have only one note being played at a time. Hence, we focus on the presence of arpeggios rather than chords. Figure 4.1 shows an example of possible arpeggios of triads, and all possible orders for the CEG case.

We consider our agent as a character in a scenario, controlled by a human player. This agent pursues some objective: for instance, collecting items or exploring the game world. The actual objective depends on the system designed using the technique, and does not affect the presented approach.

Musical cells are placed on the floor of the game environment: this environment, in turn, is divided into a grid, where each cell corresponds to a piano key. When the agent (i.e. the character controlled by the player) steps on a cell, the corresponding key plays. The grid may be invisible to the agent, and it may or may not be aware of this construction. We also consider that the agent can jump on the same place, and that would replay the same note. When placing the grid, we use a "building block", which is concatenated in all directions to cover the full scenario. This can also be seen as if the block is a torus: upon going right in the last column, the agent will reach the first column; upon going down in the last row, the agent will reach the first row. We show in Table 4.1 (a) one 3x3 block, and in Table 4.1 (b) how it would cover a 6x6 scenario.

The blocks are generated in a way that when the agent moves towards the south, it follows a sequence of thirds, and thus creates arpeggios. Similarly, if the agent moves towards the east, it follows a sequence of fifths, also creating arpeggios. For example, in the case of 7 keys, we can use the block shown in Figure 4.2 (where the colors help visualize different keys). These blocks can be generated as follows. Let

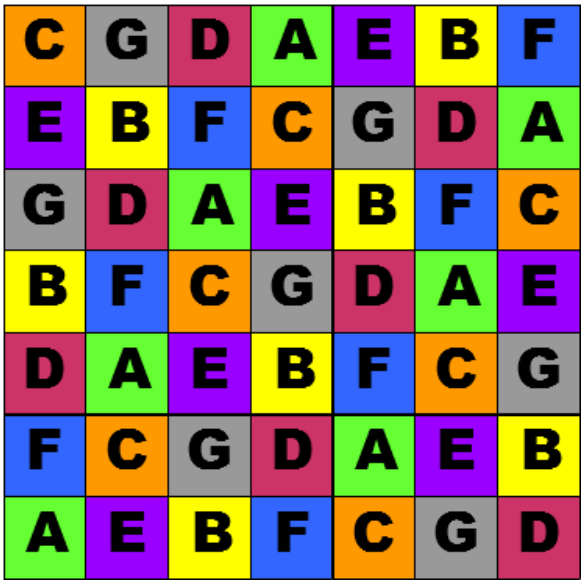


Figure 4.2: 7x7 building block.

Table 4.1: Example of a 3x3 block covering a 6x6 scenario.

A	B	C
C	A	B
B	C	A

(a) 3x3 Block

A	B	C	A	B	C
C	A	B	C	A	B
B	C	A	B	C	A
A	B	C	A	B	C
C	A	B	C	A	B
B	C	A	B	C	A

(b) 6x6 Scenario

$\mathbf{M} = \{m_1, ..., m_n\}$ be a set of notes, and B an $n \times n$ matrix. We generate our proposed block by the Algorithm 1. We start from the upper left corner, and fill in each cell of the first row in a progression of fifths (i.e., skip the next 3 elements of the set). Then, we fill all columns in a progression of thirds (i.e., skip the next element of the set).

Therefore, Figure 4.2 shows the case where $\mathbf{M} = \{C, D, E, F, G, A, B\}$. Note that in this example the start point is the C note, following the usual musical scale, but different starting notes could be used. Also, when moving north the agent will play a decreasing sequence of thirds, and likewise when moving west a decreasing sequence of fifths. As a consequence, for $|\mathbf{M}| = 7$ we also have that: (i) When moving northeast or south, the agent plays a sequence of thirds; (ii) When moving north or southwest, the agent plays a sequence of sixths; (iii) When moving west, the agent plays a sequence of fourths; (iv) When moving east, the agent plays a sequence of fifths; (v) When

Algorithm 1 Block generation algorithm.

```
1: procedure BlockGeneration
2:     B[1, 1] := 1
3:     for c := 1 ... n − 1 do
4:         B[1, c + 1] := B[1, c] mod n
5:     end for
6:     for c := 1 ... n do
7:         for l := 1 ... n − 1 do
8:             B[l + 1, c] := B[l, c] mod n
9:         end for
10:    end for
11: end procedure
```

moving southeast, the agent moves in a sequence of sevenths; (vi) When moving northwest, the agent moves one tone up. This shows that even though we emphasize the generation of thirds/fifths, the agent can still generate a great variety of notes from its current position, increasing the diversity of the musical production (in fact, for $|\mathbf{M}| = 7$, we can generate any possible note from a given position).

Additionally, in our analysis, we will also consider sets of notes of size different than 7. That could represent, for instance, notes of the next octave; or even a non-traditional division of a given frequency range in n different notes.

4.3 Analysis

We will start by analyzing the correctness of Algorithm 1. It is clear that there exists a bijective function that maps the set $\mathbf{M}$ to $\mathbb{Z}_n$. Also, the "third" of a note m_i is equivalent to the note $m_{(i+2) \bmod n}$. In a general way, a note m_i changes in a sequence of k-th to $m_{(i+k-1) \bmod n}$. Therefore, we can consider the cyclic group $(\mathbb{Z}_n, \oplus)$, with $\oplus$ representing addition modulo n, as an isomorphism to set $\mathbf{M}$ under operation of changing in a sequence of k-th.

The following theorem from Ledermann [71] will be useful to prove Algorithm 1 correctness:

Theorem 1. *Let $(G, *)$ be a cyclic group, where $|G| = n$, and $a^k = a * a * \cdots * a$ (k times). If $a \in G$ is a generator of G and k is relatively prime to n, then a^k is also a generator of G.*

Hence, considering the group $(\mathbb{Z}_n, \oplus)$, 1 is its generator. Also, we have for every integer $k > 0$ that $1^k = k \bmod n$. So, every $0 < k < n$ relatively prime to n is also a generator.

The following observation states that generating a progression of thirds and fifths in south and east direction respectively allows the agent to move as enumerated above. Consider below that an integer $k > n$ is the same as $k \bmod n$ and for any $a \in \mathbb{Z}_n$, its inverse is $-a = n \ominus a = n - a \bmod n$. Let $b = B[i, j]$.

Observation 1. *For a given element b in a matrix B, if $B[i, j+1] = b \oplus 4$ and $B[i+1, j] = b \oplus 2$, then the surrounding elements can be described in terms of b with specific offsets. Specifically, the element diagonally above and to the left, $B[i-1, j-1]$, equals $b \oplus -6$, the one directly above, $B[i-1, j]$, equals $b \oplus -2$, and the one diagonally above and to the right, $B[i-1, j+1]$, equals $b \oplus 2$. Similarly, to the left, $B[i, j-1]$ equals $b \oplus -4$, diagonally below and to the left, $B[i+1, j-1]$, equals $b \oplus -2$, and diagonally below and to the right, $B[i+1, j+1]$, equals $b \oplus 6$.*

Given that it is valid for every integer i, j, and recognizing that $B[i, j-1] = b \oplus -4$ and $B[i-1, j] = b \oplus -2$ are trivially true, we can deduce the following relationships for the matrix B:

$$\begin{aligned}
B[i-1, j-1] &= B[i-1, j] \oplus -4 \\
&= b \oplus -2 \oplus -4 \\
&= b \oplus -6, \\
B[i-1, j+1] &= B[i-1, j] \oplus 4 \\
&= b \oplus -2 \oplus 4 \\
&= b \oplus 2, \\
B[i+1, j-1] &= B[i+1, j] \oplus -4 \\
&= b \oplus 2 \oplus -4 \\
&= b \oplus -2, \\
B[i+1, j+1] &= B[i+1, j] \oplus 4 \\
&= b \oplus 2 \oplus 4 \\
&= b \oplus 6.
\end{aligned}$$

Table 4.2 shows Observation 1 applied to moves in a set $|\mathbf{M}| = 7$. Note that positive relations ($B[i-1, j+1]$, $B[i+1, j+1]$, $B[i+1, j]$, $B[i, j+1]$) will remain as in Table 4.2 for any set size n, while negative relations will change according to the calculation of the inverse $n - a \bmod n$. Also, we will use the following lemma:

Lemma 1. *If for every integer $i \in \{1, \ldots, n-1\}$ and $j \in \{1, \ldots, n\}$, $B[i+1, j] = B[i, j] \oplus 2$ and for all $j \in \{1, \ldots, n-1\}$, $B[1, j+1] = B[1, j] \oplus 4$, then: $\forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, n-1\}$, $B[i, j+1] = B[i, j] \oplus 4$.*

$\oplus 1$ B[i-1,j-1]	$\oplus 5$ B[i-1,j]	$\oplus 2$ B[i-1,j+1]
$\oplus 3$ B[i,j-1]	$\oplus 0$ B[i,j]	$\oplus 4$ B[i,j+1]
$\oplus 5$ B[i+1,j-1]	$\oplus 2$ B[i+1,j]	$\oplus 6$ B[i+1,j+1]

Table 4.2: Neighborhood of a cell as stated in Observation 1 for $|\mathbf{M}| = 7$.

Proof. We use induction on i. **Base case**: The hypothesis is given for $i = 1$. We start with $i = 2$. Thus, for every $j \leq n-1$, $B[i, j+1] = B[2, j+1] = B[1, j+1] \oplus 2 = B[1, j] \oplus 4 \oplus 2 = B[2, j] \oplus 4 = B[i, j] \oplus 4$. **Induction step**: Assume as induction hypothesis that $\forall i \in \{1, \ldots, n-1\}, \forall j \in \{1, \ldots, n-1\} : B[i, j+1] = B[i, j] \oplus 4$. Therefore, for $i = n$, we have: $B[n, j+1] = B[n-1, j+1] \oplus 2 = B[n-1, j] \oplus 4 \oplus 2 = B[n, j] \oplus 4$. □

Now we can show the correctness of Algorithm 1:

Theorem 2. *Algorithm 1 generates blocks so that the agent movement plays notes in the proposed way.*

Proof. By Observation 1, we only need to prove that Algorithm 1 generates blocks such that $B[i, j+1] = B[i, j] \oplus 4$ and $B[i+1, j] = B[i, j] \oplus 2$, for every integer i, j. At the end of line 5, we have the following postcondition:

$$B[1, 1] = 1 \wedge \forall j \in \{1, \ldots, n-1\} : B[1, j+1] = B[1, j] \oplus 4.$$

We need to show that the second for loop has the following postcondition:

$$\forall j \in \{1, \ldots, n\} : \forall i \in \{1, \ldots, n-1\} : B[i+1, j] = B[i, j] \oplus 2.$$

This is done by showing a postcondition for the innermost for loop, and then the postcondition above. At the innermost for (lines 7-9), we have the following precondition:

$$1 \leq c \leq n+1 \wedge \forall i \in \{1, \ldots, n-1\} : \forall j \in \{1, \ldots, c-1\} : B[i+1, j] = B[i, j] \oplus 2,$$

and we state the following loop invariant in the innermost for:

$$1 \leq c \leq n \wedge 1 \leq l \leq n \wedge \forall i \in \{1, \ldots, n-1\} : \forall j \in \{1, \ldots, c-1\} : B[i+1, j] = B[i, j] \oplus 2 \\ \wedge \forall i \in \{1, \ldots, l-1\} : B[i+1, c] = B[i, c] \oplus 2.$$

Initialization: Until the comparison $l \leq n-1$ at line 7, this is trivially true. **Maintenance**: At line 8, $B[i+1, c]$ becomes $B[i, c] \oplus 2$. Then, for every line i

from 1 to l, $B[i+1,c] = B[i,c] \oplus 2$. After increment of l, loop invariant is maintained. **Termination**: All lines $i < n$ in column c obeys $B[i+1,c] = B[i,c] \oplus 2$. Thus, we have as postcondition of innermost for:

$$1 \leq c \leq n \wedge \forall i \in \{1,\ldots,n-1\} : \forall j \in \{1,\ldots,c\} : B[i+1,j] = B[i,j] \oplus 2.$$

Now, for the outermost loop (lines 6-10), we state the following loop invariant:

$$1 \leq c \leq n+1 \wedge \forall i \in \{1,\ldots,n-1\} : \forall j \in \{1,\ldots,c-1\} : B[i+1,j] = B[i,j] \oplus 2.$$

Initialization: After initialization of c, loop invariant is trivially true. **Maintenance**: Loop invariant is precondition of innermost for, hence, before increment, for every line $i < n$ in column c, $B[i+1,c] = B[i,c]$. After increment, invariant is maintained. **Termination**: When c becomes $n+1$, loop invariant becomes:

$$\forall j \in \{1,\ldots,n\} : \forall i \in \{1,\ldots,n-1\} : B[i+1,j] = B[i,j] \oplus 2.$$

This proposition together with the postcondition of the first for (line 5), gives us:

$$\begin{aligned} B[1,1] = 1 \wedge \forall j \in \{1,\ldots,n\} : & \forall i \in \{1,\ldots,n-1\} : \\ & B[i+1,j] = B[i,j] \oplus 2 \\ & \wedge \forall j \in \{1,\ldots,n-1\} : \\ & B[1,j+1] = B[1,j] \oplus 4. \end{aligned}$$

Hence, by Lemma 1, we have the result. □

In the following Corollary, we show that the blocks will always by cyclic, i.e., will allow the agent to navigate as exemplified in Table 4.1.

Corollary 2.1. *Algorithm 1 generates cyclic grids for any set* $\mathbf{M}$*, and* $B[1,1]$ *set initially with any* $m \in \mathbf{M}$*.*

Proof. Algorithm 1 generates the same pattern of notes for every cell with $i,j > n$, because for every i,j :

$$B[i+n,j] = B[i+n-1,j] \oplus 4 = B[i+n-2,j] \oplus 4 \oplus 4 = \ldots = B[i,j] \oplus \bigoplus_{k=1}^{n} 4 = B[i,j],$$

and

$$B[i,j+n] = B[i,j+n-1] \oplus 2 = B[i,j+n-2] \oplus 2 \oplus 2 = \ldots = B[i,j] \oplus \bigoplus_{k=1}^{n} 2 = B[i,j].$$

□

Proposition 1. *Algorithm 1 generates complete blocks for any set* $\mathbf{M}$ *with* $|\mathbf{M}|$ *not divisible by 2 or 4 and* $B[1,1]$ *set initially with any* $m \in \mathbf{M}$*.*

Proof. By Theorem 1, we have that every $0 < k < n$ relatively prime to n is a generator of the group $(\mathbb{Z}_n, \oplus)$. Hence, for $|\mathbf{M}|$ not divisible by 2 or 4, we have that both 2 and 4 will be generators. Therefore, all rows and columns in B will have all elements in $\mathbf{M}$. □

Let's assume now random walks in our proposed blocks. We will consider two different kinds of random walks:

- From a given cell, uniform probability to move to any neighboring cell.
- Greater probability of moving in straight and lateral directions (i.e., "human"-like movement).

We will focus our analysis now in blocks where $|\mathbf{M}| = 7$.

Let's assume now random walks in our proposed blocks. We will consider two different kinds of random walks: (i) From a given cell, uniform probability to move to any neighboring cell; (ii) Greater probability of moving in straight and lateral directions (i.e., "human"-like movement). We will focus our analysis now in blocks where $|\mathbf{M}| = 7$.

In the analysis below, we define music as a repetition of a sequence of notes containing a sequence of thirds or fifths, with other notes between repetitions. In other words:

Definition 1. *Let* $N, M \in \mathbb{N}$*, and* $M \geq N$*. A sequence of notes* $\{a_i\}_{i\in\{1,\ldots,M\}} \in \mathbb{Z}_n$ *is* music *if there is a sequence of notes* $A = \langle a_1, a_2, \ldots, a_N\rangle$*, such that* $\forall i > 1 : a_{i+1} - a_i \in \{2, -2, 4, -4\}$*, and* $\{a_i\}_{i\in\{1,\ldots,M\}} = \langle A, B_1, A, B_2, \ldots, A, B_K\rangle$*, for a given* $K \in \mathbb{N}$*; and* B_i *are any sequence of notes of any size, even size zero.*

Proposition 2. *Random walks in blocks generated by Algorithm 1 have a higher probability of generating music than randomly selecting notes.*

Proof. Clearly, sequences of type $\{a_i\}_{i\in\{1,\ldots,M\}} = \langle A, B_1, A, B_2, \ldots, A, B_K\rangle$, will be generated with higher probability as the probability of generating a sequence $A = \langle a_1, a_2, \ldots, a_N\rangle$, gets higher. Hence, we focus on studying the probability of generating a sequence A. Given a sequence A of size N, the first note can be any from $\mathbf{M}$, so there are seven possibles outcomes with 1/7 probability. From our definition of music, for $i > 1$, the i-th note must be any of four possibles notes among seven from $\mathbf{M}$. Thus, the probability for generating music from this sequence randomly is $P_{r.p.}(A) = 7\frac{1}{7}\prod_{i=2}^{N}\frac{4}{7} = (\frac{4}{7})^{N-1}$, assuming uniform distribution for drawing notes.

Algorithm 1 generates 8 neighbors and the agent can repeat the same note when jumping in the same cell. Hence, random walks in the neighborhood of a cell at

every iteration has 9 possible notes and 2 repeated notes for a third above, 2 repeated notes for a third below, and one cell for a fifth above, and another cell for fifth below (see number of cells for +2, +5, +4, and +3 respectively in Table 4.2). Assuming uniform probability to move to any of these nine blocks, and that the first note is chosen randomly, the random walk probability of a sequence $A = \{a_1, \ldots, a_N\}$ is $P_{r.w.}(A) = 7p(a_1)\prod_{i=2}^{N} p(a_i|a_{i-1}) = 7\frac{1}{7}\prod_{i=2}^{N} p(a_i|a_{i-1}) = (\frac{6}{9})^{N-1}$, since, for $i > 1$: $p(a_i|a_{i-1}) = \frac{6}{9}$, if $a_i \ominus a_{i-1} \in \{2, -2, 4, -4\}$; $\frac{1}{9}$, otherwise. Hence, whatever note chosen initially, a random walk has probability $\frac{6}{9}$ at each step for generating music, because there are six directions that contribute for generating a sequence A: north, east, west, south, southwest and northeast (Table 4.2). Against $\frac{4}{7}$ probability when randomly drawing notes, it is more probable for random walking in our proposed blocks to generate music. □

Additionally, we assume that when humans are playing a game, it is more likely that they move in horizontal and vertical directions (north, south, east, west) than diagonals. For instance, there are no diagonals keys in computer keyboards, which would make these movements less likely. Therefore:

Proposition 3. *Humans have a higher probability of generating music than random walks, when moving in blocks generated by Algorithm 1.*

Proof. By the assumption above, human agents move according to the following probability: $p(move) = p + \epsilon, move \in \{\text{north, south, east, west}\}; p, \text{otherwise.}$ Additionally, we have that $\epsilon > 0, p > 0$, and:

$$9p + 4\epsilon = 1. \tag{4.1}$$

Thus, a human has, at each step, a probability of $4 \times (p + \epsilon) + 2 \times p$ to generate music. As observed in Proposition 2, random walking has probability $\frac{6}{9}$. We must have:

$$4 \times (p + \epsilon) + 2 \times p > \frac{6}{9}, \tag{4.2}$$

for human moves to be more probable to generate music. The line segment of Equation 4.1, restricted to $\epsilon > 0$ and $p > 0$, is always inside the region determined by Equation 4.2. Hence, any value of p and ϵ greater than zero that satisfies Equation 4.1, also satisfies Equation 4.2, completing the proof. □

4.4 Microbial Art

Microbial Art is an interactive experience that blends music generation with simple gameplay mechanics. In this game, players control a microbial creature to collect

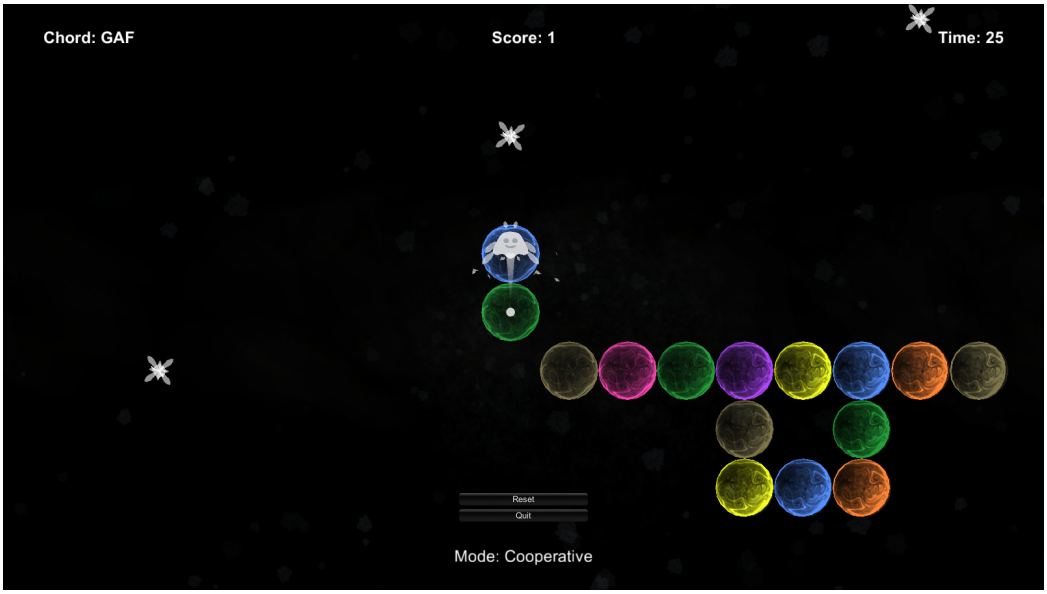


Figure 4.3: Microbial Art screenshot.

proteins to feed its body, while the environment responds with music, feeding the player's mind and creativity.

The underlying environment is a 3D-colored grid, where musical notes are placed on the cells. A musical note is assigned to each cell as the player moves. The note-generation scheme obeys a mathematical procedure that favors the generation of music, as discussed in the previous section. In other words, when the player moves, they will likely generate pleasant arpeggios. For comparison, the system also allows for two alternative procedures: one where notes are drawn uniformly random as the player moves, and another where a larger weight is placed in the random procedure to generate notes that might lead to arpeggios.

The player perceives only the cells, depicted as colored bubbles, within the environment as they move (Figure 4.3) and is unaware of the underlying notes. It is possible that the player may not even realize they are contributing to the music generation process. Hence, in this exploration of the implicit collaboration approach, the player's objective is to collect proteins for their microbial, but the grid generation scheme causes them to inadvertently generate music as a "side-effect" of their movement, resulting in an enjoyable experience.

To enhance the generation of interesting songs, beyond the grid of notes procedure presented in the previous section, three additional features are implemented: (i) The microbial movement leaves a trail, increasing with each collected item and fading over time, motivating continuous player movement. (ii) Upon collecting an item, the system adds, for a limited time, a corresponding musical element to the harmony,

allowing the creation of more complex musical pieces. (iii) After the generation of every three notes, they are played simultaneously, providing players with the experience of chords instead of just arpeggios and enhancing the sense of rhythm in the interaction between the user and the system.

While players may notice they are generating music as they move, given the association of environment colors with musical notes, it would be interesting to observe whether they forget their primary goal of collecting proteins, and finding enjoyment in the music generation process by revisiting already explored cells. This game prototype represents a more refined iteration of the implicit cooperation approach and was created to evaluate its efficiency, as will be detailed in the User Study in Section 4.5. For interested readers, a video demonstration showcasing the implicit cooperation mechanic and Microbial Art is available at `https://youtu.be/ZaOmJEPC-ZI`. Figure 4.4 presents a timeline of screenshots illustrating the demonstration of the Microbial Art system as showcased in the accompanying video.

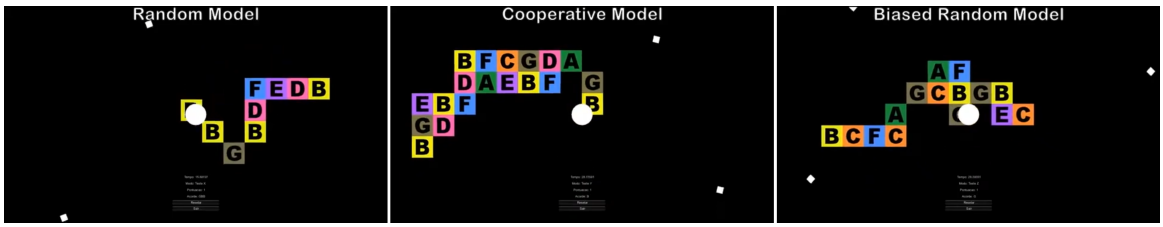


Figure 4.4: Screenshots of the Microbial Art video.

4.5 User Study

The implicit cooperation was evaluated in experiments with human players. For comparison, it was analyzed 3 different systems: (i) *Random:* Every time the agent steps in a cell, a note drawn uniformly randomly from is played (each note is selected uniformly, with all notes having an equal chance of being chosen, independent of prior selections); (ii) *Biased Random:* Similar to *Random*, but notes that are the third or the fifth of the note that was played previously are drawn with 70% probability (equally distributed), while all other notes are drawn with 30% probability (equally distributed); (iii) *Cooperative:* Follows our implicit cooperation scheme described in the previous section. Hence, in *Random* notes are drawn arbitrarily; while *Biased Random* still draws notes randomly, but following the basic principle from music theory that thirds and fifths should appear with higher likelihood, forming arpeggios.

For this assessment, we employed the Microbial Art game introduced in Section 4.4. In this game, users control a character that can freely navigate an environment and collect various objects strategically placed to encourage exploration. We randomly

selected three cells that are currently visible and filled them with objects. This selection process is repeated each time new cells become visible due to user movement. The game mechanics include displaying a score based on the number of items collected, contributing to a more immersive and game-like experience. This scoring system not only motivates users to explore the environment but also adds a layer of challenge and accomplishment to the overall gaming interaction.

The assessment sessions had 10 human evaluators, and each one played all 3 systems. We randomized the order in which each user played each system to avoid ordering issues. Additionally, the evaluators had no prior experience with the game and did not know how our system works, nor which one of the 3 systems they were currently playing (the 3 variations were presented to them as X, Y, and Z). Each variation was played for 180 seconds, and after that, they had to fill in a form about their experience, as presented in Table 4.3.

The pool of evaluators primarily consisted of undergraduate students majoring in game design from FUMEC University, in Brazil. To ensure the evaluation process adhered to time constraints determined prior to the evaluation sessions to take place, pilot tests were conducted to determine an appropriate duration for the game sessions. Based on these tests, a 10-minute duration was deemed sufficient to effectively demonstrate the systems and allow players to explore the game environments. None of the evaluators had prior experience with the systems before the evaluation sessions, and no sensitive user data was recorded. This study was approved by the ethics committee of the School of Computer and Communication at Lancaster University.

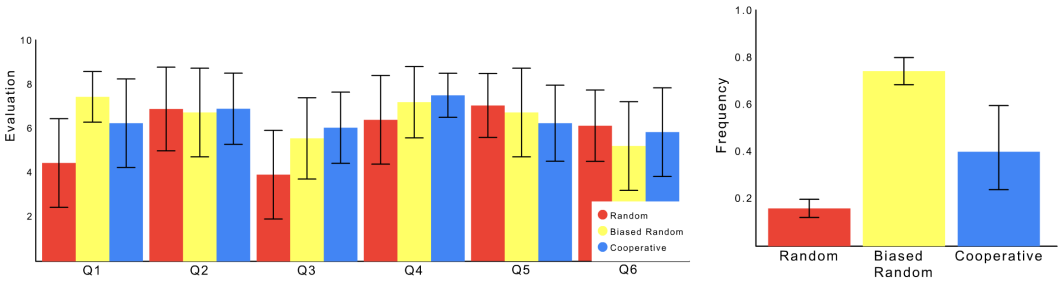
We queried the users the following questions, shown in Table 4.3.

Q1.	From 1 to 10, how do you classify the audio of the system? Very uninteresting (1) to (10) Very interesting
Q2.	How do you classify the relation between your actions and the audio of the system? Very uninteresting (1) to (10) Very interesting
Q3.	How do you classify the motivation to “compose” a song while playing? Very uninteresting (1) to (10) Very interesting
Q4.	Would you classify the sound output of the system as “music”? Very uninteresting (1) to (10) Very interesting
Q5.	How do you classify the audio experience provided by the system? Very uninteresting (1) to (10) Very interesting
Q6.	How do you classify your experience with the system as a whole? Very uninteresting (1) to (10) Very interesting

Table 4.3: Implicit Cooperation assessment questionnaire.

4.5.1 Results

The observed results are shown in Figure 4.5 (a). We can observe that humans could perceive that *Random* produced more arbitrary sounds, while the sounds produced by *Biased Random* and *Cooperative* were considered more interesting (Q1). We also noticed that users were not able to distinguish the importance of their actions in the audio generation process across the three systems (Q2). Additionally, when queried to assume that there is a relation, they seem to consider feeling a stronger motivation to generate music in *Cooperative*, even though they did not perceive that their actions had a greater effect in *Cooperative* (Q3). We also notice that the audio of *Cooperative* had the greatest tendency to be classified as "music", with *Biased Random* close behind. *Cooperative* also had the lowest variance in this aspect, indicating that users were more likely to agree in classifying the system as producing "music" than in the other systems (Q4). Interestingly, although users tended to agree more that *Cooperative* generates music, they also tended to perceive it as more "disturbing" than in the other systems (Q5). Finally, in terms of feeling engaged with the system, both *Random* and *Cooperative* had similar results, with *Biased Random* right behind (Q6).



(a) Survey comparing the different systems. Error bars show the 90% confidence interval.

(b) Frequency of triads of thirds. Error bars show SD.

Figure 4.5: Results of the experiment with real users.

For statistical analysis, we employed non-parametric statistical methods for analyzing the user responses, given the ordinal data and non-normal distribution of responses. The Mann-Whitney U test was employed to compare perceptions between pairs of systems, whereas the Kruskal-Wallis test was used for multi-group comparisons. Our findings indicated that users could discern musical quality differences between systems. Specifically, the *Cooperative* system was perceived more favorably than the *Random* system, with the Kruskal-Wallis test yielding a $p < 0.3$. Although this does not reach traditional significance levels ($p < 0.05$), it suggests a trend where the *Cooperative* output is more likely to be classified as "music".

In assessing engagement and interest in the sound produced (Q1), the Mann-Whitney U test showed a $p < 0.24$, pointing towards a preference for *Cooperative* over *Random*, although not yet at a significant level. The high p-value ($p < 0.9$) in evaluating user impact awareness on sound generation (Q2) highlighted the concept of "implicit cooperation," where the influence of user interaction is not overtly recognized. Regarding the motivation to generate music (Q3), a p-value of $p < 0.11$ was observed, indicating a stronger inclination towards music creation with the *Cooperative* system. Despite the general preference for the *Cooperative* system, it was paradoxically perceived as more "disturbing", aligning with contemporary art's potential to elicit a broad spectrum of emotional responses. We do not see that as a negative result, as art does not necessarily have to be pleasant; providing a disturbing experience is also one of the main objectives of contemporary arts [51]. Figure 4.5 (b) showcases the audio analysis, revealing the essential role of human interaction in arpeggio formation within the *Cooperative* system.

Although *Biased Random* recorded a higher arpeggio frequency, it did not necessarily correlate with a higher musical classification compared to the *Cooperative* system. We also analyzed the audio produced. In Figure 4.5 (b) we show the frequency of occurrences of triads of thirds (in any order) across 10 executions. These preliminary results suggest a nuanced understanding of the systems' capabilities and user experiences, warranting further exploration with a larger participant pool to confirm these tendencies. Notably, the significance of arpeggio presence in *Cooperative* underscores the intricate relationship between user interaction and perceived musical quality.

Random Walk refers to walking in our blocks with uniform probability to any direction (including jumping in the same cell), while *Cooperative* is the data with 10 real human users. Hence, as we can see, the presence of a human agent is essential in our system for the formation of arpeggios (which increase the sound quality), even though the user is not aware of how our system works, and is not actively trying to generate those structures. Additionally, even though *Biased Random* has a higher frequency of arpeggios, it did not have a higher tendency to be classified as music than our proposed system.

Note that it is not our objective to overpass *Biased Random* in terms of frequency of arpeggios: it could be easily tuned to generate as many arpeggios as we want, we just use it to compare the user perception; and to see the arpeggio frequency of *Cooperative* in relation to an "upper bound" where those are directly generated. In terms of power chords, we find a frequency of $24.3\%(\pm 8.9\%)$ in the real executions of *Cooperative*.

In conclusion, while the *Biased Random* system generated a higher frequency of arpeggios, it did not necessarily translate to being recognized as music more often than the *Cooperative* system. The presence of human agents in the *Cooperative* system is

vital, emphasizing the system's capacity to integrate user interactions into a coherent and engaging musical experience.

4.6 Discussion

The findings from this chapter provide evidence of the effectiveness of the Implicit Cooperation approach in emergent music generation. We expect that this study on Implicit Cooperation in music generation will contribute to the fields of computational creativity and co-creation by offering new insights and methodologies that can be applied to interactive systems. By demonstrating that non-expert users can actively participate in the music creation process through their in-game actions, this research bridges the gap between human creativity and algorithmic composition, fostering a more inclusive and engaging creative environment.

Contributions to Computational Creativity and Co-Creation: The Implicit Cooperation model represents a novel approach to computational creativity, emphasizing the seamless integration of user interactions in the creative process. This approach not only enhances the user experience by providing a sense of contribution and ownership but also enriches the creative output through the unique and unpredictable elements introduced by human interaction. In the realm of co-creation, the study underscores the potential of collaborative efforts between humans and AI in producing complex, dynamic, and aesthetically pleasing art forms.

Implications for Interactive Music Systems: The study's outcomes have implications for the design of interactive music systems, particularly in gaming and virtual environments. By leveraging the natural actions of users within a game to generate music, developers can create more immersive and responsive experiences that dynamically reflect the narrative and emotional trajectory of the game. This can lead to a deeper emotional engagement and a more personalized gaming experience, where the soundtrack evolves in real-time based on player behavior.

Future Research Directions: Future work should explore the integration of Implicit Cooperation in diverse interactive contexts, examining its adaptability and impact across various genres and media. Additionally, investigating the psychological and emotional effects of participating in implicit music generation can provide deeper insights into the user experience, guiding the development of more intuitive and satisfying co-creative systems.

4.7 Limitations

This study has some limitations. The pool of users/players for evaluating the Implicit Cooperation approach was small (n = 10). While this population size allowed us to identify trends in the approach's capacity to generate meaningful musical outcomes, it will be necessary to run experiments with a larger pool of human subjects, in order to better confirm our experimental results.

While the use of *Random* and *Biased Random* systems served as initial benchmarks, future research should incorporate more sophisticated algorithmic baselines to fully assess the capabilities of the Implicit Cooperation model. The comparison with advanced music generation algorithms will provide a clearer understanding of the model's effectiveness and its contribution to the field.

In addition, our investigation was restricted to a single genre (i.e. artgame). Future research endeavors could encompass a more extensive array of video game genres and involve larger sample sizes to ensure the broader generalizability of our findings.

It is also worth noting that our user study involved individuals exclusively from Brazil, which may have implications on how the experiences with the systems were perceived.

Subsequent research efforts should consider these factors and explore how they interact with the visual aspects of video games, ultimately shaping perceptions of music quality. These considerations are essential for gaining a more comprehensive understanding of the intricate relationship between graphics, music, and player experience in video games.

4.8 Conclusion

In this chapter, we proposed a system where a human agent collaborates in emergent music generation. However, the agent collaborates as a "side-effect" of its behavior, and does not need to be actively involved, and is not required to be a music expert. We prove the correctness of our algorithm, and study the probability of generating music, showing that it is greater with the presence of a human agent. Our experimental results also indicate a larger frequency of arpeggios when a human uses our system, which indicates musical quality. Additionally, experiments with 10 human players show that users were not aware of their impact (in comparison with randomly drawing notes), but were more likely to define the product of the system as "music" when using our approach. It is still necessary, however, to runs experiments with a larger pool of human subjects, in order to better confirm our conclusions; and to verify our assumption that humans tend to move less in diagonals.

Chapter 5

Meta-interactivity

"Images are not only visual. They're also auditory, they involve sensuous impressions, bundles of information that come to us through our senses, and mainly through seeing and hearing: the audio-visual field."

— W. J. T. Mitchell

The concept of gamified interactive models and their novel extensions, such as playification, has been widely explored to engage users across various fields. However, in domains like HCI and Computational Creativity, these approaches have not yet been applied to support users in creating different forms of artwork, such as musical compositions. These techniques, while enabling new forms of interactivity with partially-autonomous systems, could also democratize the creation of artworks, making them accessible to non-experts.

In this chapter, we introduce the concept of meta-interactivity for compositional interfaces. Meta-interactivity extends an individual's capabilities by translating their efforts into coherent musical outcomes. It can be viewed as a form of conscious production where an initial action not only achieves its primary goal but also triggers a secondary action, such as manipulating visual elements to generate musical results.

We demonstrate the effectiveness of this approach through a novel system that allows non-experts to compose coherent musical pieces using imagetic elements in a virtual environment. Our experiments, conducted with both musical experts and non-experts, show that non-experts were able to create high-quality musical productions using our interactive approach.

5.1 Introduction

Producing a coherent musical corpus in an emergent fashion is a very challenging effort for interactive musical interfaces and AI systems to overcome. Since music is a form of expression inherently associated with feeling and sentiment, it is essential for developers to conceive devices capable of supporting and transforming an expressive motivation of an individual into a sound structure that the general public might acknowledge as a human-made musical piece. In addition, partially-autonomous systems that propose to support the creation of music usually demand artists and sound designers previous knowledge in determined AI techniques, such as Machine Learning, in order for the model to be trained according to their own goals, as we can observe in Roberts et al. [103]. As for the interactive musical interfaces, aside from the entertainment-based ludic approaches from video games [119, 33], sometimes they require a good understanding of musical theory (and sometimes practice) for the user to co-create pieces that can be recognized as "music". Or yet, the interactive model reduces the compositional experience to mini-games [115, 50], in which the user's creative expressiveness does not produce a result, restricting the experience to an instant entertainment, where the usage of sounds only serves as a scoring system element. That is, the user is not really creating a new music, but trying to mimic/play an existing one according to a scoring system. Thus, in many situations, non-expert users might feel discouraged and unmotivated to have a quality creative experience with an interactive musical system.

Another obstacle in the development of emergent content approaches is finding a perfect balance between the user's freedom to express himself/herself musically and the constraints of the co-creation algorithm, like the one we are going to propose in this work, in the attempt to foster the musical product of the interactive system to be coherent and homogeneous according to a given number of variables (e.g., tone, rhythm, tempo, etc). In our case, since it is intended that a non-expert also has a satisfactory experience with our system, the efforts must be centered on providing algorithms that work underneath the player's experience layer, trying to accurately fit the user's input in a *temporal musical structure* in a discreet fashion, improving the quality of the musical outcome while preserving the user's expressivity and original intent. That is, the algorithm should be able, regardless of how the notation system will be presented (whether it is visual, audible, or tactile), to allow the user to transform his/her ideas with good accuracy through the interactive model.

Many researchers study the interaction of human agents with partially-autonomous systems for emerging artworks. For instance, Jacob and Magerko [63] propose an approach where a human and an agent collaborate to produce movement-based performance pieces. Davis et al. [26] describe a system where a user takes turns with an AI system when drawing on a canvas. Similarly, many contemporary systems

produce musical pieces by collaboration between human agents and AI agents. In all previous works, however, human agents need expertise to produce high-quality outcomes.

Thus, in this chapter, we introduce the concept of meta-interactivity, and how it can establish a relation with different expressive endeavors to homogeneously produce a desired outcome. This approach can be seen as a powerful gamified tool that can be used to explore a user's creative proficiencies, such as drawing, and extend its potential to other expressive fields, such as music production. To explore this concept, we present a novel Virtual Reality Musical Instrument (VRMI) called Bubble Sounds, an interactive art installation that enables non-experts in musical theory and performance to produce an original and pleasant ambient music piece through a friendly interface that allows him/her to use visual elements, such as colors, as tools to sculpt a coherent musical corpus.

Bubble Sounds shares conceptual similarities with interactive music systems, such as Google's Bach Doodle [17], which allows users to create music by interacting with a playful interface. However, Bubble Sounds differentiates itself by offering a more immersive experience through its Virtual Reality (VR) environment and a unique interface that emphasizes the use of visual elements as direct conduits for musical expression. Bubble Sounds provides a broader canvas for musical creativity, allowing users to create a musical piece from scratch using visual metaphors. This approach not only facilitates an intuitive understanding of musical elements but also enriches the user's creative experience by offering a multi-sensory engagement with the music creation process.

Some of the findings of the conducted user studies in this chapter are:

- Meta-interactivity is efficient in improving the user experience and supporting the generation of music.
- Meta-interactivity accomplishes its goal of extending users' proficiencies.
- Music generated by novices using this approach is perceived as equal to or superior in quality to that produced by experts.
- Music generated by novices through this approach matches music produced by experts.

We evaluate our system on its compositional capacity with human subjects, as well as the music produced through it. We show that non-expert users in musical theory and practice were capable of creating music as well as experts.

5.2 Meta-Interactivity – Overview

In this topic, we introduce the concept of meta-interactivity as a novel approach to be explored in gamified experiences and AI systems. Gamification consists of using particular game development techniques and strategies in other contexts that are not necessarily associated with gameplay mechanics themselves [49, 27]. It is a reframing of Game Design elements in order for it to be applied in different fields or domains, especially through the design of gratification-wise mechanics for improving both experiences in the concrete reality and all kinds of interactivity we can have within cyberspaces. Bringing such elements to experiences other than games, as we can observe examples for educational purposes [29], for staff training [47], and many others, fosters a huge impact in points such as engagement, productivity, and focus, thereby making tasks in any context tangible and more pleasant to achieve. According to Lucero et al. [77], features that make games and play engaging can also make other kinds of products more enjoyable and meaningful, increasing the quality of the overall user experience. Playfulness, in other words, can be a positive feature in products that go beyond pure entertainment. Similarly, the concept of playification [84] has recently been proposed as an extension to gamification. Differently from gamification, which uses specific game elements to engage individuals (e.g., score and trophy systems), playification aims to employ a gameplay mechanic by itself. This means using the actual game's interactive model instead of external rewarding elements, thus substantially approximating any experience to that of a video game. This approach was employed to reconfigure physiotherapy sessions for elderly inpatients, with the aim of increasing treatment efficiency.

With meta-interactivity, we aim to address the same goal and effect: engage users to accomplish complex tasks, such as music composition, through a ludic effort, that not only makes the tasks easier to be accomplished but also more engaging and fun. The way we propose such an effect is rather different, however. Our effort is centered on algorithms that establish bridges capable of translating different artistic endeavors, that are mainly based on expressive efforts, into new artistic instances. From this novel form of user-system relation, we expect coherent forms of artwork emerging from interfaces designed with ludic elements, that evoke an aptitude of the user and convert it into a new instance. It can be comprehended as a way to translate different creative motivations, like "drawing" to "musical composition", in order to propose different relations between the system and the users, such as "painting a music", "playing" (in the musical sense) on a canvas. Ultimately, it consists of designing minimalist, very intuitive, and easy-to-use interactive models that generate complex outcomes.

This kind of subverting behavior has been observed in classical art, more specifically in the abstractionism movement, generally understood as a form of

expression that does not represent objects in their proper form, according to our perception of concrete reality. For instance, Kandinsky connected painting with musical composition, drawing forms that refer to motion and that can be perceived as a musical notation system [23]. Bringing a similar translation model to digital environments establishes very particular challenges in order for the approach to be effectively implemented. For instance, Hunt and Kirk [60] examined many different strategies for mapping human gestures onto interactive systems in order to improve performers' expressivity in live performances, allowing them to compose music in real-time.

In Chapter 4, it was explored Implicit Cooperation, a novel form of collaboration that consists of translating the player's actions, such as exploring a virtual environment, into a musical corpus. In the approach described in that chapter, however, the user was not necessarily aware that he/she was actually contributing to the quality of the emergent music. Meta-interactivity, on the other hand, proposes that the user's focus should be on the musical production, as if they were manipulating a musical instrument. Also, the relationship we intend to establish between the user and the system should allow the emergence of coherent music through a playful interface that does not stipulate excessive restrictions on the creative side.

Other similar translation approaches have also been explored in recent works. For instance, Duckworth et al. [31] describe the design and development of Resonance, an interactive tabletop artwork that targets upper-limb movement rehabilitation for patients with an acquired brain injury. The artwork consists of several interactive game environments, which enable artistic expression, exploration, and play. Each environment aims to encourage cooperative and competitive modes of interaction for small groups of participants in the same location. This is an example of subverted game mechanics usage as tools to achieve different goals, such as people rehabilitation. The potential of similar approaches employed as a way to extend an individual's capabilities is, however, yet to be explored.

Meta-interactivity allows us, furthermore, to reflect in more depth about the role of games and play as a cultural phenomenon. Miguel Sicart [112] argues that play evokes a sense of presence in the world, it is a way to understand our surroundings. It is also a form to connect individuals, fostering interactivity. It goes beyond the game itself; it is a mode of "being human". According to the author, a theory of play does not derive from a particular object or activity, it is a tool capable of bringing complex interactions between people as an extension of their daily life activities. Thus, it is not separated from reality, but part of it. From this perspective, we see that once the right connections are established to transform individual engagement into coherent musical outcomes, it is possible to turn mundane daily activities into playful experiences. This approach enables success in various tasks without the need for time-consuming and sometimes tedious training.

On the other hand, time invested in an aptitude can be very rewarding; it is great to get through a learning process and perceive an evolution. However, in cases such as music theory and practice learning, it also leads to withdrawal, since often a satisfactory evolution is attributed to "talent" [88], and this is a factor that distances people from developing basic musical skills. We believe the meta-interactive approach can help to deal with the overwhelming steps involved in the learning process of determined tools and techniques, such as music practice and composition, giving individuals more resources to explore tasks in a more intuitive, powerful, and playful way.

In the next section, we will introduce a system developed through a meta-interactive approach, where a coherent musical compositional emerges out of a playful interaction of a user in a virtual environment. Although the user is engaged in the proposed activity (that is, the user is aware he/she/they is composing a piece of music), it is done through a very minimalistic interface that proposes the user to "play" (in a ludic sense) with a virtual world by bursting bubbles under the sea, and this action is translated to a composition effort for interesting music to emerge. That user can learn and appropriate from interactions with the environment and use it in his/her favor for a complex behavior (i.e. a new musical instance) to be created.

5.2.1 Bubble Sounds

Bubble Sounds is an interactive musical system developed as an installation for artistic environments. It offers a game-like experience, but its proposal is rather different from the common entertainment-based approach focused on reward-wise mechanics. It enables the emergence of music through the interaction of human agents with the system, but not limiting the coherent production to experts in musical theory and performance. In fact, the idea is that anyone, regardless of cultural background and expertise, should be able to create music through its novel notation system. Hence, this approach enables non-experts to compose interesting musical pieces through the usage of imagetic elements in a virtual environment, which turns the system into a kind of ludic musical instrument that explores the user's perception of visual elements and translates it into music.

It was originally developed for virtual reality devices, more specifically for the Oculus Rift, and it uses the device's accelerometer to allow a 360-degree visualization and interaction with the 3D environment. It also uses a microphone to capture sound input, being the possible agency for the user over the compositional system. This project was developed using the Unity 3D engine, and the game assets were created using Blender 3D.

The aesthetics of the environment are based on an underwater theme, where bubbles emerge from the ground all the time. It presents a peculiar form of arranging

musical notes, composed by a stack of concentric circles that works like a vortex, going all the way up from the virtual environment's floor to its top, as shown in Figure 5.1. The user is located at the center of this vortex, which can be comprehended as a stack of tracks that orbits the user's position in the 3D environment, capable of receiving musical notes. It is a tridimensional representation of a musical score, which is circularly interconnected merging the edges and forming a loop.

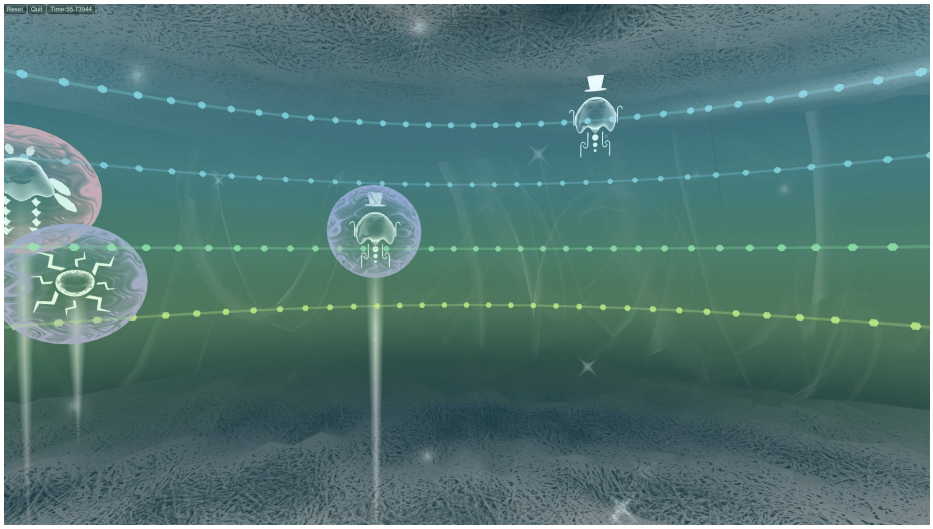


Figure 5.1: Bubble Sound's interface.

A differential of this system is how the interactive model was conceived to foster new ways in which the user dialogues with the system. As mentioned earlier, the system uses a microphone to capture the user's input, and this is how the notes are activated in the vortex. As the user directly stares at a determined bubble, he/she can clap hands, and the variation in the audio frequency will release the notes inside the bubbles, which will then start orbiting a vortex track, producing a musical loop. The tempo in which each of these notes orbits the vortex also follows a procedure that allows the emergent musical corpus to be coherent in a matter of rhythm, and it never generates cacophony (i.e. unpleasant and chaotic musical structures), as we will explain later. In this way, since the system arranges the notes in terms of tempo, allowing for coherent rhythmic structures to emerge, the clapping hands also produce a meaningful rhythm for anyone outside the experience observing the user, generating 2 layers of interactivity that produce coherent musical instances. Before going into more depth about how our approach achieves this, we will present the system overview in more detail.

In the underwater environment, the user sees 12 types of bubble colors randomly emerging from the ground all the time, each one corresponding to a musical note

that matches with a colored circular track in the vortex. Each bubble contains a microorganism, from a total of 6, that presents variation in their speeds and octave, as shown in Figure 5.2. It is up to the user to choose, according to a color procedure that matches note qualities (i.e. tones) with colors, when to release these notes from the bubbles in order to trigger its sonorities.

In other words, the colors carry an important guideline for musical composition: it establishes for the microorganism inside the bubble a range from bass to treble sounds following a visual procedure, done by relating the pitch of each note in the bubble to the color spectrum (Figure 5.2). The user can interact with the bubbles to release the 6 different types of microorganisms placing them on the vortex of tracks according to his/her own desire.

In the same way, users can also remove notes that are already orbiting according to their own will and on the fly. Thus, it is possible to test if a determined note suits the composition, and if not, it can be easily removed through a clap of hands (or a mouse click) after having this note in the focal point, and then the microorganism will fade out and disappear from the vortex.

The fastest microorganisms represent high-pitch notes, and the slowest ones the low-pitch notes. Hence, besides learning the microorganism tone and speed, users can also easily identify musical notes even before releasing them from bubbles by analyzing the speed at which bubbles are dislocating vertically. Thus, even with microorganisms idle inside the bubbles, the user might be able to understand, only by observing the visual elements (i.e. bubble colors meaning tones) and its motions (bubble speeds meaning the octaves, which is the same for the microorganism within), the sound quality of the notes that are appearing on the screen. They can then judge whether the note's characteristics fit or not the music being composed.

Like the bubbles, there are also 12 tracks in the stack (vortex), representing all the possible notes in a scale, be it a chromatic, pentatonic or diatonic, considering its accidents (C, C#, D, D#, E, F, F#, G, G#, A, A#, B), as shown in Figure 5.3. Only the commas, which are intervals smaller than a halftone, are not being approached; which gives us a good range of tonal variations from the Western musical system. As previously mentioned, the vortex works like an interactive tridimensional representation of a musical score, merged like a cylinder and with a stylized score sheet projected in its interior faces, where the user has a 360 degrees view of the whole cluster of notes orbiting him/her.

The vortex structure works as follows: the lowest notes in the same octave (i.e., in the same progression of notes, where we have the first one, the *fundamental*, and the following 7 notes, until we reach the next octave from the fundamental) orbit at the lowest part of the screen (i.e., lower tracks), and the highest notes at the top (i.e., upper circle tracks), also shown in Figure 5.3. The lowest note in the stack of tracks on the vortex is F# (our fundamental), and the highest note is F (also forming a

NOTE	COLOR
F	#A955A9
E	#3A6BA0
D#	#0499CB
D	#38B09E
C#	#56AE77
C	#8CC84F
B	#FDFD05
A#	#FFA42A
A	#FF7940
G#	#FE4A58
G	#FD3667
F#	#E13D9F

MICROORGANISM	SPEED	OCTAVE
	(slowest) 5 (00:00:16.0)	(lowest pitch) 1
	10 (00:00:08.0)	2
	20 (00:00:04.0)	3
	40 (00:00:02.0)	4
	80 (00:00:01.0)	5
	(fastest) 160 (00:00:00.5)	(highest pitch) 6

Figure 5.2: Microorganisms speeds and note-to-color match system.

loop vertically), which gets more sonorously high-pitched as it goes up. These notes are associated with a color, as mentioned earlier. E.g., F# is pink and F is lilac, as shown in Figure 5.2. Thus, besides the tonality that each note emits, which might be enough for expert users to use in his/her creative process, the color system will also support non-expert users to quickly recognize and determine the best notes to use in their composition. This translation approach of tones to colors does not only favor non-experts, as experts might also benefit from it for quick note recognition and activation.

When the bubbles are burst by the user, the microorganisms inside them start to orbit around the player in their own vortex track and in a pre-determined speed according to the microorganism's type (as shown in Figure 5.2), considering the color match of the bubble and the vortex track. Only then it start to emit its sound (while inside the bubbles, the microorganisms are idle, not producing their sonorities).

There is a 3D asset/model in the experience we call "pillar", that are faint shapes that hold the whole structure of tracks (these objects are being identified in Figure 5.3 and 5.6, along with other interactive key-elements of the system's mechanics). When a microorganism is released and starts orbiting the vortex, it collides with these pillars, triggering the note it carries. There are 3 pillars around the cylindrical musical score structure, so, in a 360^{o} turn, a microorganism will have its sonority triggered 3 times, as shown in Figure 5.3. The duration of each musical note between cycles of interaction with the pillars varies according to the quality of each microorganism and its respective speeds. For example, in Figure 5.3, we identified the rhythmic figures

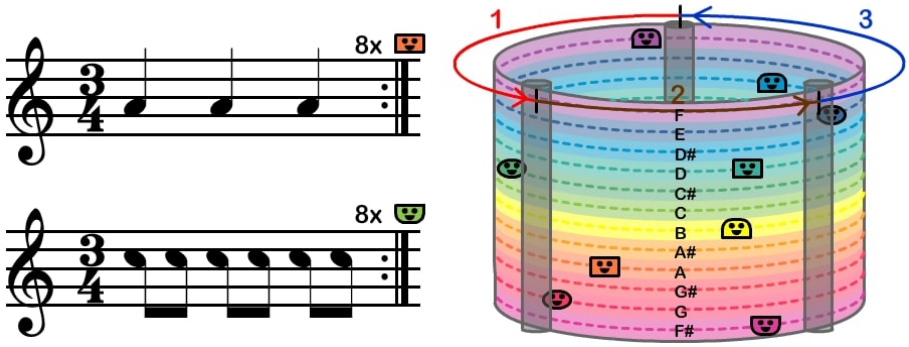


Figure 5.3: Representation of a 360º musical score, projected inside a cylindrical structure. Note that microorganisms have different speeds, as shown on the rhythmic figures presented on the compass on the left.

corresponding to the note A (orange microorganisms) and C (green microorganisms), showing that the A sustain (i.e. the time its audio signal will endure across time) will resonate for a longer period than C's.

The user has full control of the tone and pitch of the microorganisms triggered to orbit the vortex. Colors attributed to notes are fixed in the infrared to ultraviolet spectrum (e.g. *F* will always be lilac and *B* will always be yellow, as shown in Figure 5.2), so it helps to create a pattern for users to manipulate these colors in their creative process. Similarly, there are 6 octaves of the same note for users to use in their composition, which are determined by the microorganism speeds. In this way, it is possible to have many organisms orbiting the same vortex track presenting audio variation (from bass to treble), according to the 6 types of microorganisms that present their own speeds and octaves, as shown in Figure 5.4. Microorganisms orbiting the same vortex track have the same tone (e.g. only *C*, or *C#*, or *D*, etc) but they can vary in their height channels from low to high pitch. This is visually expressed by the slow and fast orbit motion of the microorganisms in the same vortex track according to their octaves (i.e., their corresponding speeds).

Hence, in short, bubble colors point to note quality, from F to F# (a complete 12-note loop) and the microorganisms point to speed (and, as such, by the octave of that note). In this way, the system not only provides a very good range of notes for the user to choose from (similarly to most pianos, that offer 7 octaves, the system is offering 6) but it also offers an organized way to visualize the cluster of notes orbiting each track, clearly separated by colors. This feature, as a compositional resource, also resembles octave pedals, used in musical instruments such as guitar and bass for more tonal diversity, since commonly string instruments are restricted to 22 or 24 frets on their necks.

Figure 5.4: All note possibilities (represented by bubble colors), varying from F to F#, and all the microorganisms octaves the bubbles can carry.

It is important to emphasize that the user does not need to burst the bubble to understand the speed of the microorganism within, since the actual speed is presented even when they are idle inside bubbles, by the speed at they are emerging from the ground (it is the same speed in which it is going to orbit its correspondent vortex track). Also important to note that all these elements from music theory, such as tone, pitch, sustain, scale, melodic and harmonic lines, etc., were completely abstracted to a purely visual system based on colors and movements, with a clean interface, that presents no buttons or complex instructions. The users just look around the

virtual environment and clap their hands to trigger the sonorities carried by the microorganisms.

The release command is made through audio recognition, in which all the time a determined bubble is on the line of sight of the user identified through the Ray Trace, that is, the vector that represents the distance between the player's location (i.e. where the user is looking at) from the location of the actual bubble, the audio frequency variation will cause the bubble to burst to release the microorganism within. As mentioned earlier, the recommended audio signal input for this action is to clap hands, as it also produces an interesting new layer of music interactivity for an external listener.

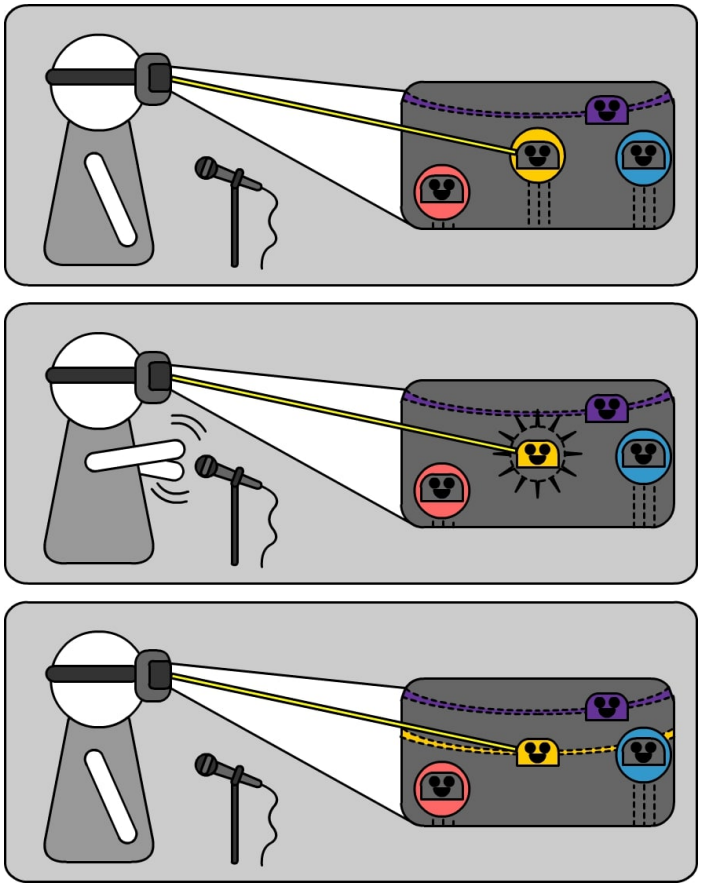


Figure 5.5: Interactive flow of Bubble Sounds.

In Figure 5.5, we identify the interactive flow for a musical note to be triggered during the experience. The sequence shown is:

- In the first frame, the user spots any emerging colorful bubbles with idle notes moving on the y-axis of the screen. We can observe that the Ray Trace is aiming

at a determined bubble (yellow line coming from the user to the spotted bubble) as it is in the user's focal point.

- In the second frame, with a bubble in the focal point (whose color determines the note), the user can release the microorganism within (which defines the octave according to its speed) by clapping hands on the microphone.
- In the third frame, after the bubble bursts, the microorganism follows a linear path to the corresponding vortex track and starts orbiting the player, producing its sonority.

As mentioned before, in the same way, users can trigger idle notes, they can also remove orbiting notes at will as they might not be wanted to a given composition or changes are required due to a natural evolution in complexity and meaning of a musical piece across time. This can be done through a similar procedure described above (i.e. the user spots an unwanted microorganism and claps their hands or left-click in the mouse button). Thus, yet minimalistic, the system's interface allows for easy customization in the emergent musical corpus.

At the exact moment, a bubble is burst, the microorganism inside follows a linear interpolation from the point it is released (considering bubbles only move in the y-axis) to the appropriate track, represented by the match of the bubble color with a corresponding circular layer in the infrared/ultraviolet spectrum. Note that the quick motion of the microorganism from the point it is released to its actual track will not affect the sound; it will only trigger the note it carries when it is actually orbiting its corresponding track, colliding with pillars. It is also important to emphasize that the encapsulated notes (the ones inside the bubbles) do not produce sounds; they only start to react when they are properly released.

A feature of the system that is not yet being fully explored in all its potential in the current version is the dynamically generated 3D assets in the environment's background, as shown in Figure 5.6. They are presented in a wide range of possible 3D assets, like algae, corals, shipwrecks, treasure chests, etc. It is intended that these assets help create soundscapes that match our actual underwater theme, also having a sound signature attached to them. However, the sound output they provide is different in length – they tend to last longer, as a subtle background harmonization is provided by the system. Since the system was designed as an interactive installation for artistic environments, it is intended for these sounds to always keep the system active, producing subtle harmonization, thus never leaving it completely silent in case no one is interacting with it. In this way, the system can still provide a pleasant audiovisual experience for those just watching or passing by it. The intent behind this module is also to dynamically create micro-narratives associated with the underwater theme. We will discuss more about this in section 5.4.

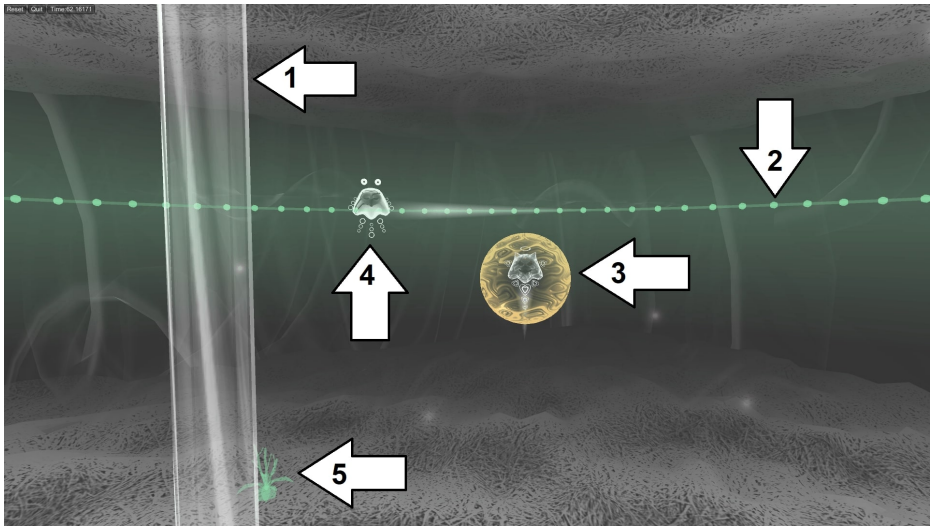


Figure 5.6: Through the arrows you can identify the following elements of the system: 1. Pillar; 2. Vortex track; 3. Bubble; 4. Microorganism; 5. Dynamically generated asset.

As previously mentioned, our system focuses on allowing non-expert users to produce a coherent musical corpus in a matter of *temporal structure*, enabling the system to work like a smart musical instrument, capable of fostering creative freedom for the user to musically express himself/herself without limiting the compositional process (thus allowing the output to result in a pleasant musical piece). This is done by a simple procedure: we make each microorganism orbit the vortex at its own speed, which follows a predetermined geometric progression (5, 10, 20, 40, 80, 160), as shown in Figure 5.2. This approach guarantees there will not be notes being played out of the tempo of the music, thus guiding non-experts to generate rhythmically coherent musical structures.

For the interested reader, a video demonstrating the system is available at `https://youtu.be/lExDgGxbeDQ`. Figure 5.7 presents a timeline of screenshots illustrating the demonstration of the Bubble Sounds system as showcased in the accompanying video.

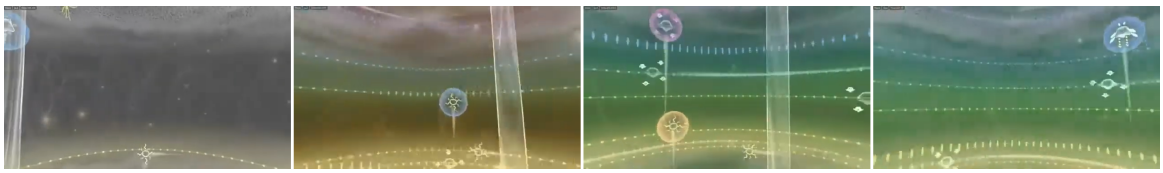


Figure 5.7: Screenshots of the Bubble Sounds video.

5.3 Experiments

Experiments were conducted in two stages. The first stage involved 42 human subjects, where 29 (69%) were non-experts in music theory, whose experience was mostly based on listening, and 13 (31%) were experts, with high knowledge in both music theory and practice. Users were trained for 3 minutes, provided with basic instructions about the system functionalities, and then allowed to freely interact with the virtual environment to get a general feel for it. Our goal at this stage was to evaluate Bubble Sounds and its meta-interactive mechanic.

The second stage involved 111 human subjects, who answered an online form that randomized three music samples from three different databases captured during the previous assessment sessions, consisting of compositions by both experts and non-experts, along with randomly generated music.

We randomly selected 10 samples created by experts, 10 by non-experts, and added 10 generated by a random script to be included in the database. These random samples were generated by an algorithm that randomly triggers notes and puts them into orbit around the vortex while a "filter" arbitrarily removes notes to simulate user agency over time.

This approach aims to keep the randomly generated musical samples less cluttered in terms of the quantity of orbiting notes. If the random algorithm continuously triggered notes, it would not only cause performance issues but also make it easy for listeners to identify these productions as artificial. Our goal was to ensure that the random creations did not sound very dissonant compared to the human productions.

Each sample on the form (random, non-expert, and expert) was presented in randomized order. For each category, we randomly chose one of the 10 samples to display to the user. Our goal at this stage was to evaluate the musical outcome of our system. Users did not interact with the Bubble Sounds interface during this stage.

In this section, we refer to ***expert/non-expert samples*** as musical samples produced by experts/non-experts, respectively, and ***expert/non-expert evaluators*** as the expert/non-expert users who evaluated those productions. This study was approved by the ethics committee of Lancaster University.

5.3.1 Stage 1 - System Evaluation

The human subjects for this session were primarily identified as game developers from FUMEC University, graduate students from IFMG, computer science students from UFMG and USP, and game enthusiasts and aspiring musicians nominated by participating individuals, all from Brazil. Each user agreed to a consent form before starting the evaluation session, with a clear explanation of the entire process. No sensitive data from any users were recorded.

Q1.	How do you classify your skills as a musician? ( ) Novice. ( ) Expert.
Q2.	From 1 to 10, how do you classify your experience with the system? Uninteresting - 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 ( ) 8 ( ) 9 ( ) 10 ( ) - Interesting
Q3.	How do you rate your own composition? Bad - 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 ( ) 8 ( ) 9 ( ) 10 ( ) - Excellent
Q4.	How much fun did you have during the experience? Little - 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 ( ) 8 ( ) 9 ( ) 10 ( ) - Plenty
Q5.	Establish a relationship between the generated music with a feeling (E.g.: calm, anguish, tension, etc).
Q6.	How do you classify the compositional interface of the system? Bad - 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 ( ) 8 ( ) 9 ( ) 10 ( ) - Excellent
Q7.	Which of the compositional elements best supported your music creation process? ( ) The relationship between colors and tones for partial identification of musical notes. ( ) The relationship between movement and tempo for rhythmic definition. ( ) The imagetic representation of low and high pitch tones in the vortex tracks. ( ) None of the above.
Q8.	Identify the main focus of your attention during the experience with the system: ( ) My focus was on sound production. ( ) My focus was on the imagery elements, such as the objects that appeared in the background. ( ) Arbitrary, I just explored the interactive possibilities provided by the system. ( ) My focus was on learning the interface and its usability. ( ) None of the above.
Q9.	Would you like to share any additional details about your experience? If so, write it below.

Table 5.1: Bubble Sounds assessment questionnaire.

The system version used for the assessment was adapted to work without both the VR device and the microphone input since the sessions were conducted online. A regular monitor with a mouse controller was used for manipulating the virtual environment view, turning the interaction into a more "video game-like" experience. After freely experimenting with the system, subjects were asked to compose a 1-minute sound piece, which was recorded for use in the second stage of the assessment. Users were queried with the questions shown in Table 5.1.

In Q1, we identified users by their proficiency in music theory and practice, enabling us to evaluate both the general scenario, considering all 42 feedbacks, and the individual scenarios from both experts (13 users) and non-experts (29 users). We considered as experts all individuals who have mastered the practice of at least one instrument, have had solo or group performance experiences, have basic knowledge of music theory (e.g., capable of reading musical scores), or who had previous compositional experience.

In Q2, we queried users about their overall experience with the system. The general scenario showed an excellent evaluation, with a $\bar{x} = 8.09$ out of 10, indicating a pleasant experience during the session, as shown in Figure 5.8 (a). In the independent scenarios, experts had a $\bar{x} = 8$, and non-experts had a $\bar{x} = 8.12$, showing that all users had satisfactory experiences regardless of their musical expertise.

In Q3, we asked users about their perception of their own composition with the system. The general scenario showed a good result, with a $\bar{x} = 7.28$, as shown in

Figure 5.8 (b). In the independent scenarios, experts had a $\bar{x} = 7.3$, and non-experts had a $\bar{x} = 7.22$. Despite only having 3 minutes of training before the sessions, users were satisfied with their compositions. This result demonstrates that our approach is promising in enabling non-experts to produce interesting musical pieces with our ludic interface in a very spontaneous fashion.

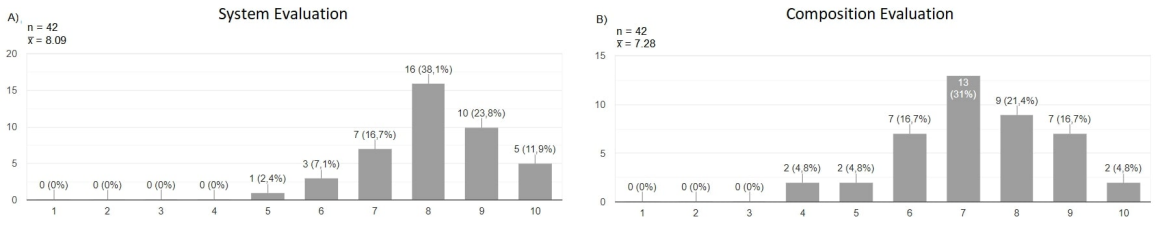


Figure 5.8: General perception over the experience with Bubble Sounds.

In Q4, we queried users about how much "fun" they had during their experience. Despite being a subjective question, our goal was to verify how well our minimalistic interface matched users' perceptions of Bubble Sounds. The general scenario showed a $\bar{x} = 7.92$, as detailed in Figure 5.9 (a). This promising result was obtained even though the version used during the sessions did not use the VR device or the microphone input. In the independent scenarios, experts had a $\bar{x} = 7.92$, and non-experts had a $\bar{x} = 7.9$. These results provide a positive glimpse of how well our approach addresses the goal of allowing non-experts to produce coherent music.

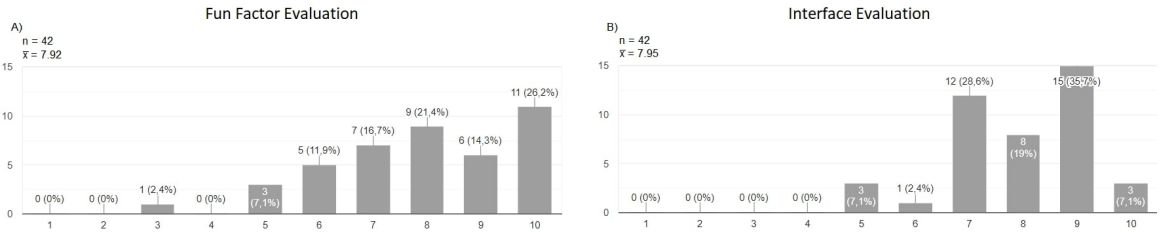


Figure 5.9: General perception over the compositions generated by Bubble Sounds.

In Q5, we asked users to identify and assign a feeling to their own composition. The responses varied widely, with "relaxation" (50%) and "tension" (30%) being the most common feelings. This contrast highlights the music comprehension as a *listening phenomenon*, where the cultural background of each individual significantly influences their perception of the musical corpus.

In Q6, we queried users about the system's interface to assess how well our approach supports the compositional process. The general scenario showed a $\bar{x} = 7.95$, as shown in Figure 5.9 (b). This positive result indicates that the interface

effectively supported user expressivity even in a simplified form. In the independent scenarios, experts had a $\bar{x} = 7.84$, and non-experts had a $\bar{x} = 7.93$. The higher mean for non-experts suggests that they relied more on the interface and its imagetic devices for fostering musical composition than experts.

In Q7 and Q8, we asked users to identify the system elements that supported their music creation process and their focus while interacting with the 3D environment. Our goal was to identify which visual elements were most important for ludic interaction.

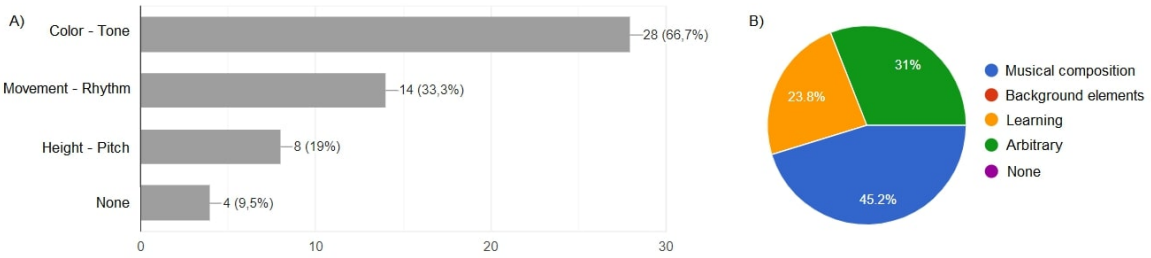


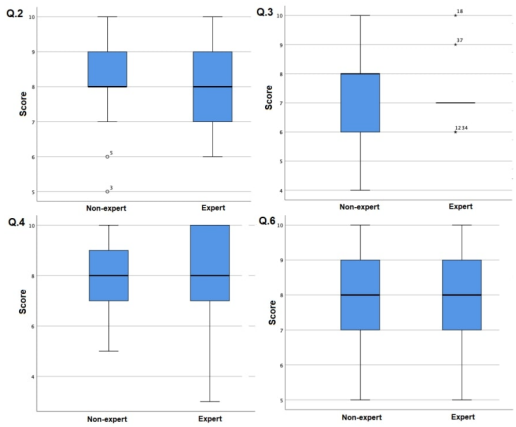
Figure 5.10: a) Users' perception of the co-creation mechanic and b) Their focus during the experience.

In Q7, the relationship between colors and tones for partial identification of musical notes (66.7%) and the relationship between movement and tempo for rhythmic definition (33.3%) were acknowledged as supportive compositional elements, as shown in Figure 5.10 (a). This trend was observed in both expert and non-expert groups.

In Q8, the majority of users (45.2%) focused on sound production, indicating engagement in the task of composing music through the imagetic interface. This result, combined with the means obtained in Q3, suggests that the approach effectively supports users in creating quality music. Another 31% of users indicated that their experience was arbitrary, highlighting the system's effectiveness in generating coherent music even when users were not focused on creating it.

A t-test was performed on our numerical variables to compare the means for both experts and non-experts in Q2, Q3, Q4, and Q6, and to verify if proficiency level influenced their perception. As shown in Figure 5.11, the high p-values ($p > 0.1$) indicated a non-significant variance in experts' and non-experts' perceptions of their overall experience, composition score, how much "fun" they had, and their evaluation of the system interface. Thus, proficiency did not play a major role in users' perceptions, and both groups provided positive feedback about their experience with the system.

A Chi-square test was performed on our categorical variables to examine the relationship between proficiency levels and perceptions in Q2, Q3, Q4, and Q6. High p-values ($p > 0.1$) indicated a non-significant variance in experts' and non-experts' perceptions, similar to the t-test results. We concluded that proficiency did not



T- test

Variables	Musical knowledge: Non-expert (N=29)	Musical knowledge: Expert (N=13)	p-value
Q2 – Experience score			
Mean	8,13	8,00	
Median	8,00	8,00	0,740
SD	1,15	1,23	
CI	(-0,651 – 0,909)		
Q3 – Composition score			
Mean	7,23	7,23	
Median	8,00	7,00	0,992
SD	1,59	1,09	
CI	(-0,980 – 0,970)		
Q4 – Fun score			
Mean	7,90	7,92	
Median	8,00	8,00	0,973
SD	1,56	2,18	
CI	(-1,191 – 1,152)		
Q6- Interface score			
Mean	7,95	7,85	
Median	8,00	8,00	0,837
SD	1,18	1,57	
CI	(-0781 – 0,960)		

Figure 5.11: Non-expert and Expert means comparison for Q2, Q3, Q4 and Q6 through a t-test.

significantly influence users' perceptions, and both groups provided positive feedback about their experience with the system.

5.3.2 Stage 2 - Music Evaluation

During this stage, human subjects answered an online form containing three audio samples captured from the first stage of the assessment. Participants did not interact with the system; they only listened to music produced through it. Before starting, each user agreed to a consent form with a clear explanation of the process. No sensitive data were recorded.

The samples and their order were randomized on the form, developed by the authors and available at `https://phersu.com.br/bubblesounds`. Users did not know whether they were listening to a random, novice, or expert sample. We queried the users with the questions shown in Table 5.2.

In Q1, we queried users about their knowledge of music theory and practice. From the 111 feedbacks received, 41 were from experts (36.9%) and 70 from non-experts (63.1%).

In Q2, we asked users to identify which sample they liked the most. Both the order of presentation and the sample itself were randomized. As shown in Figure 5.12 (a), the majority of subjects preferred the expert samples (39.6

We also analyzed the scenario considering all human-produced samples (experts and non-experts) against computer-generated samples (random). As shown in Figure 5.12 (b), human-produced samples were preferred by 76.6

In Q3, we queried users about the relationship between the random, expert, and non-expert samples. As shown in Figure 5.13, 31.8

Q1.	How do you classify your skills as a musician? ( ) Novice. ( ) Expert.
Q2.	Which music was the best? ( ) Sample 1. ( ) Sample 2. ( ) Sample 3.
Q3.	Identify the relation between samples. ( ) Sample 1 resembles Sample 2. ( ) Sample 1 resembles Sample 3. ( ) Sample 2 resembles Sample 3. ( ) The 3 samples sound diverse. ( ) The 3 samples sound similar.
Q4.	Which track sounded more professional, presenting more sophistication? ( ) Sample 1. ( ) Sample 2. ( ) Sample 3.

Table 5.2: User perception questionnaire.

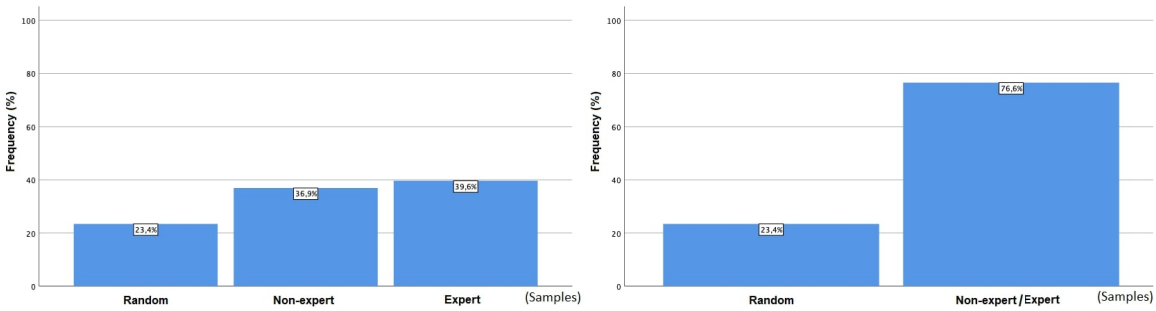


Figure 5.12: a) The best-voted samples according to the preference of both experts and non-experts. b) All samples (non-experts and experts) against random executions.

In Q4, we asked users to identify which samples sounded more "professional". As shown in Figure 5.14 (a), expert evaluators acknowledged non-expert samples as better music than expert samples, highlighting the effectiveness of our system.

Figure 5.14 (b) shows the preference of expert and non-expert evaluators for human-produced samples (experts and non-experts) versus machine-produced samples (random). Expert evaluators preferred human samples by a significant margin (85.4

We observed that users who preferred human-produced samples also rated expert samples as more "professional" with a high probability ($p \leq 0.01$ according to a

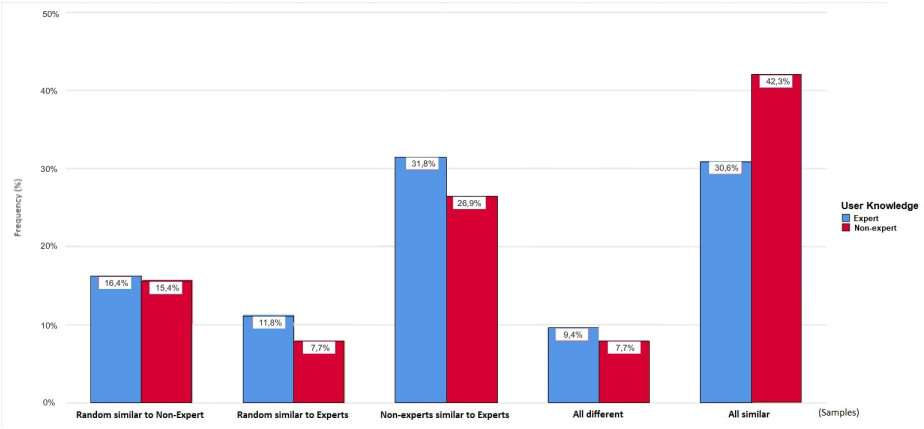


Figure 5.13: The relation between the samples according to the user evaluator's perception.

Chi-square test). This result indicates that the human factor is crucial for creating high-quality music. Our system effectively supports human-machine collaboration, empowering users in their creative efforts and achieving our goal of enhancing user compositional capabilities.

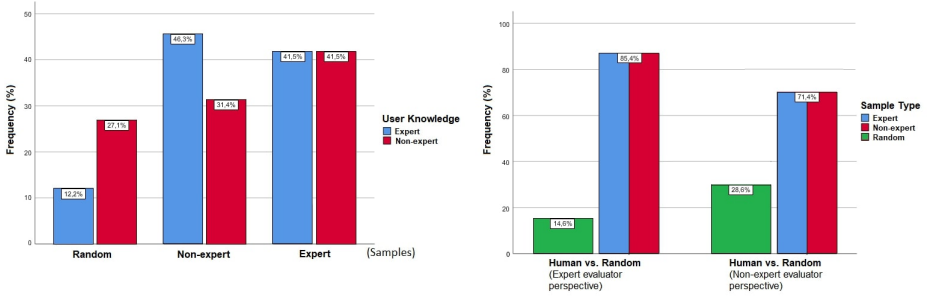


Figure 5.14: a) Individual user's perception regarding how professional each sample sounded. b) All human samples (non-experts and experts) against random executions.

5.4 Discussion and Limitations

Given the current stage of development of our system, we believe our approach showed an immense potential to allow non-expert users to create good music that might actually sound relevant to an audience. For instance, the results show that the Bubble Sounds color system helped non-experts to create high-quality productions, also without constraining the experts' expressiveness, finding a good balance between the conceived creative freedom and supportive guidelines. The equalization of these

two elements in an interactive musical experience is complex, and can be compared, in the game design process, to the narrative vs. interaction dilemma; it is not simple to devise efficient mechanisms that tell stories without compromising the interactive model (and vice-versa). Commonly, one ceases for the other to appear, as we can observe in cinematic cutscenes. In music video games, as discussed before, the compositional element is commonly reduced to musical mini-games in order for the experience to be "fun" while audio variations emerge from the player's interaction with game mechanics, thus not allowing for new music to be created from scratch.

As an extension of our evaluation, we would like to discuss further the creative freedom and support for collaborative creation in partially-autonomous systems, as well as other elements regarding our system and approach:

General considerations about meta-interactivity. Our approach dialogues with the concept of Metacreation of art using the paradigms of artificial life, or a-life, as discussed by Whitelaw, M. [120]. Metacreation is an interdisciplinary science focused on artificial systems that mimic the properties of living systems, explored in the 90s by contemporary artists who appropriated and adapted these techniques to create novel forms of art through the conception of "Cybernatures", which are interactive computational systems that simulate or "mimic" ecosystems in virtual worlds. This concept relates to our approach in the attempt to bring music performance and its physical manifestations into abstract virtual environments, bringing considerations about how algorithms can efficiently work translating ideas and endeavors into something novel and engaging. Games already do that; they establish metaphors for actions that should be performed by an avatar that represents the player within the experience. And there are protocols, that are well-conceived models that already become patterns, such as having directional buttons in the joystick (or any other device) for the player to move its character in the world. For Virtual Reality Musical Instruments (VRMIs), however, there are no pre-established interactive models or protocols to do so, and new patterns must be created in order to foster a bridge between musical instruments and their anatomic designs into novel ways to map it through common interactive devices, such as those utilized in games.

One element that separates interactive musical systems, such as Electroplankton [97] and Bubble Sounds to some form of ludic musical instruments is that it is harder to foster live improvisation with a band. This is not a creative freedom issue by itself, but a recurrent limitation in terms of collaborative creation, for example, when we think about these systems working in harmony in a live performance environment. Many works, like in Roberts et al. [103] attempted to experiment with autonomous approaches that learn from other instrumentalist's input to provide musical accompaniments for live performance, but it is not the kind of live improvisation that a musician would have when soloing over a harmonic rhythmic

base in a free jazz concert, for example. It is not spontaneous and on the fly, it demands time for training. Interactive musical systems, to the best of our knowledge, still work upon the creation of an editor for a musical corpus to emerge. These beforementioned interactive systems are not as bureaucratic and complex to manipulate as audio tools interfaces, such as FL Studio [24] or Pro Tools [41], but also not as responsive as musical instruments such as guitars or any other string instruments, that work as an extension of the human body, providing a quick conversion of ideas and feelings into a musical mass. They are somewhere in between.

Taking a meta-interactive approach to live performance is a very powerful way to experiment and test new interactive interfaces for musical creation, and definitely a path we will undertake in the future as an extension of this work. We believe that a more "performatic" version of Bubble Sounds, using both the VR device and the microphone input goes towards a more natural way to create music, allowing musicians on stage to work and interact with other musicians in harmony.

However, we also understand that evaluating the system at its maximum interactive capacity involves a considerably higher effort and investment to be implemented, which also points to a satisfactory and successful experiment carried out remotely in this version, which allowed us to glimpse many positive characteristics and some limitations of our approach.

We believe we advanced in regards to creating a ludic musical instrument by allowing users to manipulate all the 12 possible notes in a scale (considering its accidents) through our color translation approach, thus providing more resources for all kinds of users, regardless of their expertise level, to compose music. The work presented in Chapter 4 also proposed complex music to emerge from playful cooperation between the user and a system, however the user has less control of the notes that can be chosen at certain times since they use a dynamic grid on the environment's floor for notes to be triggered (thus, sometimes a determined note might not be at reach in the neighboring arrangement of notes, considering the user's current position in the grid).

According to Serafin et al. [111], the NIME (New Interfaces for Musical Expression) community has not shown the proper attention to the imagetic side of VRMIs. According to the authors, one of the reasons why VRMIs have not drawn its deserved attention in the HCI community might be due to the fact that musicians, sound designers, and enthusiasts rely mostly on some specific points when designing this kind of interactive models, such as auditory and tactile feedback, as well as in ways in which performers interact with the audience during live performances. In addition, another reason pointed out is that only recently portable visualization devices have become accessible in a matter of cost. In this regard, we highlight the importance of the convergence between VRMIs and videogames, such as Toshio Iwai's work in Electroplankton [97] and also Bubble Sounds, especially when the goal

is in proposing more ludic interactive models that aim on approximating the general public to artistic activities, such music composition. We believe this work contributed to showing that visual interfaces combined with sound interfaces can create powerful experiences in terms of improving the quality of outcomes generated by partially-autonomous systems.

It is also important to encourage thinking about new ways to conceive minimalistic interfaces for interactive musical systems, different from audio interfaces such as Pro Tools [41] and FL Studio [24], for example, which presents very complex panels with many options for the composer to create music (i.e. not friendly for lay users), as the one we proposed on Bubble Sounds, intending on providing to novice users the possibility to create complex outcomes without the necessity to manipulate many instructions at once or excessive time investment in training. Our evaluation shows that minimalistic interfaces for music creation generated very positive results in the quality of non-expert compositions if compared to experts, even with experts acknowledging non-expert production as better music in some cases. In this way, we believe we achieved interesting results in providing novel engaging ways to interact with musical systems, but there are still many obstacles to overcome. Thus, further experiments on new approaches are necessary.

Due to the remote nature of our assessment, we were unable to evaluate the learning capability of Bubble Sounds as a ludic musical instrument, particularly in terms of how users could learn to play it over time. In future evaluations, we plan to conduct pre and post-surveys to track whether the visual elements and the color-to-tone feature can foster learning, especially among novice users.

The role of sentiment in the emergence of art. Our evaluation contributed to a discussion of whether or not it is possible to dissociate music production in fully and partially-autonomous systems from human sentiment. After all, is it possible for a music piece to exist and be relevant without the human factor?

As presented in Figure 5.12 (b), we observed an expressive preference among users who participated in our evaluation process for the music created by human agents over the music generated through our random procedure. Even on a system such as Bubble Sounds, which organizes the musical corpus in a matter of temporal structure (thus fostering rhythmically coherent music to emerge even from chaotically arbitrary interaction), the perception of external listeners still acknowledged the human production as better music in both experts and non-experts scenarios.

This result empowers our system in its attempt to be a ludic musical instrument, since, differently from autonomous approaches, such as [103, 59, 15, 1], users expressivity plays major importance in the experience. In our case, the system and user depend on each other for the emergence of music, so feeling and sentiment will always be a preponderant factor in the compositions to emerge from our approach.

Thus, on Bubble Sounds, the algorithm has less to do with the music creation itself, and more as a support for users to achieve more complex results in their creative process.

Additional feedback from the 1st stage of the evaluation. During the assessment sessions with users who experimented with the system, we received feedback regarding the interface for composition, and this feedback will also be taken into account when planning the next steps in the development. For instance, it was mentioned the possibility of a dynamic generation of circular tracks in the vortex (for drumming patterns) in order to better guide non-expert users on the construction of the rhythmic structure of the music. Also, the dynamically generated images in the background are randomly generated in the current version, and this was noticed by some users. Indeed, at this moment, this resource is not yet being fully explored. Now, it does not yet perform the function of a compositional element (it is not yet triggered by the user's action, but as a collaborative complement from the system). This is a very promising module that can be improved in this work to help achieve better results, allowing the users to create even more engaging musical outcomes. We intend to create a mechanic guideline to propose the emergence of micro-narratives associated with sound qualities, making it become a stronger imagetic compositional element for the user to use in his/her expressivity (thus, empowering our translation approach).

Regarding outcomes from the assessment sessions. Although we could verify that the background knowledge did not play an influence in the results (as discussed in Q8 of the 1st assessment stage in the Results section), a variance may be observed if we consider other parameters such as age and nationality, for example. For instance, in Q5 of the first stage of the assessment in the Results section, we presented the different perceptions of the users in their attempt to designate a feeling to their own composition, and we intend, in the future, to track down what triggers these feelings in the human agents. Also, we intend to develop an effective way to assess how well the users understood the system features (such as color/tone relation) in order to measure how well they learned and applied the system's translation approach (i.e. color to tone) in their production.

Past Bubble Sounds presentations. We have had the opportunity to present Bubble Sounds on digital ateliers at the Fine Arts School of the University of Minas Gerais (UFMG), and its sounding outcome has been acknowledged by many artists that had an experience with it, either by listening or interacting with the system, as an "ambiance music", a concept-genre defined by Eno [36] as a sound mass capable to properly mix with ambient sounds, creating soundscapes.

Since this is not a very widespread genre among music enthusiasts, given its experimental nature, it can, in some way, generate a feeling of "strangeness" in people more accustomed to the traditional musical model developed through the Western notation system. We intend to evaluate Bubble Sounds in musical performance environments, where they could actually be used as a musical instrument that works in harmony with other instruments, like in a band. At this stage, also due to limited human interactions, we only evaluated the system on its solo capacities, but it will be very engaging for us to evaluate it in musical workshops along with other voices. This kind of experimentation will certainly help us plan the next system's iterations.

Similar approaches. Similar goals to our meta-interactive proposal have been attempted using different AI approaches, such as machine learning. For instance, Wekinator [44] can generate a wide range of new instruments and ways to interact with them through gestural input (users can fully customize the way the compositional mechanic should work). Our approach, on the other hand, is more focused on allowing users to create coherent harmonic structures in a more "game-like" way, through a minimalistic interface.

Additionally, Crosscale [13] employed VR devices to propose a virtual musical instrument interface that captures gesture input from users and presents customized instrumental mappings to allow performers from different expertise levels to play complex songs. In this system, musical notes are arranged in a grid that adjusts the distance between notes across the scale, thus fostering interesting progressions. This system relates to Bubble Sounds in many aspects, but we highlight the interactive simplicity; both systems aim to improve the learning curve for users from different backgrounds to efficiently produce high-quality outcomes. The way in which both systems propose the experience is rather different, however, both in their mechanics to produce music and also in the way they are presented as a VRMI. Crosscale proposes a musical experience that resembles a piano, where the left hand is used for chords and the right hand is expected to play isolated notes, like those contained in a scale. In addition, this system uses a grid for the arrangement of notes, in the same fashion as it was discussed in the work in Chapter 4. On the other hand, Bubble Sounds uses the notion of projecting a musical score in a cylindrical structure and presents a metaphor of a vortex that matches the idea of musical loops with different tempos in each layer. It is also detached from the analogous mechanics of a common known musical instrument and goes towards a NIME (New Interface for Musical Expression), resembling, in this way, works such as Iwai's Tenori-On and Electroplankton [62, 97].

As previously mentioned, we believe gamified interactive models should be used to empower users in more playful ways, thus allowing for a more engaging experience that can guide users to accomplish complex tasks. In the sense of proposing more playful interfaces, the concept of playification [84] was developed and presented to address

this purpose and achieve more engaging results. However, it is not necessarily used as a tool to mitigate, engage, and extend users' productivity in regards to the creation of works of art, like meta-interactivity. Thus, we believe interesting outcomes may emerge out of hybrid approaches of these techniques combined.

Different layers of interactivity. We can observe that the meta-interactive approach occurs at many semiotic levels. For instance, we focus on a translation approach that transforms user proficiency by providing a playful interface resembling daily activities, like games, appropriating its mechanics and aesthetics to generate emergent music as a side effect.

The thesis addresses the lack of influence of proficiency on user perception, suggesting that the system's design enables users of all skill levels to produce coherent and enjoyable music. This might imply that the system generates similar outputs regardless of proficiency, potentially limiting originality. The study acknowledges this by suggesting the need to explore expert feedback on system limitations and assess whether users would use Bubble Sounds repeatedly, evaluating its value and potential improvements for both novice and expert users.

In Bubble Sounds, the dialogue between the user and the system is bidirectional; users "play" with bubbles and colors to produce music, and in doing so, they also create a musical corpus that emerges in reality through rhythmic hand clapping. Thus, user input in the real world alters the virtual state, while the virtual world influences physical gestures, transforming concrete reality.

Typically, users wear headphones when interacting with systems like Bubble Sounds, so the sound outcome is often limited to the individual. However, since clapping occurs in the real world, an observer unaware of the virtual context might find the rhythmically coherent clapping intriguing or at least curious, as it emerges from the in-game experience.

5.5 Conclusion

The meta-interactive approach demonstrates potential in enabling novice users to effectively engage and perform well when interacting with partially-autonomous systems through a playful and intuitive interface. This concept, reminiscent of historical artistic endeavors like Kandinsky's use of abstract forms to represent musical notes, discusses a paradigm in digital media, opening up expansive possibilities for innovative developments.

Our study revealed that individuals without formal musical training could generate music of appreciable quality using our system. The creations were validated by both experts and laypeople, who recognized these outputs as cohesive musical compositions.

To further enhance the system's capabilities, including the addition of multitrack compositions that encompass rhythm and melody in addition to harmony, we plan to conduct comprehensive testing with a broader demographic. This will particularly focus on users who have interacted directly with the system, to more accurately assess the system's efficacy as an image-based musical notation tool that nurtures user creativity without compromising their original artistic vision.

Further iterations of Meta-interactivity may be able to generate compelling dynamics of human-computer interaction, offering innovative pathways for interactive experiences and reinvigorating discussions on art, music, and games. Our findings confirm the indispensable role of human involvement in art creation. We are optimistic that this research will pave the way for designing future musical interactive systems that empower individuals in their creative pursuits, facilitating musical composition without constraining their expressive potential.

Chapter 6

Machine Autonomy

"I believe in creative control. No matter what anyone makes, they should have control over it."

— David Lynch

This chapter presents an autonomous approach that explores the dynamic generation of relaxing soundscapes for games and artistic installations. Different from past works, this system can generate music and images simultaneously, preserving human intent and coherency. We present our algorithm for the generation of audiovisual instances and also a system based on this approach, verifying the quality of the outcomes it can produce in light of current approaches for the generation of images and music. We also instigate the discussion around the new paradigm in arts, where the creative process is delegated to autonomous systems, with limited human participation. Our user study (N=74) shows that our approach overcomes a deep learning model in terms of quality, being recognized as human production, as if the outcome were being generated out of an endless musical improvisation performance.

6.1 Introduction

Recently, we have been exposed to a buzz of AI applications that dynamically generate visual art [102, 93, 106]. This emerging technological paradigm has profound implications in the way we perceive and interact with artistic pieces, and it also impacts various fields of HCI [61]. However, despite the excitement surrounding AI's ability to create meaningful artworks with limited human intervention, many aspects related to these works remain unclear. Questions arise regarding whether algorithmic creation can be considered art and how it affects the role of traditional

artists, particularly in the context of ownership debates surrounding AI-generated works [99]. The inherent association of artistic expression with human emotion raises concerns about the meaning and value of content created without explicit human involvement. Furthermore, understanding these phenomena is crucial for game designers, developers, and researchers as they grapple with ethical questions and seek to incorporate AI-generated art into their projects. For example, one question to consider is: **How do humans perceive and appreciate machine-produced artistic pieces when direct human intervention is not present in the creation?**

After all, in the core of works based on autonomous approaches, usually resides a goal for them to be acknowledged as a human-made production for an external observer. In this way, a common approach is attempting to simulate a specific endeavor, like representing the art style of a given musician or genre [109, 79, 35]. Therefore, machine learning techniques can be employed for a system to learn about the specificities of a given artist or artistic movement, and then this guideline will generate an artistic piece. This is for instance what happens in works using natural language processing (NLP) [92], proposing text input to generate diverse outcomes [2, 102, 93, 106].

Although these systems generate interesting outcomes, they are not able to fully support artists and game developers in creating original artworks for their own projects [35]. These systems often require significant post-production work to address the flaws in the generated outcomes [7]. Additionally, many artists lack the expertise to train and manipulate these models, as previously discussed in Chapter 5, which adds an additional barrier as they need to invest time in learning and adapting the technology to their creative processes. The lack of autonomy in defining aesthetic standards for the outcomes further limits the artists' control over the creative process, making it essential to explore new methodologies that empower artists and integrate their creative intent more directly.

Furthermore, it is important to note that previous studies have focused on developing dynamic asset generation approaches; however, many of them have not evaluated the perceptions of human evaluators regarding the outcomes produced by these autonomous methods. It is crucial to assess the representativeness of machine-produced content by examining how human evaluators perceive and distinguish between different autonomous systems. Understanding these perceptions can pave the way for broader applications of AI-generated art, extending beyond social media play artifacts and supporting the implementation of interactive experiences, such as games. Autonomous systems have the potential to contribute to game development by assisting with specific tasks, such as music composition or background scenario creation, providing valuable support to smaller teams facing challenges in these areas.

In recent works, **images** created by humans can be algorithmically turned into

music by machines (and vice-versa). Whether to adapt to a game's narrative changes or to suit changing moods desired by the developer [101, 123], this dialogue between image and sound instances requires, as expected, that one of these manifestations must pre-exist and come first, serving as ground truth. Thus, the rule is that initial stimuli must be presented, be it based on sounds, images, or text [16, 66, 46, 96]. However, the scenario becomes more intricate when the system is required to generate music and images simultaneously while preserving their inherent meaning. Usually, it is the human factor that defines the "mood" or "what goes well together", given a determined emotion (i.e. it parts from an abstract intent). Without a guiding intent that can serve as a translation guideline, such as an image parameter that influences musical quality, establishing coherence becomes challenging. As a result, the integration between audio and image risks becoming arbitrary from the perspective of human listeners. Additionally, there is a research gap in developing systems that seamlessly integrate and synchronize both music and visuals in real-time soundscapes and multimedia generation, rather than focusing solely on generating individual components.

Given the intricacies of audio and image manipulation, our work aims to explore autonomous approaches for the real-time generation of audiovisual instances. In this paper, we present a system design that addresses the challenge of simultaneous music and video generation while preserving coherence. Our approach establishes an audiovisual dialogue that empowers developers to maintain control and shape the desired outcome in terms of intent, aesthetics, and mood. Although our primary focus is on soundscape generation, we propose a versatile solution that can be implemented across various interactive applications, including game development and dynamic art generation. By utilizing this approach, developers can generate music, landscapes, or both simultaneously, ensuring a cohesive and harmonious audiovisual experience.

Thus, this work seeks to answer the following research question: **How can audio and images be generated simultaneously and coherently in an autonomous fashion, while preserving human expressivity?**

Orbiting the main question, we also attempt to answer the following questions:

- Are human evaluators able to identify nuances and distinguish between different autonomous systems with different levels of human involvement in the production?
- Do the outcomes generated by autonomous systems elicit pleasant responses from human listeners?
- What feelings do these outcomes generated by autonomous systems convey to human observers and listeners?

To tackle these questions, we introduce a Rule-based approach capable of managing both sound and graphic elements in order for a coherent outcome to emerge. Although autonomous, our approach still allows artists and developers to be active in the process of defining the mood and the aesthetics of the emergent piece. As an example of this approach, we also present Solato, an artistic installation that establishes a convergence territory for audiovisual and AI techniques for the emergence of meaningful outcomes. This approach differs from other works in the way it fosters the dialogue between audio and image manifestations for the generation of coherent audiovisual instances. Solato challenges the common notion that video is a primary element in audiovisual creation by generating music and images simultaneously.

We evaluate our approach in two stages, consisting of an ablation study to verify its efficiency in the generation of landscapes, and also a user study (n = 74) to compare the proposed Rule-based approach against 2 external baselines in the task of generating interesting soundscapes. These baselines consist of a Biased Random approach and also a Deep Learning approach that was trained to generate outcomes with similar goals as our system, also sharing the same dataset of musical and audio assets. We compare the outcomes generated by our approach and the baselines to evaluate the perception of human evaluators in their experiences. We also discuss how these different approaches can contribute to this work's goal, emphasizing the pros and cons of each one. Finally, we discuss the feelings each model conveyed, as acknowledged by human evaluators, which allows us to glimpse many ethical-related questions that surround autonomous production, such as the ability of machine production to resemble human creation.

Our study revealed the following findings:

- Our approach achieved the goal of generating coherent outcomes.
- Solato provides a unified experience, leading human evaluators to appreciate the cohesive composition of music and images, which they perceived to be a human creation.
- Our study showed that different systems conveyed different feelings; therefore this work offers insights into the creative process of artists, designers, and game developers.
- Our study emphasizes that rule-based procedural content generation (PCG), despite the advances in Deep Learning approaches, continues to be an effective method for fulfilling this research's objectives. The outcomes suggest that utilizing such techniques in artistic creation enhances human participation, ensuring that humans remain integral to the creative process.

This chapter's contribution can be summarized in the following ways: (i) An algorithm for the simultaneous generation of coherent audiovisual instances; (ii) An in-depth discussion and evaluation of past and current approaches for dynamically generating audio and video assets and the emotions they evoke; (iii) Reflections on the human factor in autonomous AI-based systems applied in games and art, and how humans perceive the creations of such systems. Thus, the most important contribution to HCI lies in the development of three novel systems that dynamically generate soundscapes, evoking specific feelings, and their user-centric evaluation, revealing the potential of emotionally engaging interactive experiences. In this way, this work tries to fill a gap by analyzing how different autonomous approaches using PCG, Rule-based mechanics, and Deep Learning techniques can operate in the audiovisual arts and also what is the human perception of it. In addition, besides evaluating the mechanics of our approach, we also evaluate the outcomes it can produce.

6.2 Landscape and Music Generation

In this section, we discuss our algorithm for the generation of music and imagetic instances coherently. To achieve this, our approach combines rule-based mechanics and procedural content generation (PCG) techniques. We have developed an algorithm that facilitates the real-time generation of audiovisual instances, maintaining coherence between the music and images. The algorithm considers various factors, such as mood, aesthetics, and intent, allowing developers to have control over the desired outcome.

We can understand **music** and **image** as different stimuli that find convergence in **audiovisual** works, generating a third entity that contains its own meaning, detached from its original manifestations. In other words, the convergence of sounding and imagetic instances fosters the creation of a new form of artwork that may even convey different emotions, different from the original audio and sound manifestations that originated it. Usually, it is a human intent that defines a soundtrack that extends or potentializes what the image is trying to convey. Therefore, it is challenging to conceive a system whose set of rules is "generic" enough to be able to generate both images and music coherently, especially if it is expected a "dialogue" between these music and video manifestations in a sense of creating meaning. As previously discussed, past works use an initial stimulus for the generation of images or sounds [16, 66, 46, 96]. That is, essentially, one needs to pre-exist and serve as an input for the other to be generated. Our approach, on the other hand, is able to generate a solution to this problem, although it still has limitations, as we will discuss further on.

The process for generating sounds and images is usually quite different. Be it in theoretical academic works or in practical audiovisual production, it usually requires

researchers and professionals with different backgrounds and expertise to manipulate these instances. In the case of our approach, however, the same algorithm can be employed for the generation of both sound and imagetic outcomes in parallel executions, as shown in Figure 6.1. Therefore, its usage can be generic in its application, allowing artists to appropriate from it and still keep control of the generated outcomes. This process does not exclude humans from the creative process. Instead, it manages the content of the different datasets. In this way, game designers and artists still have control over the outcome's aesthetic if they want to create their own dataset assets, although they might as well use public assets if they do not desire to have partial control of how the outcomes will be assembled. The designer/artist will feed the system's datasets of music and image samples for it to generate landscapes and music, and, ultimately, a soundscape. In this way, differently from the current approaches [102, 93], this system does not work as a black box, where the designer/artist does not see what is happening – it is accessible to anyone, whether or not they wish to use their own assets to maintain control over the aesthetics or just use free or paid asset packages available online.

The analogy is that a development team might have visual assets of trees, houses, benches, telephone cabins, etc, however, they will not have a city. Our approach offers support in this endeavor; that is taking these assets, and generating coherent "clusters" or "arrangements" of them in a harmonious way. In the same way as for music, having a dataset of isolated musical note samples (e.g. piano samples of each key isolated, such as C, C#, D, D#, E, etc) does warrant a composed melody for a song. This approach intends to address the challenge of arranging these groups of notes into a pleasant musical corpus, that also dialogues and composes a harmonious relationship with the visual instances generating a human-like created soundscape.

The coherence between imagetic and sounding instances comes mainly from the way the building blocks for the generation of images and music are being proposed, as we will present in detail.

For a better understanding of the mechanic elements that we will discuss in more detail ahead, we will define the terms **"building block"** for addressing the design patterns, or grids, that were pre-created in order for music and images to emerge (these elements determine the different shapes of the "modules" of the environment's floor in which the 3D assets are instantiated upon); **"visual assets"** for the 3D art generated by human artists and that represents elements of the experience's aesthetics in our visual dataset; and **"audio assets"**, that relates to the sound samples generated by human designers that simulate instruments in our musical dataset and that are generated note by note by the system; and **"palette"**, that consists of nonvisible objects that trigger the instantiating or removal of musical and visual assets over time.

Landscapes and music are generated as follows: the building blocks, as mentioned

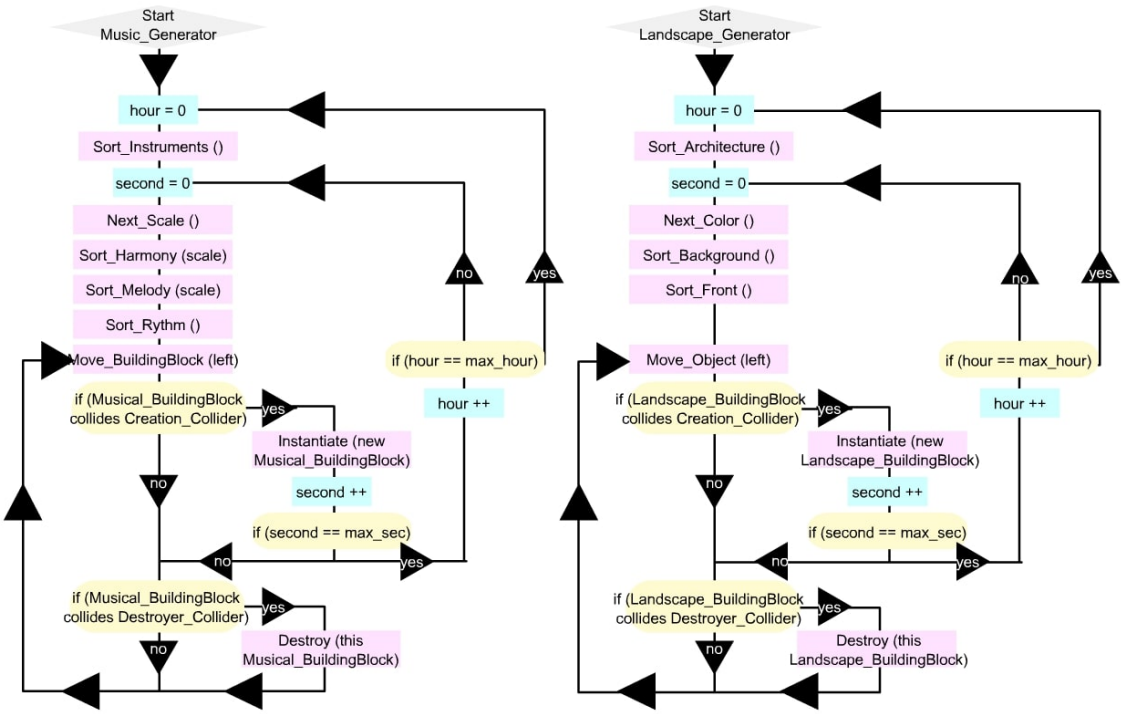


Figure 6.1: Diagrams for the music generator (left) and for the landscape generator (right) that foster the emergence of soundscapes.

above, are human-made instances that dynamically assemble visual and audio assets over time. It could be comprehended as a filter that otherwise would make the system work in a random fashion. These building blocks consist of **Musical** and **Landscape** types. Developers can create and add new shapes of building blocks according to the needs of their own experience, and also customize many aspects related to the generative process (e.g. time, frequency, visual assets size, etc). The intention is to make both the music and the landscape have so many variations to a point that each iterative loop will never look or sound repetitive.

In the experience we developed based on this approach, which will be presented ahead, we generated 8 building block shapes for Music and 8 building block shapes for Landscape (see examples in Figure 6.2.C and Figure 6.3.C). However, as mentioned above, artists and developers can create and customize the characteristics of their building blocks at will (e.g. add more shapes, customize cell sizes, etc).

Musical building blocks are allocated in 3 positions: the harmony (in the middle of the grid), the melody (in the back of the grid), and the rhythm (in the front of the grid), as shown in Figure 6.2.A. Each position (i.e. back, middle, and front) contains

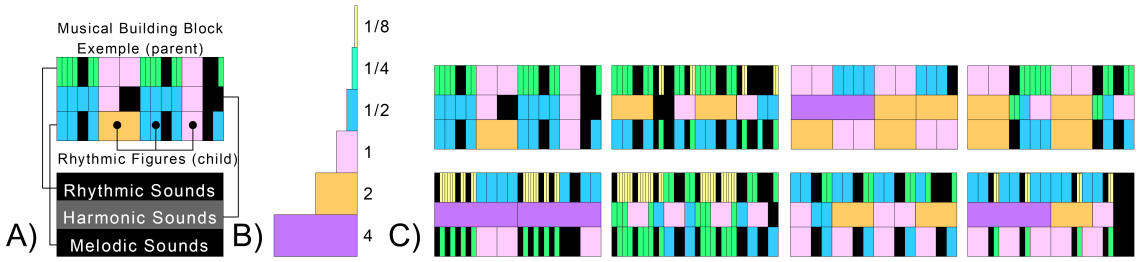


Figure 6.2: 6.2.A) The Music building block. 6.2.B) The box triggers' sizes represent rhythmic figures. 6.2.C) 8 pre-determined Musical building block patterns completed with rhythmic figures presented according to the system agency.

several trigger boxes (which work like "colliders" to detect objects overlap) of different sizes, represented by 6 different colors. These colors correspond to different rhythmic figures of a compass, such as 1/8, 1/4, 1/2, 1, 2, and 4, as shown in Figure 6.2.B. The colliders trigger different note durations, and the arrangement of these different durations produces a rhythm of the emergent music. In the same way, it also dictates different melodic and harmonic patterns. In this way, the building blocks are being played from left to right as if a virtual line (i.e. palettes, as we will define ahead) also moves from left to right, and all boxes in the same column are played simultaneously.

Each position represents a different instrumentation in the performance (which can also be comprehended as a different "voice" in this analog "band") and has its own dataset of samples. For instance, in the system we created based on this approach, we used digital piano samples for the harmonic part, piano for the melodic part, and a spatialized sample sound texture for the rhythmic part. However, artists and developers can customize their own musical datasets with different music samples (e.g. add different piano textures, beat sounds, drum samples, harmonic voices, etc).

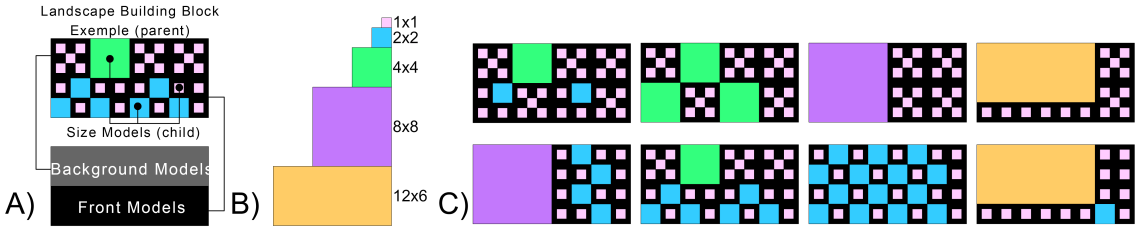


Figure 6.3: 6.3.A) The empty Landscape building block. 6.3.B) The sizes of the 4 types of 3D assets that the system can generate. 6.3.C) 8 pre-determined patterns for landscape generation that are presented by the system.

Landscape building blocks are divided into 2 different groups: **front model sizes**

and **background model sizes**. Background model sizes are larger so they can be visualized from a long distance, as shown in Figure 6.3.A, while front model sizes are smaller so they do not obstruct background objects. Each landscape building block model size draws a 3D asset of its corresponding size, as shown in Figure 6.3.B. While the 3D models (see Figure 6.7 for some examples from our own system's dataset) are instantiated in predetermined positions, as shown in Figure 6.3.C, the system randomly selects 4 possible rotations (varying in 90 degrees) to instantiate each of the front 3D models in the dataset, providing more visual diversity.

There are two palettes, one that creates and one that destroys both Musical and Landscape building blocks, shown in Figure 6.4.B. The palette that creates randomly generates pre-defined Musical and Landscape building blocks from each theme dataset. On the other hand, the palette that removes Musical and Landscape building blocks works by destroying the building blocks that are not being rendered in the scene to make the experience more fluid, reducing its computational cost. These palettes are invisible instances in the 3D environment; they can be comprehended as colliders that trigger the instantiation or destruction of the dataset assets, be it from the musical assets or image assets repositories.

In this approach, the virtual environment, composed of an arrangement of building blocks, moves from the right to the left direction, as shown in 6.4.B. Therefore, the palettes and the camera stand still, while the grids move towards the instantiating and deleting objects (i.e. the palettes). It was generated like this in order to address the convention of some platform games, in which characters usually move in this fashion (e.g. in games of the endless run genre and also in many classic games). In addition, in the experience we will present ahead, our goal was to simulate this kind of movement. However, it is important to mention that developers can customize this at will, including dynamic changes of direction (e.g. make the scenario move forward and backward).

It is also important to emphasize that although the system is autonomous to generate music, landscapes, or full soundscapes, there is a level of participation of humans in the creative process. The datasets of assets are fully customizable by artists and developers, who can decide whether to develop their own assets, purchase them, or get them from open libraries, as mentioned earlier. In addition, developers can also customize their building blocks, which will define many different aspects of the landscape that is being generated (e.g. size, layers, density, etc). Therefore, our work differs from current approaches for the dynamic generation of sound and visual instances based on high-level input instructions through keywords [2, 102, 93, 106, 96].

An accomplishment of our approach is to promote, through the same database management algorithm of artistic assets, the generation of sound and visual instances, as shown in Figure 6.2 - A and B and Algorithm 2. This means that the same process

used for the generation of music can be used in the generation of landscapes, therefore the generation of soundscapes can happen through a single parallelized process. Given the differences and intricacies that need to be handled in the process of generating audio and visual elements, our approach addresses a complex problem and offers a promising solution to cope with the dynamic emergence of soundscapes. In this way, this procedure is capable of addressing the problem of generating audio and video instances simultaneously (as we will present further on), however, it can also be used separately to generate music and background for games and other applications.

6.2.1 SOLATO

In this section, we present Solato (Soundtracks & Landscapes Tour), an autonomous approach inspired by the intent of exploring the dynamic generation of soundscapes, using the algorithm presented in the previous section. Solato is an artistic installation that generates its sounding and imagetic parts in a dynamic fashion, using a peculiar game-like colorful aesthetic. The experience was developed using the Unity engine (ver.5.4.6) and coded with C#. The 3D assets were developed using Blender (ver.3.0.1), and the sound samples were produced using FL Studio (ver.12).

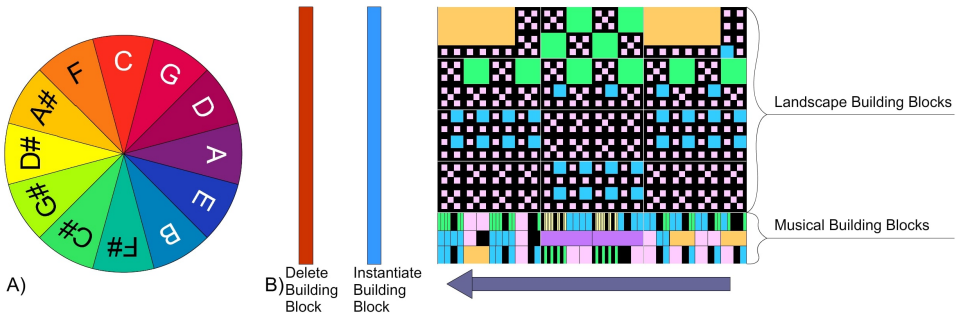


Figure 6.4: 6.4.A) A Color-Tone clock that shows the relationship between musical tones, colors, and hours. 6.4.B) The deleting (red) and the instantiating (blue) palettes of the Musical and Landscape building blocks. These nonvisible objects are responsible for instantiating visual assets that are about to enter the camera frustum and also deleting them when they are off the camera range.

The experience takes place as if the interlocutor (i.e. anyone engaged in the experience) is traveling by train and watching ethnic-urban landscapes through the window of its cabin. Our goal was to produce an interactive installation based on existing relaxation videos on streaming platforms, however, fostering a greater immersive experience.

These landscapes are dynamically generated by our algorithm infinitely. In the

same way, a musical corpus also emerges dynamically, and both of these instances (i.e. audio and visual) are supposed to coherently dialogue for the emergence of the experience's meaning. However, the system is not deterministic, and at each run, a different environment and sequence of themes are presented. Even the authors cannot predict how the music nor the virtual environment will come together, since the system is autonomous to generate meaningful outcomes. In this way, at each new execution, a new experience emerges, which is then signified in its own way by each person. Therefore, it is intended that the audience becomes an active part of the installation, giving it a personal view and meaning. Although the audiovisual instances that emerge from the system present a level of unpredictability, it does foster pleasant and coherent structures for both the musical corpus as well as for the virtual environment aesthetics.

The poetic of the experience is centered on the idea of allowing the individual to engage in a relaxing "oniric virtual tour" around the world. Some of the feelings conveyed by the system, as observed in an early version of the prototype, were the feeling of relaxation (as proposed by many video music streamings to this end); the feeling of departure, of leaving what is known and comfortable towards the unknown; and the feeling of adventure, embarking on a journey throughout the world. However, a more robust evaluation was performed to assess the feelings that the system conveys, as we will present ahead.

Figure 6.5: Screenshots of 6 different day and night moments of 3 different themes of Solato.

At each system's new run, a new theme (containing 3D assets and instrument samples specific to that theme) is generated. We currently have 3 completed themes. A complete loop of themes in the experience currently takes 72 minutes (24 minutes for each) to be concluded. However, it is not guaranteed that the expectant will see

the themes in a row since the system randomizes the order. A given theme may reoccur, although never exactly the same.

We simulate the different hours of a day through a color scheme (shown in Figure 6.5) that is presented through a color-tone *clock*, shown in Figure 6.4.A. This clock controls the change of periods of the day (i.e. color moods) and also the key that determines the scale that will be used for the system to create a musical corpus. Thus at each cycle of 12 interpolable colors, a new day cycle begins, with its own music. The transitions between these colors are smooth, in this way in the colors of the red spectrum, for example, you have a notion of a sunset. In the same way, it also generated harmonious interpolations of scales, following a sequence of fifths. Each hour of the day takes 2 minutes to be completed in real life.

For interested readers, a video demonstration showcasing Solato is available at `https://youtu.be/VexND-gZc38`. Figure 6.6 presents a timeline of screenshots illustrating the demonstration of the Magenta/NightCafe system as showcased in the accompanying video.

Figure 6.6: Screenshots of the Bubble Sounds video.

6.2.1.1 Image Presentation

As presented, Solato uses a color aesthetic to simulate different hours of the day, as shown in Figure 6.5, controlled by a color-tone clock, shown in Figure 6.4.A. The system uses a shader customized by the authors that renders the 3D objects emphasizing their edges. This is done by setting the fog (resource available on 3D engines for improving performance) with no gradient. In this way, the first plan of 3D assets is not involved by the fog, whereas the 3D assets on the back are. However, the fog does not apply to the edges of the objects; therefore, 3D assets are always visible in a "cartoonish" style. The first plan rendering system uses a method that simulates a drawing being filled with ink as the object enters a certain range in the frustum of the camera.

All assets in the scene are in 3D; there is no 2D asset being used despite the background looking like a drawing. There are 4 types of 3D assets with different sizes in our dataset. Currently, there are 32 of the 1x1 type (e.g. benches, light poles, etc),

16 of the 2x2 (e.g. telephone cabins, statues), 8 of the 4x4 (e.g. water well, kiosks), 4 of the 8x8 (e.g. churches, temples) and 2 of the 12x6 (e.g. pyramids, drawbridges, etc) for each theme, as shown in Figure 6.3.B and 6.7, all of them corresponding to the theme's aesthetics and architectures (e.g. Egypt's theme have pyramids and sphinx). Thus, we currently have developed 90 assets that are contained in our dataset for the system to manage (some examples shown in Figure 6.7). As a reference, the 1x1 size corresponds to a $1m^2$ in real life.

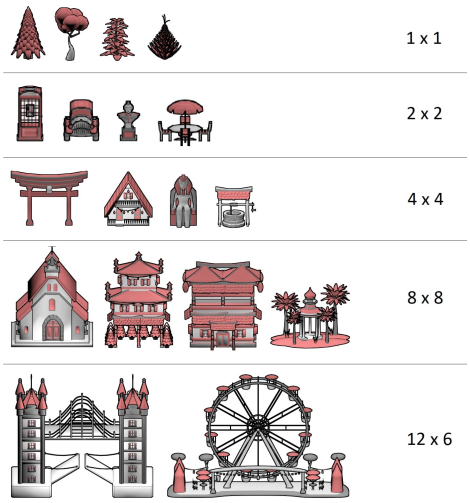


Figure 6.7: Examples of low poly 3D assets from different themes managed by the system.

Bigger objects (i.e 12x6) are always instantiated in the back, while smaller objects (i.e. 1x1) are always instantiated in the front, as shown in Figure 6.3. In this way, we guarantee the emergence of pleasant landscapes, where bigger objects do not obstruct the view of smaller assets.

6.2.1.2 Sound Presentation

Similarly to what happens for the generation of landscapes, the musical keys shown in the color-tone clock represent the scale the system will use to generate a melodic line. More precisely, it dictates the current tone of a minor diatonic scale which will be the basis for the generation of the melody. The clock controls the time-passing simulation occurs in a sequence of fifths; that is, if the current hour is defined by the Cm key, the next one will be a Gm tone/scale. Thus, the note presented in the clock, as shown in Figure 6.4.A, does not determine which note is being played; it dictates the scale over which the system will improvise over. For instance, the representation of the early stage of the night (midnight) is determined by the Am key, the color

purple. The notes in which the system will generate the melody, in this case, will be *(A, B, C, D, E, F, G)*. Thus, these notes will be present in the melody that the system will generate for the emergent music's melodic lines.

As for the harmony, the system will use the same set of notes to generate chords, that is, 3 notes being played simultaneously among all contained in the C scale, to create the harmonic lines. For the generation of chords, the system will randomly choose the first note within the scale (i.e. the fundamental). As for the second note of the chord, the system will choose a second, third, or fourth degree from the fundamental (e.g. "C" (fundamental) = degree I, "D" = degree II, "E" = degree III) from the same scale. For the third note, the system will choose a fifth (degree V), sixth (degree VI), or seventh (degree VII), following the same logic.

The rhythmic section is defined by the size of trigger boxes in the building blocks, as shown in Figure 6.4.B. In this way, as the palettes that instantiate 3D assets in the grid pass by different shapes in the building blocks, which in turn interacts with 3D assets of varying sizes, as shown in Figure 6.7, generating rhythmic patterns of different latencies (in music, latency is the delay between when a musical action is initiated and when it is heard, which can affect timing and synchronization).

Each theme has its own instrument samples, based on ethnic aspects of different places around the world. There are currently 1 rhythmic, 1 melodic, and 12 harmonic instruments. The harmonic instruments are generated considering 4 octaves with 12 musical notes (48 samples for each harmonic instrument), necessary for the formation of chords. The melodic instruments, however, only need 3 octaves with 12 musical notes (36 samples for each melodic instrument). There are, currently, 576 harmony samples, 36 melody samples, and 3 rhythm samples in our dataset.

6.2.1.3 VR Adaptation

As stated earlier, our concept was inspired by relaxing music and videos available on streaming platforms. However, Solato was developed as an art installation that also works with a VR device. There were some challenges to overcome in order to have the system prepared for a VR experience. Since our system generates the landscapes in a specific axis in the virtual environment (e.g. the scenarios are being generated along the x-axis as the camera moves in the virtual world), we had to constrain the view of the user to the perspective in which the landscapes are being generated.

As previously mentioned, the experience was designed as if the user is inside a train cabin, contemplating the generated landscapes through the train's window. In this way, the 3D model of the train cabin provided a coherent solution for displaying only the part of the scenario that was interesting to our proposal, as shown in Figure 6.8. As the users stare all around the virtual environment, they will observe details of the interior of the train cabin, such as the train seats, doors, and luggage racks.

Algorithm 2 Music generation (left) and landscape generation (right). Note the similarity in the approaches despite generating distinct instances, showcasing the versatility and adaptability of the algorithm.

```
1: function OnTriggerEnter(collider)
2:     if collider.name == "Creation_Collider" then
3:         Instantiate(Musical_Object, position, Quaternion.identity)
4:         second += 1
5:         if second == 120 then
6:             hour += 1
7:             if hour == 24 then
8:                 Sort_Instruments()
9:                 hour = 0
10:            end if
11:            Variables()
12:        end if
13:    end if
14:    if collider.name == "Destroyer_Collider" then
15:        Destroy(this.gameObject)
16:    end if
17: end function
```

```
1: function OnTriggerEnter(collider)
2:     if collider.name == "Creation_Collider" then
3:         Instantiate(Landscape_Object, position, Quaternion.identity)
4:         second += 1
5:         if second == 120 then
6:             hour += 1
7:             if hour == 24 then
8:                 Sort_Architecture()
9:                 hour = 0
10:            end if
11:            Variables()
12:        end if
13:    end if
14:    if collider.name == "Destroyer_Collider" then
15:        Destroy(this.gameObject)
16:    end if
17: end function
```

However, when looking at the window, the generated landscapes can be observed passing by seamlessly.

6.3 Qualitative Analysis

Before we evaluate the generation of soundscapes through the approach presented in section 6.2 with human evaluators, we conducted an ablation study (which is an analysis in which specific components of the system are selectively removed to help understanding their impact) to evaluate the Rule-based approach in the specific task of generating coherent landscapes. We conducted a series of Solato runs employing the approach presented in Figure 6.1 and also without it, in a randomized fashion, observing how the system behaves in the task of generating harmonious and coherent landscapes. In this session, we did not evaluate our algorithm in the task of generating

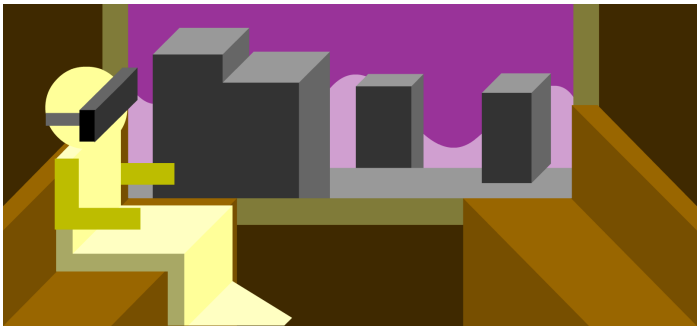


Figure 6.8: Mockup of the train cabin that constrains the user view of the landscape to the desired perspective.

music, due to the complexity and unclarity of music visualization (e.g. spectrograms and their signal strengths) and the difficulty in defining a sound quality standard that could be purely observable. However, we did evaluate the efficiency of the system in the creation of interesting music through the perception of human listeners, as we will present in the next section.

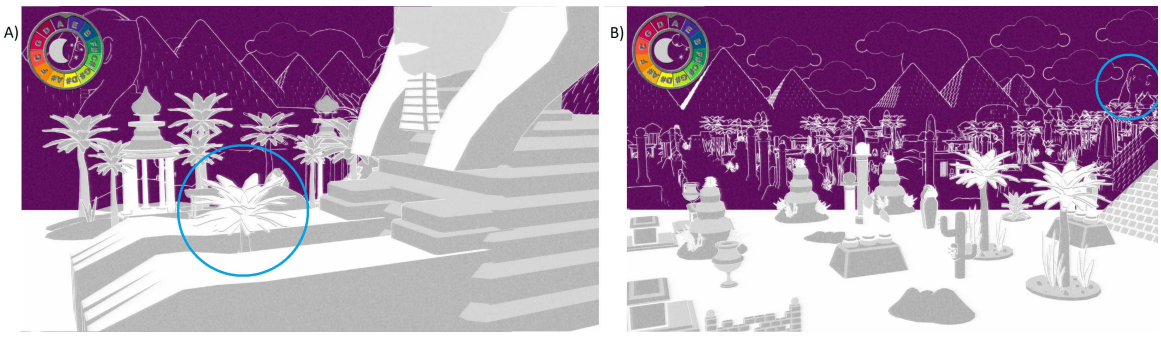


Figure 6.9: Comparison of landscape generation outcomes with and without our approach. The image on the left (without our approach) highlights issues such as overlapping 3D assets and visual clutter. In contrast, the image on the right (with our approach) demonstrates a more harmonious visual composition, with no asset overlap or 3D objects obstructing the background.

By running the system using the algorithm presented in Figure 6.1 and 6.3, it is possible to notice the effectiveness of the Rule-based approach in the generation of compelling landscapes. In the detail highlighted inside the blue circle shown in Figure 6.9.A, we show that the generation of the scene **without our approach** is disturbed by some glitches, such as overlapping of 3D objects. In this particular case, a sphinx, which is a 12x6 object, was instantiated along with foreground objects, overlaying the 3D asset of a coconut tree. As presented in Section 4, it is important to note that, in

the Rule-based approach, small 3D assets such as the coconut tree (i.e. 1x1 in size) only appear in the first layer plan, not to obstruct the urban landscapes.

As for the system execution using our approach, as shown in Figure 6.9 (blue circle in the image's upper right), it is possible to see a sphinx in its proper place, concentrated in the further layer of the background, as all the major 12x6 3D objects are. In this way, major objects do not disturb the harmony of the image obstructing the generated landscape, allowing for a "cleaner" visualization of the scene.

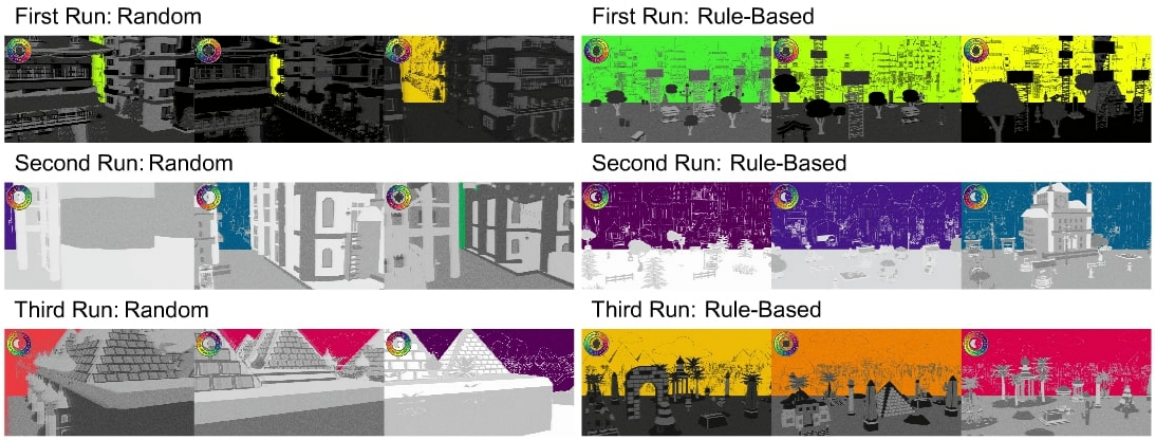


Figure 6.10: Timeline of 3 different runs of the system without our approach (left) and with our approach (right).

To provide better visualization, we generated a timeline of 3 executions of the system running using our approach, presented in Section 6.2, as shown in Figure 6.10, and also without it (which worked in a randomized fashion), for a glimpse of how the system manages the creation of meaningful landscapes. Each timeline presents 3 images captured at different daytimes (as represented in the experience), and we can clearly observe that the execution employing the Rule-based approach was able to generate more harmonious outcomes. Whether as a standalone relaxing experience or for the generation of the background for games, the Rule-based approach showed great potential to be utilized as a resource to foster unpredictable, coherent, and aesthetically interesting (although specific) effects.

6.4 Quantitative Analysis

Besides the evaluation of Solato, we also evaluated 2 variations of the system that uses different methods for the generation of its images and sounds. The first variation consists of a Solato version working under a Biased Random mechanic. The second

variation consists of two different AI approaches, one for the generation of audio and the other for the generation of images, that were employed to assemble these instances into a single and homogeneous experience.

For addressing the music, we used a Deep Learning approach based on Google Magenta [30]. For the generation of images, we used the NightCafe Machine Learning approach [106]. For the musical generation, specifically, a piece of music generated through Solato was employed as a training input for Magenta. In this way, our goal was that an external baseline was still related to our approach, to minimize differences between systems and ensure a reliable scenario for evaluation. Thus, we have **Solato**, a **Biased Random Approach**, and a **Magenta/NightCafe Approach**.

The 3 systems use the same dataset of audio samples to guarantee a fair comparison between the approaches, not allowing for aspects such as "sound textures" and timbres to generate noise in our results. Currently, to the best of our knowledge, we could not find a commercial system that fosters a similar experience as ours to the point that it could be used in our evaluation. Therefore, we created these baselines. To do so, besides preserving a link with our approach, other important criteria needed to be fulfilled: these baselines should be based on a system that dynamically generates both images and videos in an autonomous fashion; otherwise, there would be differences in timbres that could affect the evaluation. Hence, we present the differences and peculiarities of the systems below.

Solato: This is the system we developed using the approach presented in Section 6.2 and discussed in detail in Subsection 6.2.1. In this system, the building blocks that produce the music and the landscape were generated by human artists to ensure greater control over the outcomes. Therefore, although autonomous for the generation of outcomes, this system allows a more participative presence of humans in the emergence of landscapes.

Biased Random: In this version, the building blocks that generate music and landscapes have their internal shapes randomly generated. However, there is a bias in the music generation since the grids (i.e., building blocks) contribute to creating rhythmic figures and ensuring variations in the music and images. Within the parent building block, the same music and landscape pieces are randomly placed, which can cause accidental overlaps between 3D models when larger pieces occupy positions meant for smaller ones (as detailed in our ablation study in Section 6.3). A musical building block generates three positions for random rhythmic figures: rhythm, harmony, and melody. A landscape building block has four horizontal positions for different object sizes, and when a large piece is placed, the block moves to the next horizontal line to generate objects. A music video demonstrating the Biased Random approach is available at `https://youtu.be/-TpTsiq8cVo`. Figure 6.11 presents a

timeline of screenshots illustrating the demonstration of the Biased Random system as showcased in the accompanying video.

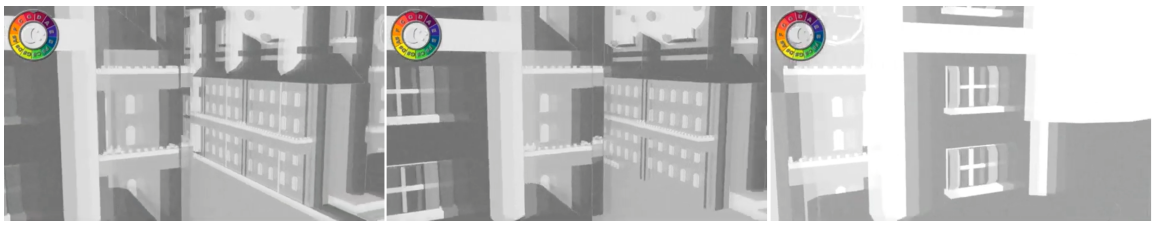


Figure 6.11: Screenshots of the Biased Random video.

Magenta/NightCafe: To make our experiments more robust, we also evaluated a different approach for audiovisual production. Since we could not find a system that proposes that exact effect as ours – that is, generating both images and sounds simultaneously in real time for an artistic end – we followed a procedure to create an external baseline. The procedure for creating this system is described as follows: First, we recorded a music sample generated by Solato in MIDI format. This MIDI track structure can be visualized in Figure 6.12 (A). Second, we used the Deep Learning approach MidiMe [30] based on Google Magenta to train a synthetic agent to generate a new melody based on the music generated through our system. From the music we created through Solato, MidiMe extract its melodic line, as shown in Figure 6.12 (B). Third, we trained the model and generated a 2:30 minutes long music loop, also in MIDI format. The reason why we chose this specific video length will be clarified at the beginning of the User Study section.

In Figure 6.12 we can visualize and compare how the original melody used as training data (A) and the resulting melody (B) are structurally related. In Figure 6.12 (C), we show how well the model was trained in reconstructing the original music (the closest to 0, the most accurate it was). After this process, we generated the final music using the same instrument samples we used in Solato, and then converted the MIDI file to an MP3 format.

Our goal in generating this external baseline was to create a work that is not only novel and interesting but also establishes a clear link with other approaches. This baseline utilizes the same audio dataset for music generation and produces colorful graphics that loop in a somewhat unpredictable manner. This approach ensures a more sophisticated musical sample from the trained version, enabling a fair comparison between systems. For the landscape component, we employed NightCafe Studio [106], a recurrent neural network system designed for generating visuals based on keywords. After conducting empirical experiments, we selected keywords – "Landscape", "Musical", and "Colorful" – that align with the Solato

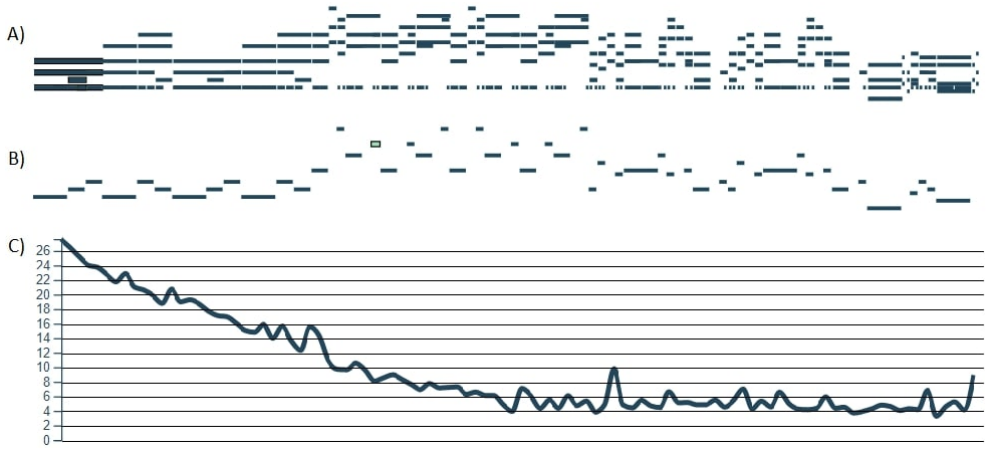


Figure 6.12: A) The original music generated through Solato. B) The melodic line extracted by the original music. C) The error rate demonstrates how well the trained model reconstructed the original music melody.

experience. Using NightCafe Studio, it was generated images, and through further experimentation, we created a small video that replicates to match the 2:30 minutes of the music. This external baseline serves various purposes, such as providing a benchmark for comparison, showcasing diversity, and contributing to the overall evaluation framework. The choice of keywords and the visual generation process ensures that the generated content aligns with the Solato experience, creating a cohesive and immersive multimedia environment. A music video demonstrating the Magenta/NightCafe approach is available at `https://youtu.be/XjPFW6ONge8`. Figure 6.13 presents a timeline of screenshots illustrating the demonstration of the Magenta/NightCafe system as showcased in the accompanying video.

Figure 6.13: Screenshots of Magenta/NightCafe.

6.4.1 User Study

In our assessment sessions, a total of 74 human evaluators participated and provided feedback through an online form. The evaluators consisted primarily of graphic design

undergraduate students and game developers from FUMEC University, artists from the School of Fine Arts at UFMG, and computer science students from UFMG, all based in Brazil. The form presented three music videos, each corresponding to the three systems discussed earlier in the previous section: **Solato**, **Biased Random**, and **Magenta/NightCafe**. These demonstration music videos were carefully curated, with each video being 2 minutes and 30 seconds in duration. They were divided into three distinct category sessions, aligned with the three systems under evaluation. To ensure a comprehensive evaluation, the music videos for Solato were specifically chosen to capture different time points within its day/night cycle. This selection was made because Solato generates unique music and soundscape variations in each run. This deliberate choice allowed us to showcase crucial details and noticeable differences between the outcomes produced by each approach, effectively highlighting their respective characteristics. Furthermore, considering the time constraints in the evaluation environments, we conducted pilot tests to determine an appropriate duration for each video. Based on these tests, we determined that 2 minutes and 30 seconds provided enough time to demonstrate the systems and their variations effectively.

The evaluators in the sessions had an average age of 24 years and were selected based on having at least a basic knowledge of sound design. Furthermore, it is worth noting that all evaluators possessed music knowledge and skills, as they had completed courses in sound design. This suggests that they had a certain level of understanding of music theory, composition techniques, and the technical aspects of sound design. Additionally, the evaluators' expertise in music composition implies their familiarity with human expressivity in music, indicating that they should be able to effectively evaluate and comprehend the emotional and artistic elements of the compositions under study. As students of art and technology-based courses, they may be able to effectively evaluate the current paradigm of AI approaches such as MidJourney [93] and its implications in arts. These observations emphasize the evaluators' qualifications and competence in assessing the impact of sound design techniques on the expressivity and perception of musical compositions.

Since the evaluation sessions were conducted remotely, the participants did not experiment with the system in its VR capacity.

Each user agreed and marked a consent form, with a clear explanation of the whole process. None of the evaluators had any experience with the systems prior to the evaluation sessions, and no sensitive data from any of the users were recorded. This study was approved by the ethics committee of the Faculty of Science and Technology at Lancaster University.

Evaluators answered the 5 questions presented in the questionnaire below (see Table 6.1) after the presentation of each video. The order of the video presentation and its respective sections in the forms was randomized to avoid any bias.

Q1.	How do you classify the music? Very uninteresting (1) to (10) Very interesting
Q2.	How do you classify the video? Very uninteresting (1) to (10) Very interesting
Q3.	How do you classify the composition between music and image? Very dissonant (1) to (10) Very consonant
Q4.	"The video clip was generated by a computer, not a human." Do you agree with this? Completely disagree (1) to (10) Completely agree
Q5.	Which feeling predominantly describes your experience with the video clip? [ ] exciting, fun; [ ] relaxing, calm; [ ] gloomy, melancholic; [ ] aggressive, hectic;

Table 6.1: Evaluation questionnaire for the 3 systems.

Due to time constraints, we could not run the experience in its integral length. The experience currently takes 72 minutes to generate a loop between all the themes it currently has. In this way, evaluators only had contact with the videos of each system to judge the questions asked in Table 6.1 for each of the 3 systems; thus, for each individual, we asked 15 questions in total. The questionnaire presented multiple choice questions with answers varying on a linear scale from 1 to 10 in questions 1 to 4. As for question 5, we presented a visual scheme (shown in Figure 6.14) showing the feelings presented in Q5 of Table 6.1 for evaluators to choose from.

Also regarding Q5, our goal was to ensure simplicity and clarity for the evaluators when assessing the emotions conveyed by the systems. We acknowledge that evaluating emotions can be complex and subjective, and the previous study in Chapter 5 has recognized that similar feelings can be categorized differently. In our study, we drew inspiration from the GEMS system [78], which provided an interesting approach for emotional assessment. However, considering the specific focus and requirements of our study, we found that the GEMS system could potentially confuse our evaluators due to its extensive range of emotions that could be interconnected. To address this concern, we conducted pilot tests to gain insights into the possible range of feelings that our system was likely to convey. This allowed us to refine and narrow down the spectrum of emotions we aimed to evoke. Through these internal lab experiments, we identified the most commonly observed feelings and organized them into small groups of abstract nouns. Consequently, we developed a straightforward model based on four main pairs of organized feelings: *exciting, fun; relaxing, calm; gloomy, melancholic; aggressive, hectic*, as presented in the questionnaire in Table 6.1. This model enables us to accommodate a wide range of interpretations, particularly for Solato, which seeks to provide a creative and relaxing experience.

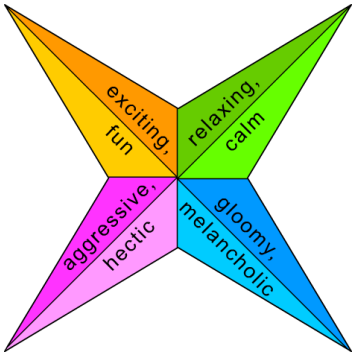


Figure 6.14: Chart available in Q5 of the evaluation questionnaire, regarding the main feeling conveyed by the system.

6.4.2 Results

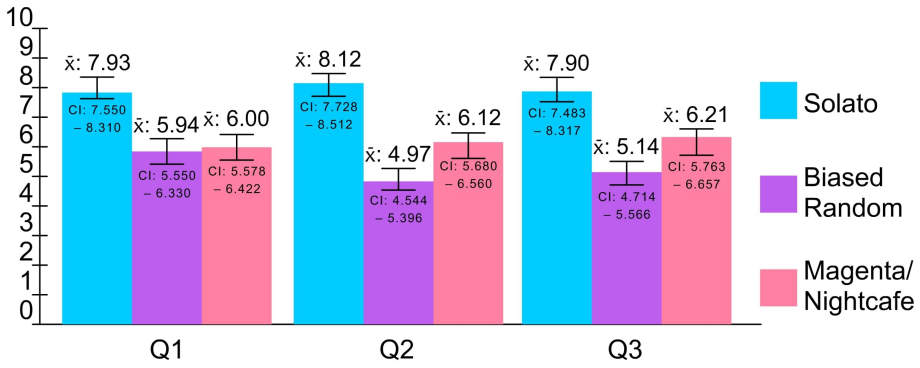


Figure 6.15: Means obtained by each system in Q1, Q2, and Q3.

In **Q1**, evaluators rated the music generated by the 3 systems according to their perception of quality. In **Solato**, we observed a $\bar{x} = 7.93$ (SD: 1.67), showing a satisfactory perception of musical quality. In the **Biased Random** we observed a $\bar{x} = 5.94$ (SD: 1.71), showing that evaluators also considered this experience fairly interesting, considering the high complexity task that is autonomous music generation, especially in a random fashion. In the **Magenta/NightCafe** we observed a $\bar{x} = 6.00$ (SD: 1.85), showing that overall users also enjoyed the outcome generated by our baseline. This information can be visualized in Figure 6.15 (Q1).

After conducting the Friedman test on the evaluators' ratings, the results indicated a significant difference among the three systems in terms of perceived music quality ($p < 0.05$). The findings suggest that the evaluators' perceptions of music quality

varied significantly depending on the system used. We also run a t-test to compare the results from Solato vs Magenta/Nightcafe regarding musical quality, and we find a $p < 0.05$ (95% CI), showing a significant difference between the scores in favor of the Solato system.

In **Q2**, evaluators rated the landscapes generated by each system according to their perception of quality. Here, for **Solato**, **Biased Random** and **Magenta/NightCafe**, we observed a $\bar{x} = 8.12$ (SD: 1.72), $\bar{x} = 4.97$ (SD: 1.87) and $\bar{x} = 6.12$ (SD: 1.93), respectively, as shown in Figure 6.15 (Q2). Once again, evaluators acknowledged the landscapes generated by **Solato** as the most interesting.

After conducting the Friedman test, the results indicated that there were significant differences in the evaluators' ratings across the three systems ($p < 0.05$). This analysis suggests that the evaluators' perceptions of landscape quality differed significantly between the three systems, indicating that there are variations in the effectiveness of each system in generating landscapes that meet the evaluators' expectations. In this way, Solato was the system acknowledged as the system that generated the best landscapes. We also run a t-test to compare the results from Solato vs Magenta/Nightcafe regarding landscape quality, and we find a $p < 0.05$ (95% CI), showing a significant difference between the scores in favor of the Solato landscape generation.

In **Q3**, regarding music and video composition, we observed a $\bar{x} = 7.90$ (SD: 1.83) for **Solato**, $\bar{x} = 5.14$ (SD: 1.87) for **Biased Random** and $\bar{x} = 6.21$ (SD: 1.96) for the **Magenta/NightCafe**, as shown in Figure 6.15 (Q3). Therefore, according to the evaluators, **Solato** presented a more natural blend between music and video.

The Friedman test results revealed a significant difference in the ratings of the three systems ($p < 0.05$). The findings indicate variations in how well the systems achieved homogeneity in the generated soundscapes. These results contribute to our understanding of the effectiveness of different systems in producing homogenous music and images, which can be valuable in applications where such composition is desired. We also run a t-test to compare the scores from Solato vs Magenta/Nightcafe to evaluate if a homogeneous composition between music and image (i.e. the fundamental factor for the emergence of harmonious soundscapes) could be observed. We find a $p < 0.05$ (95% CI), showing a significant difference between the scores in favor of Solato.

In **Q4**, we queried evaluators about whether the music video was created by a **human** or a **machine**. It was possible to choose, through a linear scale, options from 1 to 10, where 1 to 5 meant **human** and 6 to 10 meant **machine**. We observed a $\bar{x} = 4.56$ (SD: 2.23), $\bar{x} = 7.47$ (SD: 1.63) and $\bar{x} = 7.77$ (SD: 1.75) for **Solato**, **Biased Random** and **Magenta/NightCafe**, respectively. Thus, interestingly, the Solato system stayed in the "human" spectrum, although borderline, while both Biased Random and Magenta/Nightcafe stayed in the "machine" spectrum. These outcomes

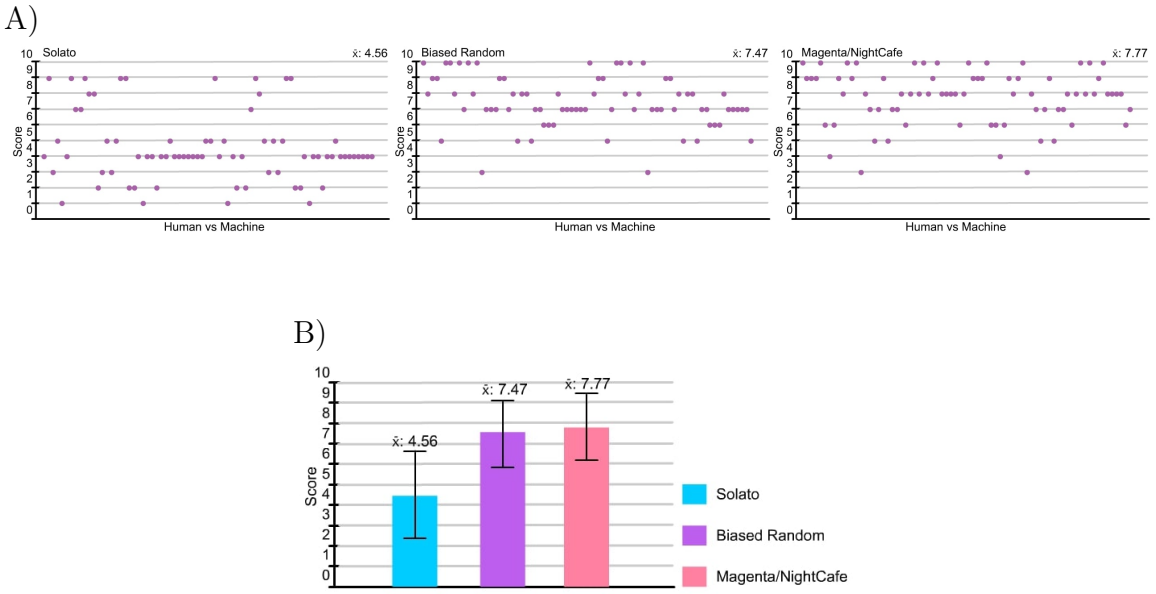


Figure 6.16: A) Dispersion graph (top) showing individual scores in the Human vs. Machine relation as acknowledged by human evaluators. B) Means (bottom) showing tendencies of the systems as acknowledged by evaluators. Values between 1 to 5 show a tendency to be a **human** production, while 6 to 10 show a tendency to be a **machine** production.

highlight that the evaluators correctly acknowledged that the outcomes were produced by autonomous systems, although there is also a tendency for Solato to be perceived as a human production. The trends between the systems are shown in Figure 6.16 - A and B, where we can observe the predominance of individual scores among our evaluators. According to our study, considering that a metric for quality in autonomous systems usually relates to mimicking human behavior, we conclude that the Solato system was the one closest to the goal, although none of the systems presented a clear prevalence to be acknowledged as a human-made production (e.g. score 3 or less).

As for the systems in the "machine" spectrum, the **Biased Random** was the one most acknowledged as a machine-made experience. Interesting to observe that evaluators, in general, could perceive the autonomous nature of this baseline, maybe through the identification of more cacophonies in the random music, as in fact there are minimum musical theory guidelines running underneath such approaches. In addition, visually, **Biased Random** contained some glitches, as shown in Figure 6.9. As for the Magenta/NightCafe approach, we believe this perception may come from the current hype between AI approaches for the generation of art, such as DALL-E [102] and Midjourney [93], that generates images through keywords. These approaches

are becoming fairly known, generating many interactions in social networks, and people might be getting used to their outcomes and general aesthetics. However, aspects that support evaluators' decision to acknowledge the music of these systems as a machine production remain unknown.

Since Solato was borderline in the human spectrum as acknowledged by evaluators (i.e. $\bar{x} = 4.56$), we ran a t-test to evaluate if we could observe a statistical difference between the scores attributed. In the scenarios of Solato vs Biased Random and Solato vs Magenta/NightCafe, we observed a statistically significant difference between the scores ($p < 0.01$ for both cases). As for the comparison between Biased Random vs Magenta/Nightcafe, the difference was not statistically significant ($p = 0.282$). These results reinforce that human evaluators were able to identify consistent traits of human production against traits of purely autonomous production, which paves the way for future discussions about what autonomous production means as a cultural phenomenon.

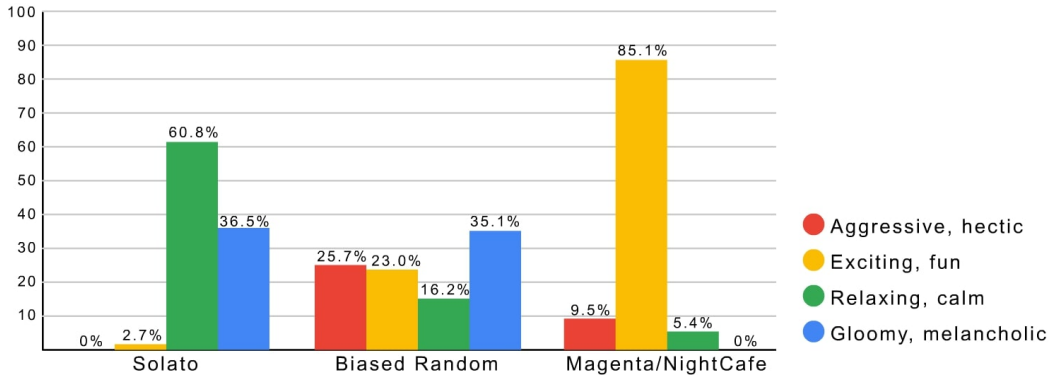


Figure 6.17: Comparison between the feelings acknowledged by evaluators in Solato, Biased Random, and Magenta/NightCafe.

In **Q5**, evaluators assigned a feeling according to their perception of the soundscapes generated by the systems, as shown in Figure 6.17. We observed an interesting trend in the different approaches to conveying different feelings. For instance, **Solato** was mainly acknowledged as generating a mixture of "relaxing/calm" outcomes (60.8%). This supports the original motivation behind the development of Solato, which was inspired by calming Ambient Music videos available on streaming platforms. However, we also observed a high recognition of the feeling "gloomy/melancholic" (36.5%). Although the choices of "relaxing" vs. "melancholic" seem contradictory, we believe this perception of "melancholy" comes from the system improvising over a minor scale, which is commonly perceived as "sad" in comparison to the major scales, that tends to be acknowledged as "happy" [19]. In addition, we also believe the sound texture of the samples we used for the system to generate music

may have an effect on this perception, although we can not confirm this assumption. In a minor portion, evaluators also identified the feeling as "exciting/fun" (2.7%).

As for the **Biased Random**, we observed the choice of a variety of feelings. The "gloomy/melancholic" was preponderant (35.1%), followed by "exciting/fun" (23%). Finally, "aggressive/hectic" (25.7%) and "relaxing/calm" (16.2%). The results for this evaluation suggest that random production was the most difficult for evaluators to assign a feeling, and, once again, this can be attributed to the fact that the guidelines for musical production in the **Biased Random** were mild, thus allowing for more unconventional musical structures to emerge.

As for the **Magenta/NightCafe**, we observed evaluators highly acknowledging the outcomes as "exciting/fun" (85.1%). A small portion also recognized the feelings of "relaxing/calm" (5.4%), and an even smaller share also identified "aggressive/hectic" (9.5%). It was quite surprising to observe how the Deep Learning approach could predominantly convey a specific feeling, contrasting directly with previously evaluated models.

Chi-square tests were conducted to examine the association between the frequencies in the four categories of feelings in each system. The results of the chi-square test for Solato indicate a statistically significant result ($p < 0.05$), pointing out that the preference for "relaxing/calm" and "gloomy/melancholic" feelings was not random. The majority of evaluators perceived a blend of "relaxing/calm" outcomes, aligning with the system's intended design focused on fostering relaxation. Additionally, a significant portion recognized the feeling of "gloomy/melancholic", which can be attributed to the system's use of a minor scale, commonly associated with "sadness". Moreover, a smaller fraction of evaluators identified "exciting/fun" feelings, indicating that Solato's emotional qualities extend beyond just calmness. On the other hand, for the Biased Random system, the chi-square test did not point to a statistically significant result ($p > 0.05$). As expected, the preference for each feeling in this random generation system appears to be due to chance. The outcomes displayed by Biased Random showcased a diverse range of feelings. The lack of statistical significance reinforces the notion that these feelings are not influenced by any specific pattern or bias, thereby highlighting the randomness of this approach. The Magenta/NightCafe system also demonstrated a statistically significant result ($p < 0.05$) in favor of "exciting/fun" feelings. The system consistently evoked a sense of excitement and fun in the evaluators, showcasing the effectiveness of the deep learning approach in generating a specific emotional quality. Though a small portion of evaluators also recognized "relaxing/calm" and "aggressive/hectic" feelings, the system's dominant focus on "exciting/fun" was evident in the statistical analysis.

The chi-square tests revealed the emotional qualities generated by each system. Solato effectively conveyed a "relaxing/calm" feeling, achieving its goal of creating relaxing soundscapes. In contrast, Magenta/NightCafe predominantly evoked "ex-

citing/fun" feelings, while Biased Random displayed a diverse range of emotions due to its randomness. These results confirm our hypothesis that Solato and Magenta/NightCafe have specific, non-random emotional preferences, highlighting the distinct emotional expressions of each approach.

6.5 Discussion

6.5.1 Research Questions

RQ1: How to generate audio and images simultaneously and coherently in an autonomous fashion, preserving human expressivity? The field of dynamic audio and image generation encompasses a wide range of methods and techniques, including machine learning, deep learning, and text-based input approaches like stable diffusion. Rather than relying on a single method, a combination of these approaches to generate artistic and creative solutions can be employed, as we demonstrated in the generation of our baseline (i.e. Magenta/NightCafe). Most importantly when designing such systems, however, is to consider the extent of freedom and human involvement they enable. Previous studies have acknowledged the complexity of incorporating human intent, which we refer to in this work as "human expressivity", into the artistic creation process, as discussed in Chapter 5. This complexity has prompted extensive discussions on the relevance of fully algorithmically generated assets as artistic works. These discussions encompass various dilemmas, such as copyright, originality, and the preservation of emotional engagement in the creative process. Amid this ongoing debate, our work introduces an approach that specifically concentrates on the generation of soundscapes. Although our approach operates autonomously, it retains a degree of human intent through customizable generated parameters and the option to incorporate human-created dataset assets, if desired. Therefore, our work successfully achieved its goal by proposing an approach that preserves human expressivity. We hope that our work serves as inspiration for further research and the development of novel systems that can generate compelling outcomes while also considering and incorporating human expressivity in the process.

RQ2: Are human evaluators able to identify nuances and distinguish between different autonomous systems with different levels of human involvement in the production? Our study provided insights into the ability of human evaluators to differentiate between soundscape compositions generated by humans and those produced by machines. While the Solato production, which had a more active human involvement, fell within the borderline range of identification ($\bar{x} = 4.56$, as shown in Figure 6.17), there was a statistically significant distance

observed from the other baselines, Biased Random and Magenta/NightCafe ($p < 0.01$ for both cases). On the other hand, no significant difference was observed among the machine-generated productions (Biased Random and Magenta/NightCafe) with a p-value of 0.282, and similar scores were attributed to both. Therefore, although Solato's classification was borderline, it still falls within the human spectrum of evaluation ($score < 5$). On the other hand, it was evident that both fully autonomous productions clearly fall into the machine spectrum ($score > 5$). Thus, based on our study, it is possible to confidently conclude that human evaluators are capable of discerning between outcomes generated by humans and those generated by machines.

RQ3: Are the outcomes pleasant to human listeners? Overall, the findings indicate that the outcomes generated by the systems were generally pleasant to human listeners. Solato consistently outperformed the other systems in terms of both music and landscape quality. These results contribute to our understanding of the effectiveness of autonomous music and landscape generation systems in meeting evaluators' expectations. The findings successfully addressed the research question, affirming that the outcomes were perceived as pleasant by human listeners, and highlighting the effectiveness of autonomous music and landscape generation systems in meeting evaluators' expectations. Furthermore, it is important to acknowledge the potential of rule-based mechanics in generating coherent artistic outcomes that align with the objective of this study, which aims to assess and explore a stronger human presence within the context of current autonomous approaches. This perspective allows us to envision a scenario where emerging creative technologies are employed to enhance human capabilities rather than instilling apprehension about their potential to replace human involvement.

RQ4: What feelings do these outcomes generated by autonomous systems convey to human observers and listeners? The evaluation of different soundscapes generated by Solato, Biased Random, and Magenta/NightCafe revealed distinct trends in the perception of feelings. Solato was mainly recognized for producing a mixture of "relaxing/calm" outcomes, aligning with its original goal. Surprisingly, evaluators also identified a significant portion of "gloomy/melancholic" feelings, which may be attributed to the system improvising over a minor scale. Biased Random exhibited a wide range of feelings, with "gloomy/melancholic" being the most prominent, followed by "exciting/fun", "aggressive/hectic", and "relaxing/calm". Evaluators found it challenging to assign a specific feeling to random production due to the system's mild guidelines. In contrast, Magenta/NightCafe predominantly conveyed "exciting/fun" feelings, showcasing the ability of the Deep Learning approach to consistently evoke a specific emotion. The results highlight the importance of the approach in shaping the emotional response of evaluators

and suggest further investigation into the factors influencing perception. Overall, the evaluation provides valuable insights into the diverse feelings evoked by different soundscapes and suggests avenues for future research. The evaluation findings directly address the research question by analyzing the perceptions of human evaluators, we gained insights into the specific feelings evoked by the soundscapes generated by Solato, Biased Random, and Magenta/NightCafe. The results revealed distinct patterns and tendencies in the assigned feelings, providing information about the emotional responses elicited by these autonomous systems.

6.5.2 Additional Questions

Potency of the approaches to convey different feelings: Although a relation was preserved between the 3 systems (i.e. Biased Random model uses the same dataset of images and music as Solato, and Magenta/NightCafe music was trained using a music sample generated by the Solato system), each system presented very particular peculiarities for the evaluators in terms of the feelings conveyed. We believe our evaluation contributes to designers, developers, and researchers in the field of affective computing who intend to develop systems capable of conveying specific feelings, complementing works such as Ferreira et al. [43].

Mimicking human behavior: In our user study, the Rule-based model generated music perceived as the most human-like creation, as acknowledged by evaluators. Hence, our study suggests that Rule-based mechanics is more efficient to tackle scale progressions, harmonic fields, etc. According to our study, the Magenta approach showed limitations in aspects such as producing outcomes that sounded more "organic" in a way to resemble human production, as shown in Figure 6.16 A and B of our user study. Although autonomous for the generation of outcomes, our approach allows developers to customize the shapes and sizes of building blocks, thus fostering a human touch in autonomous creation. In this way, our study relates to the approach discussed in Chapter 5, where it was shown that it is challenging to neglect the human factor for the emergence of meaning in digital artworks, and that the absence of the human factor is perceived by the audience.

Scalability and adaptability: In addition to its primary focus, Solato holds potential for utilization in therapeutic contexts, such as relaxation, meditation, and stress relief. The results observed in Subsection 6.4.2 and depicted in Figure 6.17 highlight the prominent presence of the feeling of "relaxing/calm", acknowledged by the majority of human evaluators, accounting for 60.08% of responses. These findings instill confidence in Solato's ability to effectively achieve specific therapeutic goals by generating relaxing environments with calming music. In addition, our system holds

the potential to serve as a learning tool, enriching the musical experience for users. Learning music can often be overwhelming, as discussed in Chapter 5, especially for newcomers. Thus, our system has the capability to provide valuable visual guidance, aiding users in understanding and applying musical theory principles. For instance, we have incorporated color aesthetics that are associated with musical tones, offering learners an additional layer of information to easily comprehend the rules and concepts of musical theory. By drawing parallels between the progression of musical notes and color tones, our system facilitates the formation of meaningful associations, enhancing the learning experience.

6.5.3 Limitations

The work presented in this chapter has some limitations. In this section we will discuss those, some of our solutions to minimize their impacts, and also how we intend to address these challenges in future works.

Evaluation video samples consisted of a clipping of the experience Solato is a long experience. Therefore, we could not evaluate a complete day/night loop cycle in our user study. As presented in our User Study, evaluators watched and heard a 2:30-minute-long audiovisual piece, which may not cover all of the system's capabilities for the generation of soundscapes. However, we are sure that the evaluation successfully satisfied the main questions of this work, although this evaluation can be expanded in the future to test new features of the system, such as evaluating the impact of the passage of time simulation on the perception of soundscapes.

External baseline Originally, our goal was to have an external commercial work based on a deep learning model that generates the same effect as our system. However, to the best of our knowledge, we could not find one that addresses the same challenge of generating audio and visuals simultaneously for the generation of soundscapes. Thus we created our own baselines (i.e. Biased Random and Magenta/Nightcafe). Although this baseline fulfilled the demands of our assessment and was generated under strong rigor, in the future, we also intend to expand our evaluation by adding more baseline systems, such as systems with similar goals and also human-made production through partially-autonomous approaches. However, currently, existing systems could generate noise in the comparison between samples (e.g. samples in which the timbres or "sound textures" were not too different to affect evaluators' perception). Therefore, the presented baselines supported us to overcome these obstacles and also presented compelling experiences.

Country and cultural background diversity Currently, our evaluations are limited to specific groups in Brazil, like game developers, independent artists, graphic designers, and computer science students. However, for future studies, we plan to diversify our participant pool, encompassing individuals from various backgrounds and countries. This expansion aims to enhance the reproducibility of our findings and explore diverse patterns of emotions and feelings perceived by evaluators. Cultural background may influence these perceptions, and including a broader range of participants will provide deeper insights into these potential influences.

6.6 Conclusion

In this work, we presented a system capable of generating music and image instances simultaneously and coherently preserving meaning, also presenting a level of unpredictability. Our study shows that, by using our approach, it is possible to harness musical theory and image generation directly, and thus generate harmonious and coherent outcomes acknowledged as an interesting soundscape by human evaluators.

Our user study provided insights into a Rule-based method for the production of coherent and meaningful audiovisual outcomes. Our study proposes that Rule-based mechanics still offer valuable contributions to this work's goal, surpassing state-of-the-art approaches that use techniques such as Deep Learning. Solato mechanics was able to overcome a Deep Learning model in the quality of the emergent artistic pieces, also being recognized as human-made production (i.e. human evaluators acknowledged the meaning and human intent behind the creation, a common goal of AI systems). In other words, our study shows that autonomous approaches that rely on the more active presence of humans operating "behind the curtains" in production, such as Solato, still present effective solutions compared to current techniques.

We also provide a glimpse of how AI models along with HCI techniques applied for the generation of artworks may advance in the creation of more engaging experiences, in a way not to replace human expressivity but to extend it. For instance, our results suggest that different aspects of the different approaches presented can be merged for the creation of more robust system mechanics, allowing complex expressive outcomes to emerge out of autonomous systems. We also show promising results in terms of generating music and images simultaneously in a dynamic fashion, in addition to evaluating the outcomes beyond the approaches themselves, filling a gap in AI applications toward audiovisual productions.

Our work also provided insightful perspectives about abstract questions such as feelings conveyed by outcomes of autonomous systems, complementing current works in the field by showing how systems can employ music and image stimuli to convey different emotions to a general audience.

Chapter 7

Discussion

In Chapters 4, 5, and 6, this thesis explored different creative experiences that employ game-based mechanics, focusing on the generation of music in both implicit and explicit ways, and the creation of visual outcomes like soundscapes. Although each chapter concentrates on a distinct type of human-machine collaboration, common insights emerge from the specificities of the works presented.

7.1 Exploring Human-Machine Collaboration for the Dynamic Generation of Assets

This thesis introduced a main Research Question (MRQ): ***How do different levels of human-machine collaboration, ranging from partially-autonomous to fully-autonomous approaches (i.e. implicit cooperation, meta-interactivity, and machine autonomy) affect the quality, user experience, and aesthetic properties of music produced in virtual environments?*** The varying degrees of human-machine collaboration explored reveal a dynamic interplay between creative agency, user experience, system guidelines, and aesthetic intentions.

Implicit Cooperation, demonstrated through Microbial Art, achieves a balance where human creators actively or passively influence music composition. Machine intervention enriches the music quality and user experience, granting players the flexibility to either engage with the gameplay mechanics or focus on creating musically interesting pieces. Regardless of user intent, the systems developed promote engaging experiences that captivate users' interest.

Meta-interactivity, demonstrated through Bubble Sounds, fosters a real-time dialogue between humans and game mechanics, enhancing both user experience and music quality. This method minimizes machine intervention to focus on human expressivity, creating a controlled yet expressive environment similar to playing

a musical instrument. The experience stimulates creativity through its intuitive interactions, such as its color-to-notes translation.

In contrast, machine autonomy, demonstrated through Solato, prompts considerations about freedom and creative control while still retaining a subtle level of human influence. This approach exemplifies the potential for autonomous systems to produce aesthetically interesting musical outcomes.

Throughout this exploration, the collaborative approaches underscore the role that game mechanics and intelligent systems can have as creative allies, empowering individuals in artistic practices like music composition. By including novice users, these approaches democratize artistic creation, enhancing accessibility while preserving human intent in the creative process.

Thus, this thesis contributes to the ongoing debate on human-machine collaboration in the arts, underscoring the vital role of human input in augmenting creative outcomes. For example, the meta-interactivity approach presented in Chapter 5 illustrates how game mechanics can support human creativity without overshadowing it, fostering engaging interactive experiences.

7.2 The Interplay Between User Expressivity and System Guidance

This thesis explored the balance between user expressivity and gameplay mechanics for artistic creation. It investigates how intelligent systems can offer both guidance for novices and creative freedom for skilled composers, a dichotomy that demands individuals feel in control of the creative process while still benefiting from the system's capabilities and expertise.

For instance, the music generation mechanisms detailed in Chapters 4 and 5 use implicit cooperation and meta-interactivity approaches to allow users to influence musical compositions profoundly. These systems empower users to select note heights within a musical scale, enabling the creation of original music without requiring detailed knowledge of music theory. Users can trust their intuition in selecting notes, with the system ensuring that their choices result in harmonious outcomes rather than dissonant sounds.

In the implicit cooperation system described in Chapter 4, users receive guidance as the system presents notes compatible with their current selections in the game. This guided approach ensures coherence in the musical creation process, steering users towards harmonious compositions. In contrast, the meta-interactivity approach outlined in Chapter 5 provides users with greater freedom to experiment and make mistakes, allowing even potentially cacophonic structures to be adjusted for a more pleasant auditory experience. For instance, the system ensures the coherence of the

tempo in the emergent musical corpus, showcasing a dynamic and interactive feature so discreetly that users may not even realize its operation. In this way, the approach enables users to craft unique and expressive music while still benefiting from the system's expertise in music theory and composition.

7.3 Beyond Play and Creativity

This thesis demonstrates that gameplay mechanics, when integrated into interactive systems, can augment the creative process. These mechanics not only allow individuals to focus on the essential elements of their artistic endeavors but also potentially enhance productivity and the depth of artistic outputs. Such integration shifts traditional perceptions of game mechanics from mere entertainment to powerful tools that enhance creativity.

Furthermore, the collaborative approaches explored in this thesis extend their utility beyond mere entertainment; they serve as vital educational tools. These systems are particularly advantageous for individuals tackling the steep learning curves associated with musical education or other sophisticated artistic practices. By alleviating common challenges and frustrations, these interactive systems help prevent discouragement and dropout, thus making the creative process more accessible and engaging.

Ultimately, this research contends that the interplay between human creativity and machine efficiency enriches the artistic journey far beyond simple enhancements in output. It fosters a symbiotic partnership that elevates the creative process. While machines are shown to amplify human capabilities, the quintessence of creativity—deeply rooted in human experience and emotion—remains distinctly human. This partnership not only preserves but also celebrates the inherent human elements of art, championing a vision where technology and creativity merge seamlessly.

7.4 Art or Design?

The advancement of autonomous approaches in creative domains challenges traditional views of art and design. Systems like Midjourney [93] challenge the conventional belief that machines cannot create art without explicit human expressiveness. Instances where human curation misinterprets the source as originating from a human emphasize the substantial progress achieved by autonomous systems. However, according to the research undertaken in this thesis, autonomous systems do not replace the human touch that is central to the artistic experience. Instead, they can assist in the creative process, such as in creating music.

As we integrate autonomous approaches into creative fields, distinguishing between the machine's role in supporting creativity versus creating standalone pieces is crucial. This distinction ensures that art remains a reflection of human experience and emotion, preserving its unique value in society.

Art has historically served as a medium for humanity to convey emotions, provoke thought, and engage in discourse on various issues, from politics to minority rights. However, autonomous machines, in their current state, fall short of this potency. Observations across domains, such as in games, reveal people leveraging AI as a resource to meet specific immediate demands – artists find inspiration, refine concepts, or create purpose-driven pieces like character mockups and web-based publicity campaigns. Machines lack the ability to contextualize independently, steering creations towards addressing specific demands rather than embodying the pure "imaginary" essence associated with art.

This raises questions about the accuracy of etymologies, such as the term "AI" itself, initially coined for impactful promotion rather than a direct reflection of intelligence (which is a quality unique to mankind). Similarly, attempting to define AI-generated outcomes as "art" prompts further inquiry. Arguably, current applications of autonomous AI align more closely with "design" than "art", although this classification can be contentious. The term "digital art" appears fitting; however, it risks being confused with existing notions of digital art generated through 2D, 3D, and musical editing tools. Consequently, the ongoing discourse in this new digital era necessitates the development of more precise terms to delineate the nuances within these generative fields. Processes incorporating human-in-the-loop dynamics within creative endeavors, as explored in this thesis, should exercise caution in assessing whether human expressivity and intent have been maintained amid collaboration and machine intervention.

7.4.1 Limitations and Future Work

7.4.1.1 Implicit Cooperation

The examination of Implicit Cooperation in Chapter 4 involved a modest user pool (n = 10), highlighting the need for expanded experiments to strengthen the discussion of its findings. In forthcoming evaluations, a comprehensive assessment with a larger participant base is underway, aiming for a deeper exploration into the dynamics of the approach to gain a better understanding of its effectiveness. These ongoing assessments hold the potential to provide insights into the scalability and generalizability of the approach, thereby laying the groundwork for broader applications.

In future work, we aim to investigate the extent of human and machine contributions within the Implicit Cooperation framework, exploring methods such

as listener evaluations and compositional pattern analysis to quantify how closely machine-generated outputs align with human creativity and to better understand the collaborative dynamics of co-creation in music.

Future work will also explore alternative configurations of the building blocks and dynamic grids used in the current algorithm. By experimenting with different configurations, it may be possible to analyze how variations impact both player movement and the quality of the generated music. We expect that this exploration might provide deeper insights into optimizing the system for diverse gameplay scenarios and enhancing its applicability across various game environments. Additionally, testing these configurations could further validate the system's capacity to produce musically coherent outputs in response to player interactions.

7.4.1.2 Meta-interactivity

The assessment in Chapter 5 employed an adapted version of Bubble Sounds, maintaining all its original functionalities while allowing control through a regular mouse. This adaptation addresses practical accessibility concerns, especially given the limited mainstream adoption of VR devices. In future evaluation sessions, recording experiments on video will enable a more thorough analysis of participants' interactions with the system, providing valuable insights into their spontaneous reactions and experiences.

Additionally, we plan to establish more robust baselines to compare the music created with Bubble Sounds against other systems that operate in a similar manner. These comparisons will leverage the same dataset of musical samples and timbres, allowing us to evaluate the effectiveness of the system in fulfilling its role as a playful musical instrument.

The user study also indicated that non-experts can produce music comparable to that of experts, aligning with the project's objectives. Crucially, the musical output is distinguishable from random, confirming that user interaction influences the results. While this consistency supports accessibility, we recognize the need to enhance personalization and originality. Expanding the diversity of musical samples and instruments could further improve the uniqueness of user creations.

However, the difference in evaluation scores suggests that, while users enjoy the experience and find the interface intuitive, the musical output may not always meet their expectations for quality. This does not imply that the system produces unsatisfactory music but rather highlights an opportunity to refine the output to better align with users' creative aspirations. Enhancing the diversity and sophistication of musical options could further elevate the system's artistic potential.

7.4.1.3 Machine Autonomy

In Chapter 6, while VR evaluation was not feasible in the current study, future iterations of Solato aim to assess its VR capacity, considering factors such as immersion, comfort, and perceptual differences compared to non-VR experiences.

In subsequent iterations, Solato will incorporate intuitive interfaces or controls, allowing users to manipulate parameters that directly impact the emergent audiovisual outcomes. These parameters may include tempo, mood, color palette, visual effects, audio effects, and the balance between sound and image manifestations, thereby facilitating interactive experimentation and human agency in the creative process.

We also plan to conduct novel assessments using stronger baselines. With the rapid advancement of generative co-creative techniques based on Large Language Models (LLMs), these new baselines will likely include more recent Deep Learning approaches. In Chapters 3, 4, and 5, we primarily used random approaches as baselines. It is important to note that random approaches in music remain a relevant production method. Temperley's work on music and probability, which explores the random and probabilistic nature of human decision-making in music perception, supports this view [118].

In future work, we aim to investigate why users perceive the Biased Random approach as machine-made, focusing on its lack of coherence. This highlights the expectation that human-made outputs should be coherent and typical. We need to explore why users see Biased Random as machine-made and Solato as human-made, considering both quality and unique style. Although the soundscape is dynamically generated, it was carefully crafted by an artist. This raises important questions: Is dynamic generation as critical as the setup? Which aspect shapes users' perceptions of human vs. machine-made creations more? Addressing these will provide deeper insights into the interplay between dynamic generation and artistic setup in co-creative systems. Additionally, we plan to conduct experiments with creators who can modify the dataset of images and sounds to validate Solato's generalization by ensuring it can adapt to various inputs and still produce coherent and meaningful outputs.

Chapter 8

Conclusions

This thesis explored the intersections between human creativity and machine capabilities, specifically focusing on the roles of partially-autonomous and autonomous systems in music co-creation. We explored the transformative potential that game mechanics hold for augmenting artistic expressions without diminishing the essence of human creativity.

The research demonstrated that intelligent systems and game mechanics, when integrated thoughtfully within collaborative frameworks, can enhance human artistic endeavors. This was particularly observed through the concepts of Implicit Cooperation and Meta-interactivity, where gameplay mechanics have been shown to not only support but also enrich the creative process. Even in the context of Autonomous Creation, it is possible to foster collaboration and space for human intervention. By fostering an environment where human sentiment and expressivity are preserved, these systems ensure that technology complements rather than replaces human creativity.

The contributions of this thesis are multifaceted. Firstly, it established that partially-autonomous and autonomous approaches can be a potent ally for artistic expression, enabling individuals to push beyond their perceived creative limits. This benefits both artists and enthusiasts, from experts to novices, expanding their capabilities and introducing them to new forms of artistic expression. Secondly, the emotional depth of the works produced through these collaborative systems has been critically assessed. The findings reveal that the technical outcomes of such collaborations are aesthetically pleasing and resonate on an emotional level with human audiences.

Furthermore, the implications of this research extend beyond music generation, influencing industry practices, particularly in the game sector. The integration of game design and gameplay mechanics in music co-creation can lead to more engaging, immersive, and creatively stimulating environments, enhancing both the creator's and the user's experience. This thesis posits that the future of creative industries lies in

leveraging the synergy between human creativity and interactive systems to foster innovative outcomes that are both technically sound and emotionally engaging.

In conclusion, the research presented in this thesis advocates for a balanced approach to human-machine integration in creative fields, promoting systems that enhance human capabilities without undermining their expressivity and creativity. As we witness the complexities of AI insertion in creative contexts, it is important to maintain this balance to ensure that digital arts remain a human-centric field. This work hopes to contribute to the ongoing dialogue about the role of technology in art, encouraging further exploration, discussion, and presentation of novel systems that explore the integration of not only emergent AI techniques but also the creative implementation of game mechanics capable of establishing rich ways in which any individual can express themselves.

References

[1] Nipun Agarwala, Yuki Inoue, and Axel Sly. "Music composition using recurrent neural networks". In: *CS 224n: Natural Language Processing with Deep Learning, Spring* 1 (2017), pp. 1–10.

[2] Andrea Agostinelli et al. *MusicLM: Generating Music From Text*. 2023. arXiv: 2301.11325 [cs.SD].

[3] Maximilian Altmeyer et al. "Here Comes No Boom! The Lack of Sound Feedback Effects on Performance and User Experience in a Gamified Image Classification Task". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517581. URL: https://doi.org/10.1145/3491102.3517581.

[4] Charles Ames. "The Markov Process as a Compositional Model: A Survey and Tutorial". In: *Leonardo* 22.2 (1989), pp. 175–187.

[5] Willi Apel. *The Harvard dictionary of music*. Vol. 16. Harvard University: Harvard University Press, 2003.

[6] Mark Applebaum. *The mad scientist of music*. TEDxStanford Conferences. 2012.

[7] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. "Tools for placing cuts and transitions in interview video". In: *ACM Transactions on Graphics (TOG)* 31.4 (2012), pp. 1–8.

[8] John A. Biles. "GenJam: A Genetic Algorithm for Generating Jazz Solos". In: *Proceedings of the International Computer Music Conference*. 1994.

[9] Renaud Bougueng Tchemeube, Jeffrey John Ens, and Philippe Pasquier. "Calliope: A co-creative interface for multi-track music generation". In: *Proceedings of the 14th Conference on Creativity and Cognition*. 2022, pp. 608–611.

[10] Joana Braguez. "AI as a Creative Partner: Enhancing Artistic Creation and Acceptance". In: *BAMC2023 Conference Proceedings*. IAFOR. 2023, pp. 121–131. DOI: 10.22492/issn.2435-9475.2023.11. URL: https://papers.iafor.org/submission72833/.

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

[12] Daniel Brown. "Mezzo: An adaptive, real-time composition program for game soundtracks". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 8. 4. AAAI, 2012, pp. 68–72.

[13] Marcio Cabral et al. "Crosscale: A 3D virtual musical instrument interface". In: *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE. Arles, France: IEEE Xplore, 2015, pp. 199–200.

[14] Filippo Carnovalini and Antonio Rodà. "Computational creativity and music generation systems: An introduction to the state of the art". In: *Frontiers in Artificial Intelligence* 3 (2020), p. 14.

[15] Pablo Samuel Castro. "Performing Structured Improvisations with pre-trained Deep Learning Models". In: *arXiv preprint arXiv:1904.13285* 1 (2019), p. 7.

[16] Matteo Casu, Marinos Koutsomichalis, and Andrea Valle. "Imaginary Soundscapes: The SoDA Project". In: *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*. AM '14. Aalborg, Denmark: Association for Computing Machinery, 2014. ISBN: 9781450330329. DOI: 10.1145/2636879.2636885. URL: https://doi.org/10.1145/2636879.2636885.

[17] *Celebrating Johann Sebastian Bach*. https://www.google.com/doodles/celebrating-johann-sebastian-bach. Accessed: 2023-09-21. 2019.

[18] Miguel Civit et al. "A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends". In: *Expert Systems with Applications* 209 (2022), p. 118190. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.118190. URL: https://www.sciencedirect.com/science/article/pii/S0957417422013537.

[19] William G Collier and Timothy L Hubbard. "Musical scales and evaluations of happiness and awkwardness: Effects of pitch, direction, and scale mode". In: *American Journal of psychology* 114.3 (2001), pp. 355–375.

[20] Nick Collins and Julio d'Escriván. *The Cambridge Companion to Electronic Music*. Cambridge University Press, 2008.

[21] David Cope. *Experiments in Musical Intelligence*. A-R Editions, Inc., 1996.

[22] Mark D'Inverno and Jon McCormack. "Heroic versus Collaborative AI for the Arts". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, pp. 2438–2444. ISBN: 9781577357384.

[23] Magdalena Dabrowski. "Kandinsky Compositions: The Music of the Spheres". In: *MoMA* 5.4 (1995), pp. 10–13. ISSN: 08930279. URL: http://www.jstor.org/stable/4381285.

[24] Didier Dambrin. *FLStudio*. 1997.

[25] Nicholas Davis. "Human-computer co-creativity: Blending human and computational creativity". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 9. 6. 2013, pp. 9–12.

[26] Nicholas Davis et al. "Empirically Studying Participatory Sense-Making in Abstract Drawing with a Co-Creative Cognitive Agent". In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. IUI '16. Sonoma, California, USA: Association for Computing Machinery, 2016, pp. 196–207. ISBN: 9781450341370. DOI: 10.1145/2856767.2856795. URL: https://doi.org/10.1145/2856767.2856795.

[27] Sebastian Deterding et al. "From game design elements to gamefulness: defining "gamification"". In: *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. Tampere, Finland: Association for Computing Machinery – ACM, 2011, pp. 9–15.

[28] Prafulla Dhariwal et al. "Jukebox: A generative model for music". In: *arXiv preprint arXiv:2005.00341* 1.1 (2020), p. 20.

[29] Darina Dicheva et al. "Gamification in education: A systematic mapping study". In: *Journal of Educational Technology & Society* 18.3 (2015).

[30] Monica Dinculescu, Jesse Engel, and Adam Roberts, eds. *MidiMe: Personalizing a MusicVAE model with user data*. 2019.

[31] Jonathan Duckworth et al. "Resonance: an interactive tabletop artwork for co-located group rehabilitation and play". In: *International Conference on Universal Access in Human-Computer Interaction*. Germany: Springer, 2015, pp. 420–431.

[32] Kemal Ebcioglu. "An Expert System for Harmonizing Four-Part Chorales". In: *Computer Music Journal* 12.3 (1988), pp. 43–51.

[33] Eran B Egozy et al. *Systems and methods for simulating a rock band experience*. US Patent 8,663,013. Apr. 2014.

[34] Arne Eigenfeldt and Philippe Pasquier. "Considering Vertical and Horizontal Context in Corpus-based Generative Electronic Dance Music". In: *Proceedings of the Fourth International Conference on Computational Creativity*. ICCC. Sydney, Australia: International Conference on Computational Creativity, 2013.

[35] Ahmed M. Elgammal et al. "CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms". In: *CoRR* abs/1706.07068 (2017). arXiv: 1706.07068. URL: http://arxiv.org/abs/1706.07068.

[36] Brian Eno. "Ambient music". In: *Audio Culture. Readings in Modern Music* 9497 (2004).

[37] Andromeda Entertainment. *SoundSelf: A Technodelic*. 2020.

[38] Mário Escarce et al. "Emerging Sounds Through Implicit Cooperation: A Novel Model for Dynamic Music Generation". In: *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment)*. Snowbird, Little Cottonwood Canyon, Utah, USA: AAAI, 2017, pp. 186–192. URL: https://aaai.org/ocs/index.php/AIIDE/AIIDE17/paper/view/15887.

[39] Mário Escarce Junior et al. "A Meta-Interactive Compositional Approach That Fosters Musical Emergence through Ludic Expressivity". In: *Proc. ACM Hum.-Comput. Interact.* 5.CHI PLAY (Oct. 2021). DOI: 10.1145/3474689. URL: https://doi.org/10.1145/3474689.

[40] Mário Escarce Junior et al. "The Aesthetics of Disharmony: Harnessing Sounds and Images for Dynamic Soundscapes Generation". In: *Proc. ACM Hum.-Comput. Interact.* 7.CHI PLAY (Oct. 2023). DOI: 10.1145/3611045. URL: https://doi.org/10.1145/3611045.

[41] Peter Gotcher Evan Brooks. *Pro Tools*. 1991.

[42] Jose D Fernández and Francisco Vico. "AI methods in algorithmic composition: A comprehensive survey". In: *Journal of Artificial Intelligence Research* 48 (2013), pp. 513–582.

[43] Lucas N. Ferreira and E. James Whitehead. "Learning to Generate Music With Sentiment". In: *ArXiv* abs/2103.06125 (2019).

[44] Rebecca Anne Fiebrink. *Real-time human interaction with supervised learning algorithms for music composition and performance*. Princeton University, New Jersey, United States: Citeseer, 2011.

[45] William W Gaver et al. "The drift table: designing for ludic engagement". In: *CHI 04 – Extended Abstracts on Human Factors in Computing Systems*. New York, NY, United States: ACM, 2004, pp. 885–900.

[46] Lauryn Gayhardt and Maya Ackerman. "SOVIA: Sonification of Visual Interactive Art." In: *ICCC*. 2021, pp. 391–394.

[47] Sarah Victoria Gentry et al. "Serious gaming and gamification education in health professions: systematic review". In: *Journal of medical Internet research* 21.3 (2019), e12994.

[48] Darrell Gibson and Richard Polfreman. "A framework for the development and evaluation of graphical interpolation for synthesizer parameter mappings". In: *Proceedings of the 16th Sound and Music Computing Conference*. Nottingham, United Kingdom: SMC2019, 2019, p. 8.

[49] Juho Hamari, Jonna Koivisto, and Harri Sarsa. "Does gamification work?–a literature review of empirical studies on gamification". In: *2014 47th Hawaii international conference on system sciences*. Waikoloa, HI, USA: IEEE, 2014, pp. 3025–3034.

[50] Harmonix. *Rockband*. 2007.

[51] David Henley. "Art of Disturbation: Provocation and Censorship in Art Education". In: *Art Education* 50.4 (1997), pp. 39–45.

[52] Simon Holland et al. "Music interaction: understanding music and human-computer interaction". In: *Music and human-computer interaction*. London, UK: Springer, 2013, pp. 1–28.

[53] T. Holmes. "Art games and Breakout: New media meets the American arcade". In: *Computer Games and Digital Cultures Conference*. Finland, 2002.

[54] Amy K Hoover and Kenneth O Stanley. "Exploiting functional relationships in musical composition". In: *Connection Science* 21.2-3 (2009), pp. 227–251.

[55] Amy K Hoover, Paul A Szerlip, and Kenneth O Stanley. "Functional scaffolding for composing additional musical voices". In: *Computer Music Journal* 38.4 (2014), pp. 80–99.

[56] Amy K. Hoover, Michael P. Rosario, and Kenneth O. Stanley. "Scaffolding for Interactively Evolving Novel Drum Tracks for Existing Songs". In: *Applications of Evolutionary Computing*. Ed. by Mario Giacobini et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 412–422. ISBN: 978-3-540-78761-7.

[57] Amy K. Hoover, Michael P. Rosario, and Kenneth O. Stanley. "Scaffolding for Interactively Evolving Novel Drum Tracks for Existing Songs". In: *Applications of Evolutionary Computing*. Ed. by Mario Giacobini et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 412–422. ISBN: 978-3-540-78761-7.

[58] Andrew Horner and David E. Goldberg. "Machine Tongues XVI: Genetic Algorithms and Evolutionary Music Composition". In: *Computer Music Journal* 19.3 (1995), pp. 17–34.

[59] Allen Huang and Raymond Wu. "Deep Learning for Music". In: *ArXiv* abs/1606.04930 (2016), p. 8.

[60] Andy Hunt and Ross Kirk. "Mapping strategies for musical performance". In: *Trends in gestural control of music* 21.2000 (2000), pp. 231–258.

[61] Kori Inkpen et al. "Where is the human? Bridging the gap between AI and HCI". In: *Extended abstracts of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–9.

[62] T. Iwai. *Electroplankton*. 2005.

[63] Mikhail Jacob and Brian Magerko. "Interaction-based Authoring for Scalable Co-creative Agent". In: *Proceedings of the Sixth International Conference on Computational Creativity*. Utah, USA: ICCC, 2015, pp. 236–246.

[64] Sergi Jordà et al. "The reacTable". In: *In Proceedings of the International Conference on Computer Music – ICMC*. New York, NY, United States: Association for Computing Machinery – ACM, 2005.

[65] Anna Jordanous. "Co-creativity and perceptions of computational agents in co-creativity". In: (2017).

[66] Yuma Kajihara, Shoya Dozono, and Nao Tokui. "Imaginary Soundscape: Cross-Modal Approach to Generate Pseudo Sound Environments". In: *Proceedings of the Workshop on Machine Learning for Creativity and Design (NIPS 2017), Long Beach, CA, USA*. 2017, pp. 1–3.

[67] David Kanaga and Ed Key. *Proteus*. 2013.

[68] Otto Karolyi. *Introducing Music*. Penguin Books; Reissue edition, 1965.

[69] Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription*. United States: Springer-Verlag US, 2007, p. 440. ISBN: 978-1-4419-4035-3.

[70] R. Koster. *A Theory of Fun for Game Design*. Paraglyph Press, 2004.

[71] Walter Ledermann. *Introduction to the theory of finite groups*. Oliver and Boyd, 1949.

[72] George E. Lewis. "Too Many Notes: Computers, Complexity and Culture in "Voyager"". In: *Leonardo Music Journal* 10 (2000), pp. 33–39.

[73] Pinyao Liu et al. "Virtual Transcendent Dream: Empowering People through Embodied Flying in Virtual Reality". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517677. URL: https://doi.org/10.1145/3491102.3517677.

[74] Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. "Targeting horror via level and soundscape generation". In: *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*. 2015.

[75] Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. "Modelling affect for horror soundscapes". In: *IEEE Transactions on Affective Computing* 10.2 (2017), pp. 209–222.

[76] Phil Lopes, Antonios Liapis, and Georgios N. Yannakakis. "Framing Tension for Game Generation". In: *Proceedings of the International Conference on Computational Creativity*. 2016.

[77] Andrés Lucero et al. "Playful or Gameful? Creating Delightful User Experiences". In: *Interactions* 21.3 (May 2014), pp. 34–39. ISSN: 1072-5520. DOI: 10.1145/2590973. URL: https://doi.org/10.1145/2590973.

[78] Michael Lyvers, Samantha Cotterell, and Fred Thorberg. ""Music is my drug": Alexithymia, empathy, and emotional responding to music". In: *Psychology of Music* 48 (Dec. 2018), p. 030573561881616. DOI: 10.1177/0305735618816166.

[79] Cong Hung Mai et al. "Learning of art style using AI and its evaluation based on psychological experiments". In: *International Conference on Entertainment Computing*. Springer. 2020, pp. 308–316.

[80] Teemu Mäki-Patola et al. "Experiments with virtual reality instruments". In: *Proceedings of the 2005 conference on New interfaces for musical expression*. Vancouver, BC, Canada: New Interfaces for Musical Expression (NIME05), 2005, pp. 11–16.

[81] William P Malm. *Music cultures of the Pacific, the Near East, and Asia*. Vol. 2. New Jersey, USA: Pearson College Division, 1996.

[82] Y. Mann. *AI Duet*. 2016. URL: https://experiments.withgoogle.com/ai/ai-duet/view/.

[83] Elizabeth H. Margulis. *On repeat: how music plays the mind*. New York, NY: Oxford University Press, 2014.

[84] Elena Márquez Segura et al. "Playification: the physeear case". In: *Proceedings of the 2016 annual symposium on computer-human interaction in play*. New York, NY, United States: Association for Computing Machinery – ACM, 2016, pp. 376–388.

[85] Georgia Rossmann Martins, Mário Escarce Junior, and Leandro Soriano Marcolino. "Jikan to Kukan: A Hands-On Musical Experience in AI, Games and Art". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, Arizona, USA: AAAI, 2016, pp. 4377–4378. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12123.

[86] J. Kent McAnally. "The Bulletin of Historical Research in Music Education". In: *SAGE Publishing* 20.13 (1995), pp. 19–58.

[87] Anthony McCosker, Rowan Wilken, and Michael Arnold. *Machine Vision in Everyday Life: Public Information and Visual Analytics*. Palgrave Macmillan, 2019.

[88] Gary E McPherson. "Giftedness and talent in music". In: *Journal of Aesthetic Education* 31.4 (1997), pp. 65–77.

[89] Chase Mitchusson. "Indeterminate Sample Sequencing in Virtual Reality". PhD thesis. USA: Louisiana State University – LSU, 2020, pp. 233–236. URL: https://doi.org/10.5281/zenodo.4813332.

[90] Shigeru Miyamoto. *Mario Bros.* 1983.

[91] Julian Moreira, Pierre Roy, and François Pachet. "Virtualband: Interacting with Styslistically Consistent Agents". In: *Proceedings of the 14th International Society for Music Information Retrieval Conference*. Curitiba, Brazil: ISMIR, 2013.

[92] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. "Natural language processing: an introduction". In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 544–551.

[93] Jonas Oppenlaender. "The Creativity of Text-based Generative Art". In: *arXiv preprint arXiv:2206.02904* (2022).

[94] François Pachet et al. "Reflexive loopers for solo musical improvisation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, United States: Association for Computing Machinery – ACM, 2013, p. 4.

[95] Philippe Pasquier et al. "An Introduction to Musical Metacreation". In: *Comput. Entertain.* 14.2 (Jan. 2017). DOI: 10.1145/2930672. URL: https://doi.org/10.1145/2930672.

[96] Nikolaos Passalis and Stavros Doropoulos. "deepsing: Generating sentiment-aware visual stories using cross-modal music translation". In: *Expert Systems with Applications* 164 (2021), p. 114059.

[97] Martin Pichlmair. "Electroplankton revisited: A Meta-Review". In: *Eludamos – Journal for Computer Game Culture* 1.1 (2007), p. 5.

[98] Vincent Barreau Pierre Barreau Denis Shtefan. *AIVA*. 2016.

[99] Luke Plunkett. *AI Creating 'Art' Is An Ethical And Copyright Nightmare*. 2022.

[100] Michael A. Policastro. *Understanding How to Build Guitar Chords and Arpeggios*. Mel Bay, 1999, p. 168.

[101] Anthony Prechtl. *Adaptive music generation for computer games*. Open University (United Kingdom), 2016.

[102] Mr D Murahari Reddy et al. "Dall-e: Creating images from text". In: *Dogo Rangsang Research Journal - DRSR* (2021).

[103] Adam Roberts et al. "Interactive musical improvisation with Magenta". In: *Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems*. (Demonstration). Barcelona, Spain: NIPS, 2017.

[104] Juan G Roederer. *The physics and psychophysics of music: An introduction*. University of Alaska, Fairbanks, AK, USA: Springer Science & Business Media, 2008.

[105] Katja Rogers et al. "The Potential Disconnect between Time Perception and Immersion: Effects of Music on VR Player Experience". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Virtual Event Canada: Association for Computing Machinery – ACM, 2020, pp. 414–426.

[106] Angus Russell. *NightCafe Studio*. 2019.

[107] P. Schaeffer, C. North, and J. Dack. *In Search of a Concrete Music*. California studies in 20th-century music. University of California Press, 2012. ISBN: 9780520265745. URL: `https://books.google.com.br/books?id=6nTruQAACAAJ`.

[108] Raymond Murray Schafer. *The new soundscape*. BMI Canada Limited Don Mills, 1969.

[109] Emery Schubert et al. "Algorithms can mimic human piano performance: the deep blues of music". In: *Journal of New Music Research* 46.2 (2017), pp. 175–186.

[110] Marco Scirea et al. "Affective Evolutionary Music Composition with MetaCompose". In: *Genetic Programming and Evolvable Machines* 18.4 (Dec. 2017), pp. 433–465. ISSN: 1389-2576. DOI: 10.1007/s10710-017-9307-y. URL: https://doi.org/10.1007/s10710-017-9307-y.

[111] Stefania Serafin et al. "Virtual reality musical instruments: State of the art, design principles, and future directions". In: *Computer Music Journal* 40.3 (2016), pp. 22–40.

[112] Miguel Sicart. *Play matters*. London, England: MIT Press, 2014.

[113] Bob L Sturm et al. "Machine learning research that matters for music creation: A case study". In: *Journal of New Music Research* 48.1 (2019), pp. 36–55.

[114] Inc. Survious. *Survios*. 2018.

[115] Harmonix Music Systems. *Guitar Hero*. 2005.

[116] Koray Tahiroğlu et al. "Digital Musical Instruments as Probes: How computation changes the mode-of-being of musical instruments". In: *Organised Sound* 25.1 (2020), pp. 64–74.

[117] Kıvanç Tatar and Philippe Pasquier. "Musical agents: A typology and state of the art towards musical metacreation". In: *Journal of New Music Research* 48.1 (2019), pp. 56–105.

[118] David Temperley. *Music and probability*. USA: MIT Press, 2007.

[119] Ubisoft. *Rocksmith*. 2011.

[120] Mitchell Whitelaw. *Metacreation: Art and Artificial Life*. English. USA: MIT Press, 2004, p. 296. ISBN: 9780262731768.

[121] Duncan Williams et al. "Investigating affect in algorithmic composition systems". In: *Psychology of Music* 43.6 (2015), pp. 831–854.

[122] Duncan Williams et al. "AI and automatic music generation for mindfulness". In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society. 2019.

[123] Kun Zhao et al. "An emotional symbolic music generation system based on LSTM networks". In: *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE. 2019, pp. 2039–2043.