The Aesthetics of Disharmony: Harnessing Sounds and Images for Dynamic Soundscapes Generation

MÁRIO ESCARCE JUNIOR*, Lancaster University, United Kingdom
GEORGIA ROSSMANN MARTINS*, Phersu Interactive, Brazil
LEANDRO SORIANO MARCOLINO, Lancaster University, United Kingdom
ELISA RUBEGNI, Lancaster University, United Kingdom

This work presents an autonomous approach that explores the dynamic generation of relaxing soundscapes for games and artistic installations. Differently 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 current deep learning models in terms of quality, being recognized as human production, as if the outcome were being generated out of an endless musical improvisation performance.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → *Arts and humanities*; • **Computing methodologies** → Artificial intelligence.

Additional Key Words and Phrases: interactive musical systems, emergent art, procedural content generation, soundscapes

ACM Reference Format:
Mário Escarce Junior, Georgia Rossmann Martins, Leandro Soriano Marcolino, and Elisa Rubegni. 2023. The Aesthetics of Disharmony: Harnessing Sounds and Images for Dynamic Soundscapes Generation. (November 2023), 33 pages.

1 INTRODUCTION

Recently, we have been exposed to a buzz of AI applications that dynamically generate visual art [52, 59, 64].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 [35]. 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 [57]. 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

*Both authors contributed equally to this research.

Authors' addresses: Mário Escarce Junior, m.escarce@lancaster.ac.uk, Lancaster University, United Kingdom; Georgia Rossmann Martins, georgia@phersu.com.br, Phersu Interactive, Brazil; Leandro Soriano Marcolino, l.marcolino@lancaster.ac.uk, Lancaster University, United Kingdom; Elisa Rubegni, e.rubegni@lancaster.ac.uk, Lancaster University, United Kingdom.

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 [19, 46, 67]. 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) [51], proposing text input to generate outcomes that are generic in their output [2, 52, 59, 64].

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. These systems often require significant post-production work to address the flaws in the generated outcomes. Additionally, many artists lack the expertise to train and manipulate these models [23], 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.

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 [58, 74], 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 [11, 26, 37, 55]. However, the scenario becomes significantly 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 interface 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 intent involved 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 also 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 effectively achieved the goal of generating coherent outcomes;
- Solato is able to foster a homogeneous experience, where human evaluators acknowledged interesting composition between the emergent music and images as if it was created by humans.
- Our study showed that different systems conveyed different feelings; therefore this work offers insights for the creative process of artists, designers, and developers, identifying trends when developing music and/or image instances for games, art, and other interactive experiences through autonomous approaches.
- Given the current state-of-the-art trend research in Deep Learning, our study shows that rule-based PCG mechanics still offer valuable resources to address this work's goal. Furthermore, the findings of the present study demonstrate that the application of such techniques for art creation shows great potential for enhancing human capabilities, rather than generating concerns about the replacement of human roles.

This work'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.

2 RELATED WORK

This work is related to the study of HCI for games and artistic environments; procedural content generation; computational creativity; and the study of AI for the dynamic generation of music and art for games and interactive experiences.

HCI approaches. Interactive design approaches aim to bring daily life experiences into the virtual environment. For instance, SoundSelf: A Technodelic [20] provides the user with a guided meditation experience. Through the usage of voice input, it is possible to explore a sound-based virtual world where modulations of the user's voice create the feeling of presence in an abstract environment. Similarly, Liu et al. [40] propose a VR self-transcending flying interface that evokes a feeling of confidence that contributes to an individual's overall well-being. Previous works also explore possibilities for new musical practices to emerge through novel mechanic design. For instance, Koray et al. [70] use the notion of "cultural probe" – a technique used to gather inspirational ideas from individuals' lives – as input to digital musical instruments. These works dialogue with our approach in the way they attempt to bring "meditative" experiences to foster the emergence of a relaxing and oniric experience, as we will present further on.

In music and interactivity, Holland et al. [29] introduce past research in Music and Human-Computer Interaction, also known as Music Interaction, and discuss its role in HCI. According to the authors, current musical activities include a wide range of expressive forms, besides performance and composition; it also includes collaborative music-making and human and machine improvisation. From this perspective, specific issues may arise, including (i) whether Music Interaction should be easy for newcomers; (ii) what can be learned from the experience from those "in the groove", that is, users that already find themselves in a flow of the experience, according to Mihaly Csikszentmihalyi's definition of immersion [14]; and (iii) what can be learned from the commitment of novices in musical theory and practice with the system mechanics and interface. In performance, Astaire [75] proposes a collaborative hybrid VR dance game for two players sharing the same environment. Rostami and McMillan [63] investigate mixed-reality performances using VR, discussing how these practices can be employed in the design of interactive systems. Although the system developed through our approach does not employ a complex interactive model, it also invites the human to the emergence of its meaning, fostering particular experiences for each individual experiencing it.

Many HCI works also seek a better understanding of how music impacts players' behavior in their experience while playing games. The understatement of music's impact on human feeling, especially in its conversion with imagetic instances, is also in the scope of this work. This work aims to contribute to the analysis of the impact the current different approaches for music generation generates in the outcomes and how these are perceived by human listeners. For instance, Rogers et al. [62] assess how much music in games contributes to the player experience, more specifically considering VR approaches. In their assessment, the authors observed that players tend to perceive time passing significantly quicker when music was being played in the proposed VR experience. This is a very meaningful outcome, given the usually reduced role of audio in comparison to the visual elements in video games (and the audiovisual production). For instance, nowadays it is common to use headsets for team communication in online versus games,

reducing the role of music and its impacts on the experience [23]. Altmeyer et al. [3] investigate the impact of sound effects in a gamified image classification task. According to the author's findings, although its common usage in interactive systems, sound effects might be a less deciding factor in a user experience based on task performance. The work opens up a path for further studies of sounds in other types of interactive experiences, and we believe our work extends this analysis by verifying how music and images generated in an autonomous fashion come together according to human perception.

Additionally, Lucero et al. [44] investigate if playful interactive models based on game elements, such as gameplay mechanics, can support the creation of better engaging experiences. According to the authors, although playful mechanics for interactive experiences are important elements to foster engagement, it is often neglected. In other words, playfulness should be more explored as a feature to improve the quality of many products and processes that goes beyond entertainment-based outcomes, such as games. In the approach we will present, we aim for this effect – that is, fostering engagement in an artistic and reflexive experience that can be also extended to be used as a tool for the development of music and background scenarios for games and other immersive experiences.

Works also investigate the impacts of the convergence between HCI and AI. Inkpen et al. [35] explore a wide range of topics around HCI's engagement with AI systems. The work's discussion goes through an in-depth analysis of the AI impact in arts, especially in regard to how artworks are now being perceived. Still related to the conversion between HCI and AI, Cimolino and Graham and Escarce et al. [12, 22] discuss the concept of shared control, a paradigm in which a human and an AI partner collaboratively control a system. In our work, however, we intend to explore scenarios of autonomy for the machine to generate art, which allows us to discuss the role of humans both in their creative and contemplative capacities.

Music Generation. Works have been exploring different ways in which users interact with virtual musical instruments. These approaches are related to our work, especially in regard to the concern about how human listeners perceive the sounding outcomes generated out of the interaction between autonomous and partially-autonomous systems. For instance, Cabral et al. [8] 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 musical 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. Mitchusson [49] 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. Escarce et al. [23] propose a system for the ludic creation of music, where users are able to deal with complex musical notation elements and create music in a playful way. Gibson and Polfreman [27] 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.

Herremans et al. [28] present a functional taxonomy for music generation systems with reference to existing systems. The taxonomy organizes systems according to the purposes for which they were designed. It also reveals the inter-relatedness amongst the systems. This design-centered approach contrasts with predominant methods-based surveys and facilitates the identification of grand challenges to set the stage for new breakthroughs.

Hoover et al. [33] presents a new musical representation that allows even musically untrained users to generate polyphonic textures that are derived from the user's own initial compositions. This representation, called functional scaffolding for musical composition (FSMC), exploits a simple yet powerful property of multipart compositions, where the pattern of notes and rhythms in different instrumental parts of the same song are functionally related.

Lyvers and Thorberg [45] performed an evaluation with people by employing the GEMS (Geneva Emotional Music Scales) retrospectively by rating the felt intensity of 45 music-related emotions based on what they typically experienced when listening to their favorite music. The system is based on 9 emotion factors, which are Wonder, Transcendence, Tenderness, Nostalgia, Peacefulness, Power, Joyful Activation, Tension, and Sadness [36]. In our evaluation, we narrowed the range of feelings. For this, we attempt to identify the most prevalent moods we would like to convey to the audience and we create a kind of "grayscale" between them to support evaluators in such a subjective effort.

AI Approaches. Music and image generation are important topics in AI. Many works create music without human intervention by applying Machine Learning and Deep Learning techniques over a musical corpus [18]. Aligned with the proposal of this work, Williams et al. [72] present a Machine Learning approach for the creation of emotionally congruent music for mindfulness. Kajihara et al. [37] propose a new interactive model for the generation of soundscapes through the visualization of panoramic photos, using a crossmodal approach between sounds and images that provide an experience of hearing pseudo environmental sounds.

Biot et al. [6] 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 [69].

A common approach is also to use Deep Learning models to support the generation of music and new composition methods. E.g., Roberts et al. [60] enables the integration of generative models through the use of Google Magenta to arouse musicians' creativity. This system was tested on a live jazz performance with a piano and drums duet. Ferreira et al. [25] explore how Deep Learning models can also 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) [56] generates emotional music for media such as games, movies, and other endings. Huang and Wu [34] creates coherent music with melodic and harmonic structures that can be acknowledged as pieces composed by a human agent. Castro [10] trained Deep Learning models to perform coherent improvisations. Agarwala et al. Agarwala et al. [1] use generative Recurrent Neural Networks to create musical sheets without predefining compositional rules for the models. Jukebox [15] generates music with singing tracks in the raw audio domain. MusicLM [2] proposes a model for generating high-fidelity music from text descriptions, working in a similar fashion as stable diffusion approaches for the generation of images.

As for the dynamic generation of images, Dall-e [59] generates images from textual descriptions as input. Deepsing [55] 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.

Rodriguez et al. [24] discuss different computational techniques related to Artificial Intelligence that has 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 work 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 are able to address music and image databases in order for coherent outcomes to emerge.

Computational Creativity. Other works, however, see the creative process for the generation of artworks as a collaboration between a human and an AI system [17, 50, 53]. These systems, however, require humans to explicitly join the music generation process, also demanding some prior knowledge of music theory. Other works also explore an implicit cooperation approach for music generation. For instance, some works propose a system based on a dynamic grid where the soundtrack of a game is generated on the fly while the player is exploring a game environment [21, 48]. Carnovalini and Rodà [9] 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 picture 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. [54] 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 [71] 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. [73] mention that there has been a significant amount of work implementing systems for algorithmic composition with the intention of targeting specific emotional responses in the listener, however, a full review of this work is not currently available. This gap creates a shared obstacle for those entering the field. Lopes et al. [42, 43] 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.

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 [38], 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 [7] is a computer program designed that procedurally writes Romantic-Era style music in real-time to accompany computer games. Lopes et al. [41] 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. [30] introduce a novel approach based on evolutionary computation called functional scaffolding for musical composition (FSMC), which helps the user to explore potential accompaniments for existing musical pieces or scaffolds. Similarly, Hoover and Stanley [32] proposes 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 [31] that interactively evolves a type of artificial neural network called a compositional pattern-producing network, which represents the functional relationship between the instruments and drums. Hoover et al. [33] present a new musical representation that contains almost no built-in knowledge using Functional Scaffold that allows even musically untrained users to generate polyphonic textures derived from the user's own initial compositions. Scirea et al. [68] present the MetaCompose music generator, a compositional framework for affective music composition. In this work context "affective" refers to the music generator's ability to express emotional information.

In general, many works discussed in this section explore a wide variety of approaches from autonomous to collaborative systems between computers and humans. However, none of these works have attempted to generate complete loops of music and video instances in a dynamic fashion and on the fly, without one instance (i.e. image or video) serving as a guideline for the generation of the other. Therefore, in this work, we propose an approach where music and images are generated simultaneously and in real-time, maintaining coherency and autonomy for humans to determine and help shape the aesthetic outcome. Our approach can be employed and executed in parallel for the generation of music and video simultaneously and also independently, according to the artist or developers' needs.

3 BACKGROUND

In this section, we briefly outline important factors related to the experiences we will discuss in detail further on. We will present aspects related to the musical theory that is being approached in the system we developed and also enlighten some of the motivations behind the development of this work.

3.1 Musical Background

The knowledge of the concepts from music theory that are being approached in this work is not required to enjoy the experience with the systems we will present further on; it is being presented to support the understanding of the system mechanics we will explore ahead. **Readers that are familiar with basic music theory concepts can skip this section.**

According to some definitions [4, 39, 47, 61], 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).
- 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. There are many types of scales, and the most common ones that will be mentioned in this work 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.

3.2 Soundscapes and Concrete Music

This work is based on the concept of soundscape, as defined by R. Murray Schaeffer, whose work developed a critical sense of the sound environment that composes our daily lives. A soundscape can be comprehended as a composition of different sounds that define a given environment, whether these sounds originated from natural or synthetic sources, or even by abstract constructions [66]. It is also inspired by the concept of Concrete Music, an experimental technique for musical composition using recorded sounds of common world things as sound creation elements. This technique was also developed by the French composer Pierre Schaeffer [65].

3.3 Humans, AI, and Expressivity

In recent years, thousands of people simultaneously turned on to live streaming videos such as the Cafe Music BGM channel [5] that provides long loops of pre-composed music along with pre-existing sequences of images that aim to match the audio. It has been a valuable resource for individuals looking for interesting music as well as calm stimuli to study, meditate, and even sleep. This work was motivated by this intent and the desire to extend these experiences to an artistic installation.

Usually, streaming videos for this end present static images or short loops of animations (e.g. waterfalls, trees swaying in the wind, etc), generally preserving the same camera framing. The initial stimulus for the development of this work came from the intention to extend these approaches, promoting a stronger sense of movement in an urban landscape, that could be generated with some level of unpredictability. Similarly and in parallel, we also envisioned a way to dynamically generate a soundtrack that sets a mood for the emergent environment, in a similar way as with games and their various genres that explore dynamic changes in ambiance to generate a stronger sense of presence in the virtual environment.

As an art installation, we are also motivated to promote different levels of interactivity among people and the system. For instance, for someone just standing outside the artistic environment, we want to foster an endless musical improvisation performance that would sound pleasant. For those inside the artistic environment, watching the full landscapes emerging, our goal is to establish images and music that could be coherent and diverse across time, with unpredictability for the generation of soundscapes. In this way, people who would leave the environment and later come back would have a different experience. In addition, we would also like this music to match the images in a way that conveys an emotional experience for each individual engaged in the experience, establishing a particular dialogue with each one.

Since the system works in an autonomous fashion, we would also like to support promoting the discussion around the ethical questions related to machine autonomy in the generation of art and its implications for humans. The development of this experience fostered a territory to discuss and reflect on the new paradigm we are living in, where

AI is more present in the arts. Although autonomous, in our work we are concerned about preserving human intent and feeling behind the dynamic creation, and understanding how those are being received by a general audience.

In this way, inspired by works that established a collaboration between humans and a system [12, 22], this work intends to propose that autonomous systems may be used in a way to empower developers to improve their workflow, automating only part of the process and not the entire aesthetic product. We also evaluate the role of human expressivity in machine-generated content, and whether or not a more active presence of humans in autonomous production is perceived by humans engaged in the experience provided by such systems.

4 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 foster 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 [11, 26, 37, 55]. 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 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 user 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 [52, 59], this system does not work as a black box, where users do 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 and 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 coherency 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.

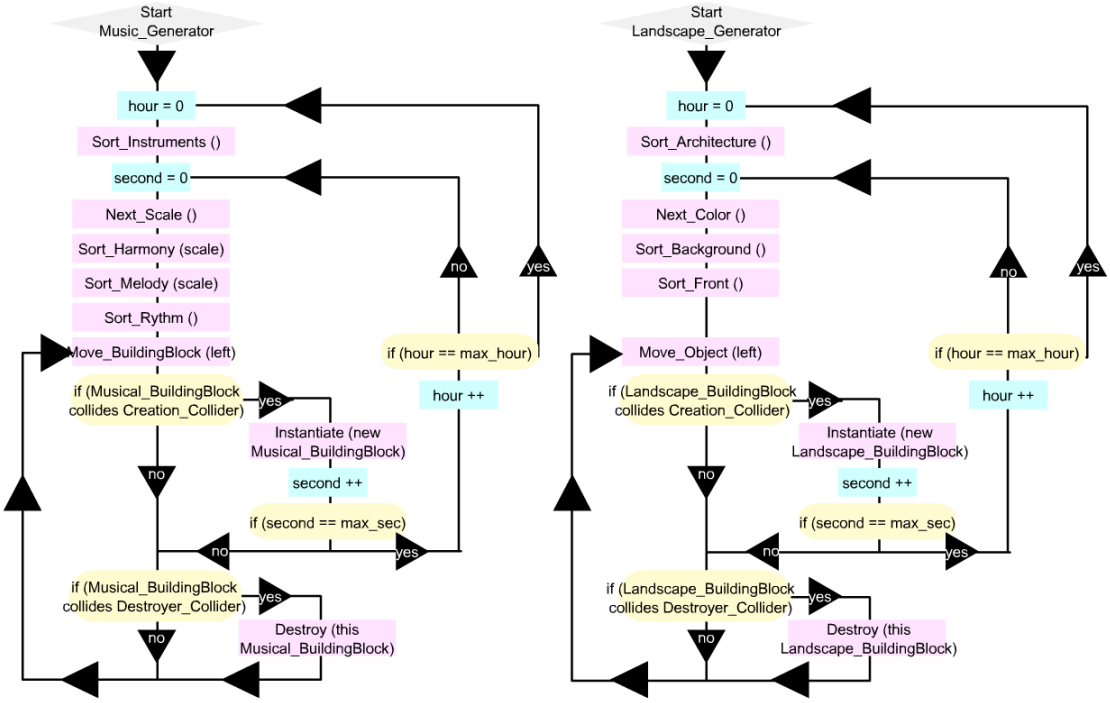

Fig. 1. Diagrams for the music generator (left) and for the landscape generator (right) that foster the emergence of soundscapes.

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 humans 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 timer.

Landscapes and music are generated as follows: the building blocks, as mentioned 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 need 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 2.C and Figure 3.C). However, as mentioned above, artists and developers can create and customize the characteristic of their building blocks at will (e.g. add more shapes, customize cell sizes, etc).

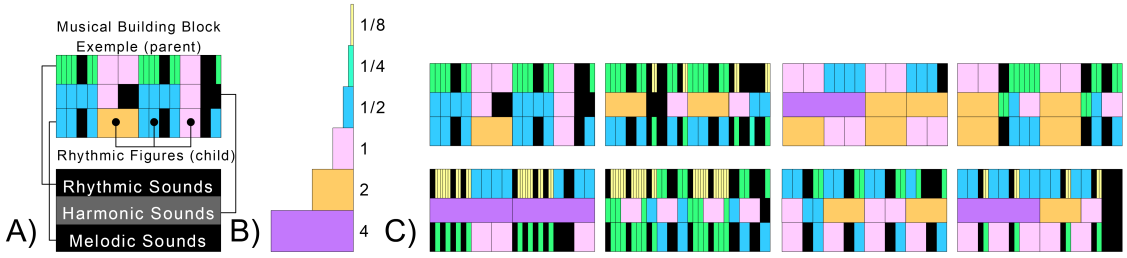

Fig. 2. 2.A) The empty Music building block. 2.B) The box triggers' sizes represent rhythmic figures. 2.C) 8 pre-determined Musical building block patterns completed with rhythmic figures presented according to the system agency.

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 2.A. Each position (i.e. back, middle, and front) contains 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 2.B. The colliders trigger different note lengths, and the arrangement of these different duration produces a rhythm of the emergent music. In the same way, it also dictates different melodic and harmonic patterns.

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).

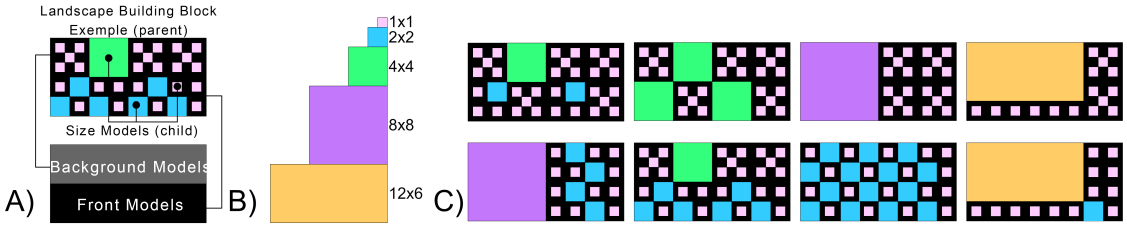

Fig. 3. 3.A) The empty Landscape building block. 3.B) The sizes of the 4 types of 3D assets that the system can generate. 3.C) 8 pre-determined patterns for landscape generation that are presented by the system.

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 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 3.B. While the 3D models (see Figure 6 for some examples from our own system's dataset) are instantiated in predetermined positions, as shown in Figure 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 5.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 work 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 destroy 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 5.B. Therefore, the palettes and the camera are actually standing still, while is the grids that 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 forwards 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, 52, 55, 59, 64].

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 2 - A and B and Algorithm 1. 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 to deal in the process to generate audio and visual elements, this composes a complex problem and a promising solution to cope with the dynamic emergence of soundscapes. In this way, this procedure can address the problem of generating audio and video instances simultaneously, however, it can also be used separately to generate music and background for games and other applications.

Resources such as our dataset of assets and source code will be released at https://github.com/MarioEscarce.

4.1 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).

Fig. 4. Screenshots of 6 different day and night moments of 3 different themes of Solato.

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 actually in a constant dialogue for the emergence of the experience's meaning. However, the system is not linear, 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 vision and meaning. Although the audiovisual instances that emerge from the system present a level of unpredictability, it does guarantee 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.

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 for a total of 4. 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 4) that is presented through a color-tone *clock*, shown in Figure 5.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,

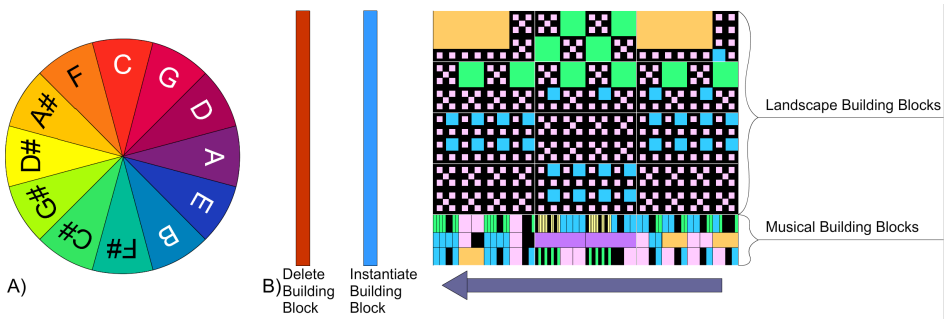

Fig. 5. 5.A) A Color-Tone clock that shows the relationship between musical tones, colors, and hours. 5.B) The deleting (red) and the instantiating (blue) palettes of the Musical and Landscape building blocks. These nonvisible objects are responsible to instantiate visual assets that are about to enter the camera frustum and also deleting them when they are off camera range.

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.

4.1.1 Image Presentation. As presented, Solato uses a color aesthetic to simulate different hours of the day, as shown in Figure 4, controlled by a color-tone clock, shown in Figure 5.A. The system uses a shader customized by the authors that render 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 3.B and 6, 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). As a reference, the 1x1 size corresponds to a $1m^2$ in real life.

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 3. In this way, we guarantee the emergence of pleasant landscapes, where bigger objects do not obstruct the view of smaller assets.

4.1.2 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 5.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

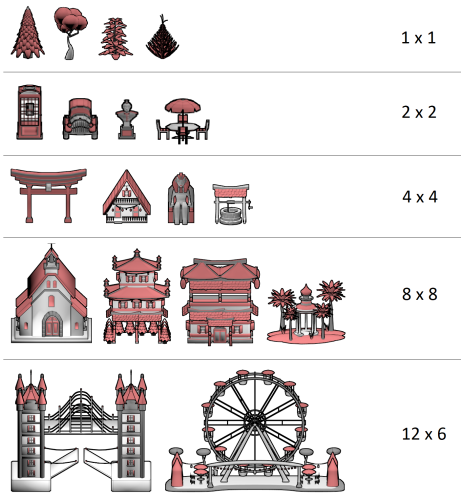

Fig. 6. Examples of low poly 3D assets from different themes managed by the system.

generate the melody and the harmony, 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 5.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, 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.

4.1.3 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.

Algorithm 1 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.

```
function OnTriggerEnter(collider)
    if collider.name == "Creation_Collider" then
        Instantiate(Musical_Object, position, Quaternion.identity)
        second += 1
        if second == 120 then
            hour += 1
            if hour == 24 then
                Sort_Instruments()
                hour = 0
            end if
            Variables()
        end if
    end if
    if collider.name == "Destroyer_Collider" then
        Destroy(this.gameObject)
    end if
end function
```

```
function OnTriggerEnter(collider)
    if collider.name == "Creation_Collider" then
        Instantiate(Landscape_Object, position, Quaternion.identity)
        second += 1
        if second == 120 then
            hour += 1
            if hour == 24 then
                Sort_Architecture()
                hour = 0
            end if
            Variables()
        end if
    end if
    if collider.name == "Destroyer_Collider" then
        Destroy(this.gameObject)
    end if
end function
```

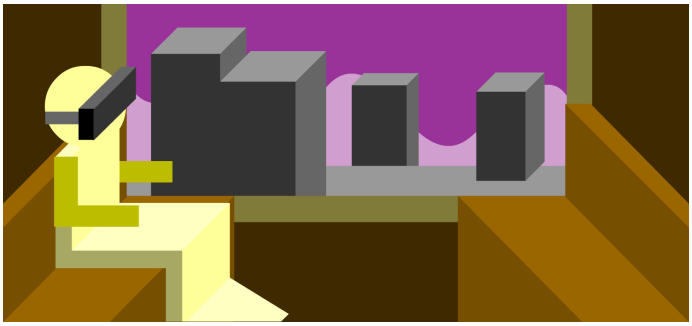

Fig. 7. Mockup of the train cabin that constrains the user view of the landscape to the desired perspective.

As mentioned earlier, the experience was designed as if the user is inside a train cabin, contemplating the generated 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 7. In this way, as the users stare all around the virtual environment, she/he/they will observe details of the interior of the train cabin, such as the train seats, doors, and luggage racks. However, when looking at the window, the generated landscapes can be observed passing by seamlessly.

5 QUALITATIVE ANALYSIS

Before we evaluate the generation of soundscapes through the approach presented in section 4 with human evaluators, we conducted an ablation study (which is an analysis in which specific components of the system are selectively removed to understand 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 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 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.

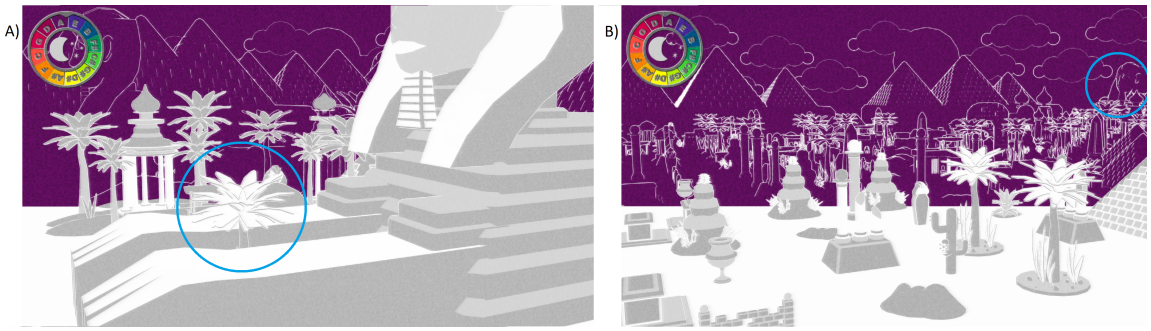

Fig. 8. Comparison between the effects our approach fosters in the generations of landscapes. The image on the left (without our approach) shows issues related to the overlap of 3D assets and visual pollution. The image on the right (with our approach) shows a more harmonious visual outcome with no asset overlap or 3D objects obstructing the background.

By running the system using the algorithm presented in Figure 1 and 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 8.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 8 (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.

To provide better visualization, we generated a timeline of 3 executions of the system running using our approach, presented in Section 4, as shown in Figure 9, 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 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

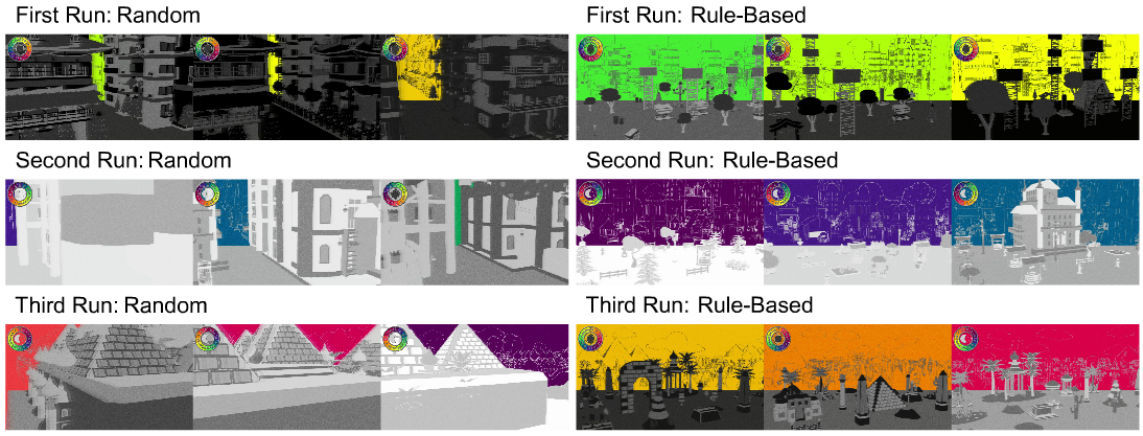

Fig. 9. Timeline of 3 different runs of the system without our approach (left) and with our approach (right).

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 [16]. For the generation of images, we used the NightCafe Machine Learning approach [64]. 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 significant 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 difference and peculiarities of the systems below.

Solato: This is the system we developed using the approach presented in Section 4 and discussed in detail in Subsection 4.1. 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. For the interested reader, a music video demonstrating our system is available at https://www.youtube.com/watch?v=P-hR1ygeMdw.

Biased Random: In this version, the building blocks that generate music and landscape have their internal shapes generated randomly. We consider there is a bias for the music generation, since the grids (i.e. building blocks) contribute to the generation of rhythmic figures and guarantees variations in the music and the images. Inside the parent building block, the same pieces of music and landscapes are randomly selected in arbitrary positions, so there may be accidental overlaps between 3D models when larger pieces appear in positions destined for smaller pieces (as presented in detail

in our ablation study, in Section 5). A Musical building block generates 3 positions of random rhythmic figures: the rhythm, the harmony, and the melody. A Landscape building block has 4 horizontal positions for different object sizes, and when a large piece is drawn on any of these positions, the landscape building block leaves this horizontal line and starts generating objects in the upcoming one. A music video demonstrating the Biased Random approach is available at https://www.youtube.com/watch?v=zpBIkMHPyoo.

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 10.A. Second, we used the Deep Learning approach MidiMe [16] 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 10.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.

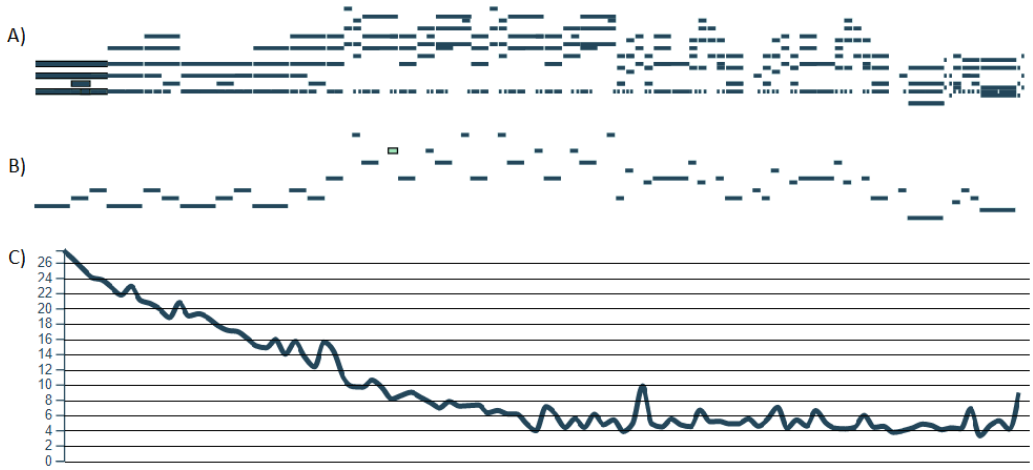

Fig. 10. A) The original music generated through Solato. B) The melodic line extracted by the original music. C) The error rate demonstrates how well the trained model reconstructed the original music melody.

In Figure 10 we can visualize and compare how the original melody used as training data (A) and the resulting melody (B) are structurally related. In Figure 10.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, then converted the MIDI file to an MP3 format.

Our goal in the generation of this external baseline was to have a work that was novel, interesting, and that have a clear link with the other approaches. In this case, it uses the same dataset of audio for generating its music and also produces colorful graphics that works in a loop, in a somewhat unpredictable manner. In this way, we had a more sophisticated musical sample from the trained version, allowing for a fair comparison between systems. As for the landscape part, we generated images using NightCafe Studio [64], a recurrent neural network system for generating visuals based on keywords. After some empirical experimentations, the keywords chosen to match the Solato experience

and generate the most suitable images were: “Landscape”, “Musical”, and “Colorful”. Thus, we generated a small video that was replicated to match the 2:30 minutes of the music. A music video demonstrating the Magenta/NightCafe approach is available at https://www.youtube.com/watch?v=DQyam87cmbc.

6.1 User Study

In our assessment sessions, a total of 74 human evaluators participated and provided feedback through an online form. 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 their attendance of sound design classes at the level of their courses, indicating their average expertise in musical composition. 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 could effectively evaluate and comprehend the emotional and artistic elements of the compositions under study. As students of art and technological-based courses, they may be able to effectively evaluate the current paradigm of AI approaches such as MidJourney [52] 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. 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.

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 from Lancaster University.**

Evaluator answered the 5 questions presented in the questionnaire below (see Table 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.

Due to time constraints, we could not run the experience in its integral length. As mentioned in section Solato, 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 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

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 statement? Completely disagree (1) to (10) Completely agree
Q5.	Which feeling predominantly describes your experience with video clip? [ ] exciting, fun; [ ] relaxing, calm; [ ] gloomy, melancholic; [ ] aggressive, hectic;

Table 1. Evaluation questionnaire for the 3 systems applied to the evaluators.

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 11) showing the feelings presented in Q5 of Table 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 previous studies have recognized that similar feelings can be categorized differently [23]. In our study, we drew inspiration from the GEMS system [45], which provided an interesting approach to 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 1. This model enables us to accommodate a wide range of interpretations, particularly for Solato, which seeks to provide a creative and relaxing experience.

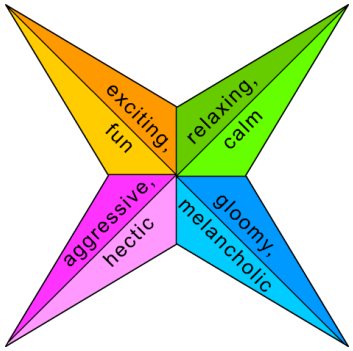

Fig. 11. Chart available in Q5 of the evaluation questionnaire, regarding the main feeling conveyed by the system.

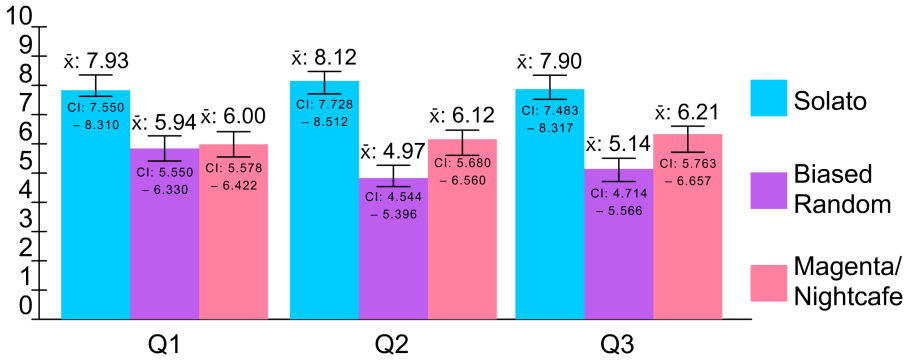

Fig. 12. Means obtained by each system in Q1, Q2, and Q3.

6.2 Results

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 12 (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 12 (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 12 (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 suggest significant differences in the ratings of the three systems, indicating variations in how well they 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.

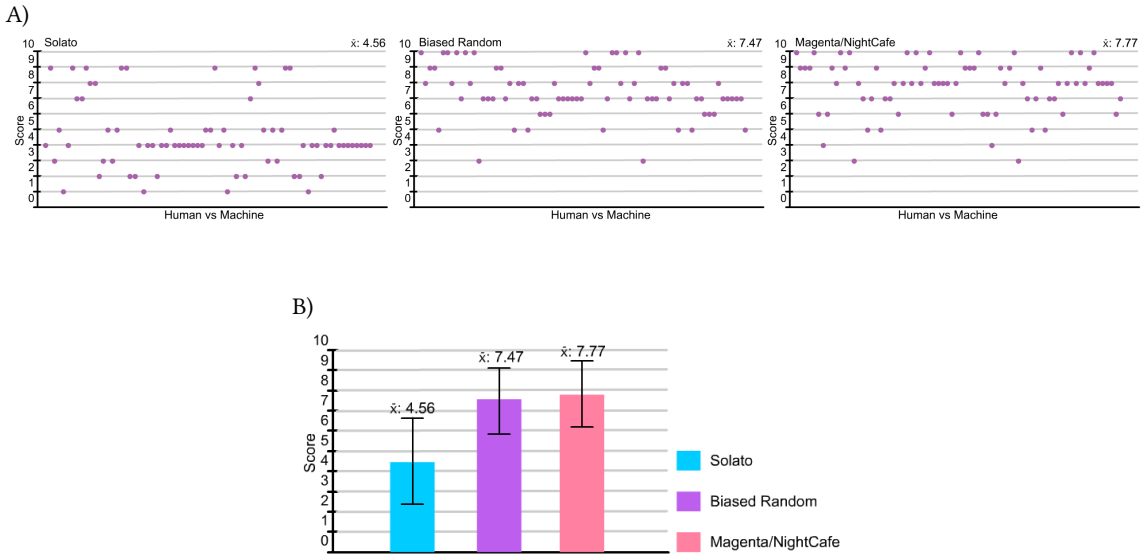

Fig. 13. A) Dispersion graph (top) showing individual scores in the Human vs. Machine relation as acknowledged by human evaluators. B) Means (bottom) showing tendencies of the systems as acknowledged by evaluators. Values between 1 to 5 show a tendency to be a **human** production, while 6 to 10 show a tendency to be a **machine** production.

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 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 13 - 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 8. 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 [59] and Midjourney [52], 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.

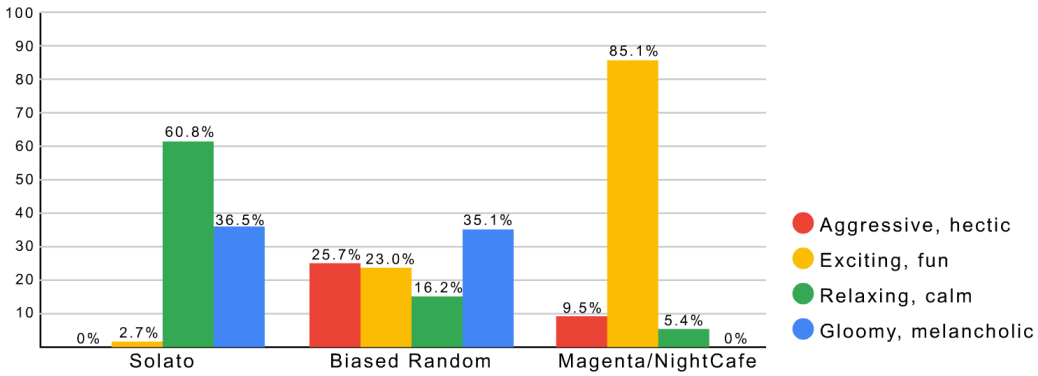

Fig. 14. Comparison between the feelings acknowledged by evaluators in Solato, Biased Random, and Magenta/NightCafe.

In **Q5**, evaluators assigned a feeling according to their perception of the soundscapes generated by the systems, as shown in Figure 14. 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" [13]. 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 "excitement/fun" (23%). Finally, "agitated/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 "agitated/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 "agitated/hectic" feelings, the system's dominant focus on "exciting/fun" was evident in the statistical analysis.

In summary, the chi-square tests offered valuable insights into the emotional qualities generated by each system. Solato demonstrated a tendency to convey the feeling of "relaxing/calm", successfully achieving its intended goal of creating soundscapes conducive to relaxation. Meanwhile, Magenta/NightCafe predominantly evoked "exciting/fun" feelings, and Biased Random displayed a diverse range of emotions due to its random nature. These results confirm our hypothesis that the preference for specific feelings in Solato and Magenta/NightCafe is not random, highlighting the distinct emotional expressions conveyed by each approach.

7 DISCUSSION

7.1 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 important 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 [23]. 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. In the midst of 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 intent involved in the production? *Our study provided significant 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 14), 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 "excitement/fun," "agitated/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.

7.2 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. [25].

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 13 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 proposed by Escarce et al. [23], 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 the Results section and depicted in Figure 14 highlight the prominent presence of the feeling of "relaxing/calm," acknowledged by a significant 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 [23], 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.

8 FUTURE WORK

Evaluate outcomes identifying expert musicians and artists: In future work, we would like to perform an evaluation with expert musicians and plastic artists to verify if we can observe variations compared to the current outcomes. In addition, expert evaluators may also provide a more sophisticated look at the outcomes produced by Solato and the other systems. For instance, since Ambient Music does not present the same common structures of other genres, we believe expert users may have more resources to decompose a musical corpus in ways to provide even more refined feedback, especially in terms of identifying human traits in autonomous production.

Generation of soundtracks and Background scenarios for games: We also intend on evaluating Solato in its capacity to generate both music and scenarios for games. We believe our approach may be efficient for the creation of background scenarios for 3D platform games, given its effect of generating urban landscapes seamlessly. Although our approach is presented running in parallel for managing images and music datasets together, we emphasize that **it can also be effective running separately**, as mentioned earlier. These scenarios could present an interesting level of unpredictability, varying according to the dataset size, supporting developers by optimizing their workflow. In addition, according to our evaluation, the different evaluated models conveyed different feelings in their music, and this can be further explored by game developers attempting to dynamically create a determined mood in the game.

In this way, this work invokes a discussion around many PCG works that support game developers and artists in order to foster an improved creation process for specific tasks, such as music creation, that may be used as background music for games [30, 33, 68]. Also important to emphasize that the goal is not to replace the importance of the artist and human expressivity, and instead to propose a collaborative process between humans and autonomous systems.

Evaluating Solato with adapted to a VR device: Although not a crucial factor for this work's goal, we intend to evaluate Solato in its VR capacity in the future. Since the evaluation was carried on remotely, this could not be taken into consideration. Some information we would like to gather from a VR evaluation would be if it improves immersion, if it causes any sort of discomfort (like motion sickness), or if the time-passing representation using our colorful mechanics through the VR device would bring any different perception compared to those that just watched the videos.

Collaborative experiences: Furthermore, Solato exhibits the potential for supporting collaborative experiences, enabling the participation of multiple users in real-time interactions that influence the emergence of audiovisual outcomes. It also offers the opportunity for artists, musicians, or participants from diverse locations to contribute to the creation of a collective audiovisual piece, even through remote participation. These collaborative features foster creativity, exploration, and cooperation, expanding the possibilities for shared artistic expression and engagement.

Interface for parameters customization: In a future iteration of Solato, we are aiming to allow users to manipulate parameters that directly impact the emergent audiovisual outcomes of Solato through a graphical interface. These parameters may encompass essential elements such as tempo, mood, color palette, visual effects, audio effects, and the balance between sound and image manifestations. By introducing intuitive interfaces or controls, artists, and developers can engage in interactive experimentation and fine-tune the system's behavior, actively participating in the creative process. This level of control not only may enhance the user experience but also facilitates the realization of their artistic vision and creative intent.

8.1 Limitations

Our work 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 also 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 sounds 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.

9 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 no 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.

Acknowledgments: The authors would like to thank Andrezza Gonzalez and Jalver Bethonico for their support in the development of this work, particularly in the recruitment process and in the development of the assessment tools used in the evaluation sessions. Special thanks are also extended to all students who participated in our user study. We are also grateful for the studentship granted by the Faculty of Science and Technology from Lancaster University, which allowed us to conduct this research. Thank you!

REFERENCES

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

[2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325 [cs.SD]

[3] Maximilian Altmeyer, Vladislav Hnatovskiy, Katja Rogers, Pascal Lessel, and Lennart E. Nacke. 2022. 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 193, 14 pages. https://doi.org/10.1145/3491102.3517581

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

[5] BGMC. 2014. Cafe Music BGM channel. https://www.youtube.com/channel/UCJhjE7wbdYAae1G25m0tHAA

[6] Jean-Pierre Briot and François Pachet. 2020. Deep learning for music generation: challenges and directions. *Neural Computing and Applications* 32, 4 (2020), 981–993.

[7] Daniel Brown. 2012. 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. AAAI, 68–72.

[8] Marcio Cabral, Andre Montes, Gabriel Roque, Olavo Belloc, Mario Nagamura, Regis RA Faria, Fernando Teubl, Celso Kurashima, Roseli Lopes, and Marcelo Zuffo. 2015. Crosscale: A 3D virtual musical instrument interface. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, IEEE Xplore, Arles, France, 199–200.

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

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

[11] Matteo Casu, Marinos Koutsomichalis, and Andrea Valle. 2014. Imaginary Soundscapes: The SoDA Project. In *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound* (Aalborg, Denmark) *(AM '14)*. Association for Computing Machinery, New York, NY, USA, Article 5, 8 pages. https://doi.org/10.1145/2636879.2636885

[12] Gabriele Cimolino and T.C. Nicholas Graham. 2022. Two Heads Are Better Than One: A Dimension Space for Unifying Human and Artificial Intelligence in Shared Control. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 8, 21 pages. https://doi.org/10.1145/3491102.3517610

[13] William G Collier and Timothy L Hubbard. 2001. Musical scales and evaluations of happiness and awkwardness: Effects of pitch, direction, and scale mode. *American Journal of psychology* 114, 3 (2001), 355–375.

[14] Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. Harper & Row, USA.

[15] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* 1, 1 (2020), 20.

[16] Monica Dinculescu, Jesse Engel, and Adam Roberts (Eds.). 2019. *MidiMe: Personalizing a MusicVAE model with user data*.

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

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

[19] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068* (2017).

[20] Andromeda Entertainment. 2020. SoundSelf: A Technodelic.

[21] Mário Escarce, Georgia Rossmann Martins, Leandro Soriano Marcolino, and Yuri Tavares dos Passos. 2017. 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)*. AAAI, Snowbird, Little Cottonwood Canyon, Utah, USA, 186–192. https://aaai.org/ocs/index.php/AIIDE/AIIDE17/paper/view/15887

[22] Mário Escarce Junior. 2022. Meta-Interactivity and Playful Approaches for Musical Composition. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 58, 4 pages. https://doi.org/10.1145/3491101.3503819

[23] Mário Escarce Junior, Georgia Rossmann Martins, Leandro Soriano Marcolino, and Elisa Rubegni. 2021. A Meta-Interactive Compositional Approach That Fosters Musical Emergence through Ludic Expressivity. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 262 (oct 2021), 32 pages. https://doi.org/10.1145/3474689

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

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

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

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

[28] Dorien Herremans, Ching-Hua Chuan, and Elaine Chew. 2017. A Functional Taxonomy of Music Generation Systems. *ACM Comput. Surv.* 50, 5, Article 69 (sep 2017), 30 pages. https://doi.org/10.1145/3108242

[29] Simon Holland, Katie Wilkie, Paul Mulholland, and Allan Seago. 2013. Music interaction: understanding music and human-computer interaction. In *Music and human-computer interaction*. Springer, London, UK, 1–28.

[30] Amy K. Hoover, Michael P. Rosario, and Kenneth O. Stanley. 2008. Scaffolding for Interactively Evolving Novel Drum Tracks for Existing Songs. In *Applications of Evolutionary Computing*, Mario Giacobini, Anthony Brabazon, Stefano Cagnoni, Gianni A. Di Caro, Rolf Drechsler, Anikó Ekárt, Anna Isabel Esparcia-Alcázar, Muddassar Farooq, Andreas Fink, Jon McCormack, Michael O'Neill, Juan Romero, Franz Rothlauf, Giovanni Squillero, A. Şima Uyar, and Shengxiang Yang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 412–422.

[31] Amy K. Hoover, Michael P. Rosario, and Kenneth O. Stanley. 2008. Scaffolding for Interactively Evolving Novel Drum Tracks for Existing Songs. In *Applications of Evolutionary Computing*, Mario Giacobini, Anthony Brabazon, Stefano Cagnoni, Gianni A. Di Caro, Rolf Drechsler, Anikó Ekárt, Anna Isabel Esparcia-Alcázar, Muddassar Farooq, Andreas Fink, Jon McCormack, Michael O'Neill, Juan Romero, Franz Rothlauf, Giovanni Squillero, A. Şima Uyar, and Shengxiang Yang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 412–422.

[32] Amy K Hoover and Kenneth O Stanley. 2009. Exploiting functional relationships in musical composition. *Connection Science* 21, 2-3 (2009), 227–251.

[33] Amy K Hoover, Paul A Szerlip, and Kenneth O Stanley. 2014. Functional scaffolding for composing additional musical voices. *Computer Music Journal* 38, 4 (2014), 80–99.

[34] Allen Huang and Raymond Wu. 2016. Deep Learning for Music. *ArXiv* abs/1606.04930 (2016), 8.

[35] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric PS Baumer. 2019. 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*. 1–9.

[36] Patrik N Juslin and John Sloboda. 2011. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press.

[37] Yuma Kajihara, Shoya Dozono, and Nao Tokui. 2017. 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*. 1–3.

[38] David Kanaga and Ed Key. 2013. Proteus.

[39] Anssi Klapuri and Manuel Davy. 2007. *Signal processing methods for music transcription*. Springer-Verlag US, United States. 440 pages.

[40] Pinyao Liu, Ekaterina R. Stepanova, Alexandra Kitson, Thecla Schiphorst, and Bernhard E. Riecke. 2022. Virtual Transcendent Dream: Empowering People through Embodied Flying in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 133, 18 pages. https://doi.org/10.1145/3491102.3517677

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

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

[43] Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. 2017. Modelling affect for horror soundscapes. *IEEE Transactions on Affective Computing* 10, 2 (2017), 209–222.

[44] Andrés Lucero, Evangelos Karapanos, Juha Arrasvuori, and Hannu Korhonen. 2014. Playful or Gameful? Creating Delightful User Experiences. *Interactions* 21, 3 (May 2014), 34–39. https://doi.org/10.1145/2590973

[45] Michael Lyvers, Samantha Cotterell, and Fred Thorberg. 2018. "Music is my drug": Alexithymia, empathy, and emotional responding to music. *Psychology of Music* 48 (12 2018), 030573561881616. https://doi.org/10.1177/0305735618816166

[46] Cong Hung Mai, Ryohei Nakatsu, Naoko Tosa, Takashi Kusumi, and Koji Koyamada. 2020. Learning of art style using AI and its evaluation based on psychological experiments. In *International Conference on Entertainment Computing*. Springer, 308–316.

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

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

[49] Chase Mitchusson. 2020. *Indeterminate Sample Sequencing in Virtual Reality*. Ph. D. Dissertation. Louisiana State University – LSU, USA. https://doi.org/10.5281/zenodo.4813332

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

[51] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association* 18, 5 (2011), 544–551.

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

[53] François Pachet, Pierre Roy, Julian Moreira, and Mark d'Inverno. 2013. Reflexive loopers for solo musical improvisation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery – ACM, New York, NY, United States, 4.

[54] Philippe Pasquier, Arne Eigenfeldt, Oliver Bown, and Shlomo Dubnov. 2017. An Introduction to Musical Metacreation. *Comput. Entertain.* 14, 2, Article 2 (jan 2017), 14 pages. https://doi.org/10.1145/2930672

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

[56] Vincent Barreau Pierre Barreau, Denis Shtefan. 2016. AIVA.

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

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

[59] Mr D Murahari Reddy, Mr Sk Masthan Basha, Mr M Chinnaiahgari Hari, and Mr N Penchalaiah. 2021. Dall-e: Creating images from text. *Dogo Rangsang Research Journal - DRSR* (2021).

[60] Adam Roberts, Jesse Engel, Curtis Hawthorne, Ian Simon, Elliot Waite, Sageev Oore, Natasha Jaques, Cinjon Resnick, and Douglas Eck. 2017. Interactive musical improvisation with Magenta. In *Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems*. NIPS, Barcelona, Spain. (Demonstration).

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

[62] Katja Rogers, Maximilian Milo, Michael Weber, and Lennart E Nacke. 2020. 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*. Association for Computing Machinery – ACM, Virtual Event Canada, 414–426.

[63] Asreen Rostami and Donald McMillan. 2022. The Normal Natural Troubles of Virtual Reality in Mixed-Reality Performances. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 132, 22 pages. https://doi.org/10.1145/3491102.3502139

[64] Angus Russell. 2019. NightCafe Studio.

[65] P. Schaeffer, C. North, and J. Dack. 2012. *In Search of a Concrete Music*. University of California Press. https://books.google.com.br/books?id=6nTruQAACAAJ

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

[67] Emery Schubert, Sergio Canazza, Giovanni De Poli, and Antonio Rodà. 2017. Algorithms can mimic human piano performance: the deep blues of music. *Journal of New Music Research* 46, 2 (2017), 175–186.

[68] Marco Scirea, Julian Togelius, Peter Eklund, and Sebastian Risi. 2017. Affective Evolutionary Music Composition with MetaCompose. *Genetic Programming and Evolvable Machines* 18, 4 (dec 2017), 433–465. https://doi.org/10.1007/s10710-017-9307-y

[69] Bob L Sturm, Oded Ben-Tal, Úna Monaghan, Nick Collins, Dorien Herremans, Elaine Chew, Gaëtan Hadjeres, Emmanuel Deruty, and François Pachet. 2019. Machine learning research that matters for music creation: A case study. *Journal of New Music Research* 48, 1 (2019), 36–55.

[70] Koray Tahiroğlu, Thor Magnusson, Adam Parkinson, Iris Garrelfs, and Atau Tanaka. 2020. Digital Musical Instruments as Probes: How computation changes the mode-of-being of musical instruments. *Organised Sound* 25, 1 (2020), 64–74.

[71] Kıvanç Tatar and Philippe Pasquier. 2019. Musical agents: A typology and state of the art towards musical metacreation. *Journal of New Music Research* 48, 1 (2019), 56–105.

[72] Duncan Williams, Victoria Hodge, Lina Gega, Damian Murphy, Peter Cowling, and Anders Drachen. 2019. AI and automatic music generation for mindfulness. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society.

[73] Duncan Williams, Alexis Kirke, Eduardo R Miranda, Etienne Roesch, Ian Daly, and Slawomir Nasuto. 2015. Investigating affect in algorithmic composition systems. *Psychology of Music* 43, 6 (2015), 831–854.

[74] Kun Zhao, Siqi Li, Juanjuan Cai, Hui Wang, and Jingling Wang. 2019. An emotional symbolic music generation system based on LSTM networks. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE, 2039–2043.

[75] Zhuoming Zhou, Elena Márquez Segura, Jared Duval, Michael John, and Katherine Isbister. 2019. Astaire: A Collaborative Mixed Reality Dance Game for Collocated Players. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Association for Computing Machinery – ACM, New York, NY, United States, 5–18.