Embodied exploration of deep latent spaces in interactive dance-music performance

Sarah Nabi, Philippe Esling, Geoffroy Peeters, Frédéric Bevilacqua

Embodied exploration of deep latent spaces in interactive dance-music performance

Sarah Nabi
sarah.nabi@ircam.fr
UMR 9912 STMS-IRCAM-CNRS-Sorbonne Université
Paris, France

Philippe Esling
philippe.esling@ircam.fr
UMR 9912 STMS-IRCAM-CNRS-Sorbonne Université
Paris, France

Geoffroy Peeters
geoffroy.peeters@telecom-paris.fr
LTCI, Télécom-Paris, Institut Polytechnique de Paris
Palaiseau, France

Frédéric Bevilacqua
frederic.bevilacqua@ircam.fr
UMR 9912 STMS-IRCAM-CNRS-Sorbonne Université
Paris, France

ABSTRACT

In recent years, significant advances have been made in deep learning models for audio generation, offering promising tools for musical creation. In this work, we investigate the use of deep audio generative models in interactive dance/music performance. We adopted a performance-led research design approach, establishing an art-research collaboration between a researcher/musician and a dancer. First, we describe our motion-sound interactive system integrating deep audio generative model and propose three methods for embodied exploration of deep latent spaces. Then, we detail the creative process for building the performance centered on the co-design of the system. Finally, we report feedback from the dancer's interviews and discuss the results and perspectives. The code implementation is publicly available on our github[1].

CCS CONCEPTS

• **Human-centered computing** → **Sound-based input / output**; **Gestural input**; **Auditory feedback**; **Collaborative interaction**; • **Applied computing** → **Sound and music computing**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

dance-music-AI performance, HCI, motion-sound interaction, deep learning, generative models, embodied exploration, latent space

ACM Reference Format:
Sarah Nabi, Philippe Esling, Geoffroy Peeters, and Frédéric Bevilacqua. 2024. Embodied exploration of deep latent spaces in interactive dance-music performance. In *9th International Conference on Movement and Computing (MOCO '24), May 30-June 2, 2024, Utrecht, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3658852.3659072

[1]https://github.com/ircam-ismm/embodied-latent-exploration

MOCO '24, May 30-June 2, 2024, Utrecht, Netherlands

ACM ISBN 979-8-4007-0994-4/24/05…$15.00
https://doi.org/10.1145/3658852.3659072

1 INTRODUCTION

In recent years, Artificial Intelligence (AI) for generative tasks has achieved impressive results. This trend in generative AI has now begun to influence creative endeavours among both the music [17] and dance communities [3] alike. These models have also raised ethical inquiries within both the artistic and research communities [11, 34] regarding data collection [44], environmental impact of their large computational costs [14], and engineering AI approaches to autonomously perform human musical tasks [17, 44]. Regarding this last concern, we aim to foster *critical design practices* [11] to investigate generative AI in creative workflows by directly including artists in the design process.

In this work, we leverage an art-research collaboration with the dancer/choreographer Marie Bruand (M.B.) to investigate the creative potential of deep audio generative models in interactive dance/music performance. Our goal was to experiment how deep generative models could be 'creatively' explored through embodied interaction and, in particular, how different motion-sound interaction strategies could be built with such models. For this, we adopted a *performance-led research* design approach [2] to conjointly build the system and create a live performance. This project also allows us to initiate discussion and reflections on dance-musical-AI practices.

From an AI perspective, we are conducting research on deep audio generative models. These models typically involve a time-consuming learning phase based on a large set of recorded sound and music material. Nevertheless, once trained, some model such as RAVE [8] enable real-time high-quality sound synthesis with low latency on standard laptop CPU, making it suitable for music performances. Specifically, the learning process produces a parametric representation of the sound database, called *latent space*, which can be used for real-time sound control. However, controlling such a synthesis process is challenging as these latent representations are very abstract and generally too high-dimensional to be directly interpretable. Hence, existing approaches usually rely on bypassing the latent information through pre-defined control attributes to explicitly condition the generation [13, 43]. However, these methods require massive sets of labeled examples and fail to address the need for intuitive and personalized control. Interestingly, RAVE embeds the possibility to work with reduced latent spaces that can be directly controlled through external sensors, which opens stimulating opportunities for interaction.

Hence, alternatively to *conditioning* methods, we propose to develop strategies for what we call an *embodied exploration* of the latent space, considered as a *raw* sound design space. The general aim is, therefore, to examine how we can discover and learn these abstract latent parameters from a bodily and sensitive perspective. We hypothesize that such an embodied approach can stimulate new music-dance-AI design practices towards creative processes.

Other recent works have also proposed to explore deep audio latent spaces for creative endeavours. These relied on interpolation strategies with audio recordings [46], soundwalking designs to directly explore RAVE latent spaces in virtual world [42], or interactive machine learning approaches to iteratively craft a mapping between the RAVE latent space and a 2-dimensional performance space for gestural control [47]. Yet, to the best of our knowledge, this is the first attempt to investigate embodied latent space exploration in designing an interactive dance/music performance.

Our contributions can be summarized as follows:

- We propose a motion-sound interactive system integrating deep audio generative model and describe three embodied interaction methods to explore deep audio latent spaces. The code implementation is available on our github[2].
- We report a documented art-research collaboration to co-design the interactive system and create the performance.
- We discuss the use of interactive deep audio generative models on dance-musical-AI practices, highlighting current limitations and open questions to investigate in future works.

First, we introduce the background and related works (Sec. 2) and explain the multiple objectives of this art-research collaboration (Sec. 3). Then, we introduce our motion-sound interactive system and the three embodied exploration methods (Sec. 4). Finally, we detail the methodology used to create the performance (Sec. 5) and discuss the use of generative AI in artistic practices (Sec. 6).

2 BACKGROUND

This project falls within several research fields as it aims to incorporate deep audio generative models into motion-sound interactive systems for a dance/music performance. We relied on embodied interaction design approaches from the HCI community systems [32].

2.1 Dance and technology

Since the pioneering piece of Merce Cunningham's *Variation V* (1965) [26], a tremendous amount of works linking dance and technology have been proposed, fostering collaboration among artists, researchers and engineers. These works span various artistic and scientific communities, such as HCI communities and the MOCO conferences, and relate to an extensive body of literature that we cannot fully cover here [5, 9, 39, 50]. Among many others, Giomi provides an interesting review on interactive music/dance systems in [26] and investigates somatic sonification and sensorimotor learning through the use of interactive auditory feedback in dance practices. The third wave of HCI [6] emphasizes expressive and embodied forms of interaction, and stimulated the development of new methodologies for the design of movement-based interactions [23]. Theses advances in interaction design have significantly impacted the dance community. Benford et al. [2] introduced the *performance-led research in the wild* design, which states that performance enables to experiment and study how humans interact with technologies. This method has been widely adopted to study the impact of technology, and especially interactive systems, in dance performance and practices in several recent papers directly related to our research [1, 4, 19, 23, 28].

Recently, the advent of AI for creative endeavors through deep generative models has received increasing attention among both artists and researchers [10, 17].These approaches have also raised several ethical and environmental concerns [11, 14], questioning their integration into artistic design practices. Recently, generative AI has gained interest among the dance community [3]. Our work falls within this growing research field and aims to investigate embodied interaction with deep audio generative models in interactive music/dance systems through performance-led research design.

2.2 Motion-sound interactive systems

Over the past decades, the concomitant advances in wearable sensing technologies, such as *Inertial Measurement Unit* (IMU) sensors, and sound synthesis techniques have paved the way to the development of motion-sound interactive systems. Designing the mapping between movement and sound is essential for interactive audio applications and extensively studied in the *New Interfaces for Musical Expression* (NIME) community, involving digital musical instrument design [25, 35, 49] possibly using machine learning [21, 30, 45].

In initial gesture-sound interfaces, artists and performers relied on *explicit* wiring of sound synthesis parameters to manually-selected gestural inputs [38]. Explicit mapping techniques have then evolved towards *implicit* strategies. These rely on an intermediate model that learns the motion-sound relationships directly from examples [29]. Inspired from early work in HCI [18], various frameworks of *Interactive Machine Learning* (IML) enable users to design data-driven interactions and build custom gesture recognition and sonification systems [20, 22, 47, 48]. The user-centered *Mapping-by-Demonstration* approach relies on IML to design the motion-sound mapping from user demonstrations using the action-perception loop [24]. This approach consists of two phases. During the *training* phase, the user synchronously records sounds and movements to constitute a set of multimodal sequences composed of temporally-aligned features extracted from both signals. A set of Hidden Markov Regression (HMR) models [22] are trained on these examples to capture the temporal dynamics and variations of the movement-sound relationship. During the *performance* phase, the system interprets the user's movement in real-time. A pre-trained classifier selects the related HMR model, which continuously estimates the associated sound features to re-synthesize the pre-recorded sample using corpus-based synthesis [41].

2.3 Latent-based audio generative models

Since the pioneering autoregressive model WaveNet [36], significant advances have been made in deep generative models for raw waveform synthesis [8, 12, 15, 16]. In particular, the RAVE model [8] allows to generate high-quality audio samples in real-time with low latency on standard laptop CPU. These methods rely on *latent-based* generative models such as Variational Auto-Encoders

[2]https://github.com/ircam-ismm/embodied-latent-exploration

(VAE) [31] and Generative Adversarial Networks (GAN) [27]. The goal is to model the underlying data distribution $p(\mathbf{x})$ of a given set of training examples $\mathbf{x} \in \mathbb{R}^{d_x}$ to generate new samples with similar properties. To do so, they introduce *latent* variables $\mathbf{z} \in \mathbb{R}^{d_z}$ in a lower-dimensional space ($d_z \ll d_x$), assumed to be responsible for most of the variations in $\mathbf{x}$. Hence, this so-called *latent space* can provide high-level features to condition the generation process. VAE provides a trainable analysis-synthesis framework using two parametric neural networks. The *encoder* infers a latent representation by approximating $p(\mathbf{z}|\mathbf{x})$, while the *decoder* models $p(\mathbf{x}|\mathbf{z})$ to generate new data from a given $\mathbf{z}$. RAVE combines VAE with adversarial fine-tuning to compress the waveform into a continuous latent space in which signals can be sampled at approximately 20Hz. Once trained, we can directly sample latent trajectories to generate new sounds similar to those of the training set, or perform *timbre transfer* by providing another audio input which is re-synthesized by our model. For instance, if we trained RAVE on violin samples and use a drum loop as input, the decoder will produce a new violin sound with the same rhythmic pattern.

However, controlling these models remains challenging. The latent representation is very abstract and still too high-dimensional to be directly interpretable, which precludes straightforward and intuitive control. To address this, existing approaches mainly consist in *conditioning* using additional control inputs [13, 16, 43] or *disentanglement* techniques to directly model the assumed independent underlying factors of variations [33, 37]. However, these methods require massive sets of labeled examples and prior assumptions.

As the latent variables highly depend on the initial dataset, an alternative approach would be to embrace exploration strategies to search the latent space and discover the learned synthesis parameters. Some recent works have begun to investigate audio latent space exploration from a design perspective. Tatar et al. [46] proposed three interpolation strategies using audio recordings to leverage raw audio VAE in live coding performance. Scurto and Postel [42] designed a virtual environment to explore RAVE latent space through soundwalking. Finally, Vigliensoni and Fiebrink [47] leaned on IML to map RAVE latent space to a 2-dimensional performance space from user demonstrations. These approaches rely on the nn~ external[3] to use pre-trained RAVE models in Max/MSP for real-time AI audio processing. As the learned representation is still too high-dimensional for human-machine interaction (with 128 latent variables), Caillon and Esling [8] introduced a post-training analysis to only keep the most informative latent variables and expose only 4 to 8 input control dimensions. Three modalities of interaction with RAVE are initially proposed in nn~. First, we can explore different latent parameter combinations to generate sounds with a slider for each informative dimension. Second, we can perform live timbre transfer using a live audio stream or an offline audio file as input. Finally, we can do unconditional generation with a temporal auto-regressive model pre-trained on latent representations, that autonomously produces an audio stream. Thanks to nn~, RAVE has been actively used among both artists and researchers. Yet, to the best of our knowledge, this is the first attempt to bind it with motion sensors and study embodied latent space exploration for musical creation in interactive dance/music performance.

[3]https://github.com/acids-ircam/nn_tilde

3 THE RESEARCH AND CREATION CONTEXT

3.1 Artistic objectives and collaborative context

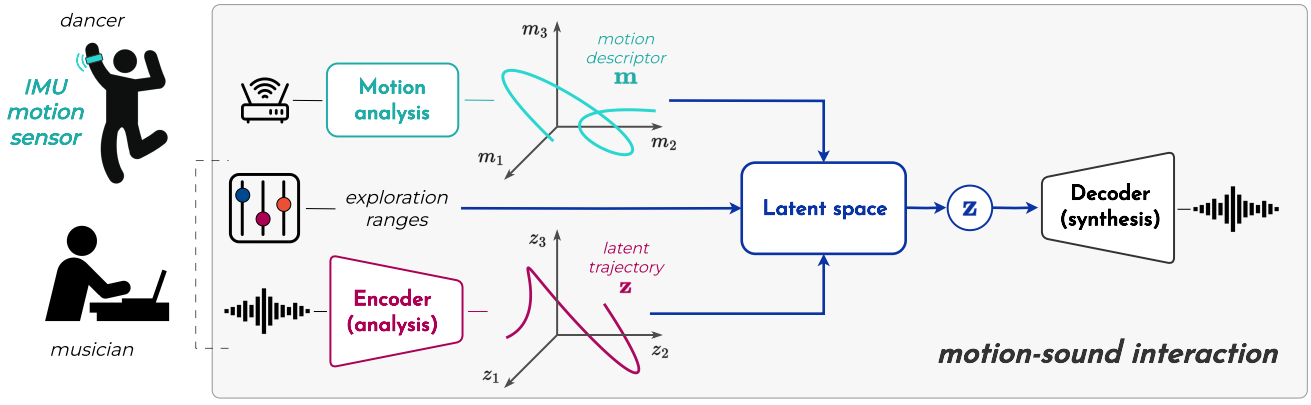
The artistic project was conceived and submitted as an art/science collaboration, between the first author of the paper (AI researcher and musician) and the dancer/choreographer Marie Bruand with practice in contemporary dance, but without any prior knowledge on AI nor motion-sound interactive systems. The initial shared goal of the performance was based on the idea of generating sounds and music through the dancer's movements. Both the researcher/musician and dancer were interested in exploring connections between music and dance, and, in particular, investigating the musicality of the moving body. Thus, the aim was not only to merely build a movement-based musical interaction but also question how such a system could offer new perspectives on movement and sound perceptions for both the performers and the audience: is it possible to 'listen' to movement as well as to 'watch' music ? After such early discussions, a live dance/music performance was submitted and accepted for the art festival *Nuit Blanche* in Paris. It also provided an interesting venue to engage in a broader discussion regarding the use of technologies in musical and dance practices, in particular, how the use of AI, here deep generative sound synthesis, could bring novel perspectives in the dance, music & technology field.

3.2 General scientific research questions

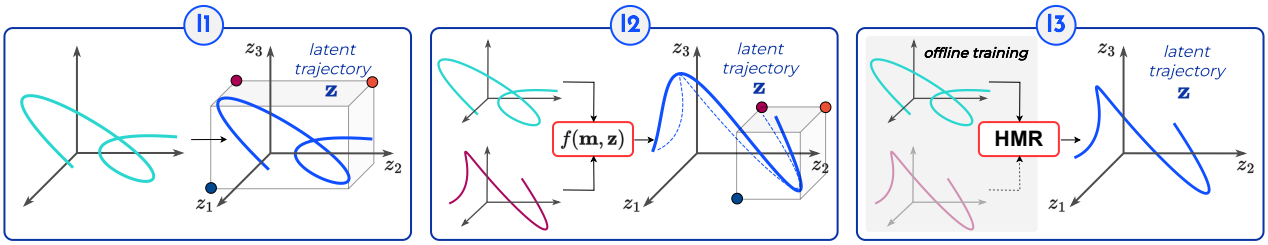
This art-research collaboration was also associated to the research topic of the researcher/musician on the use of deep audio generative models for musical creation in live performance (first author PhD project, co-supervised by the authors bringing their expertise in movement-sound interaction, audio technology and deep learning). As explained previously, these models can be controlled through the use of an abstract latent space. In terms of human-machine interaction, it is intriguing to investigate whether and how such an abstract parametric space could be associated to a movement parameters space and explored through embodied interaction. Therefore, scientific research questions concern both the development of the technological system 1) how can we associate motion sensing and deep generative sound synthesis system ? and questions relative to human-machine interaction 2) how can such a system be explored, learned and played through movement ?

3.3 Methodology

We followed a practice-based approach during three months. The co-design of the interactive system and of the piece was at the core of the creative process. The researcher/musician and dancer iteratively alternated phases of technological development, experimentation and co-creation of the piece throughout the 3-month collaboration, with the support of the supervisors and research team. Moreover, the performance and creative choices were made iteratively, in parallel with the building of the interactive system, collaboratively refining various specific aims at each step. Thus, we can describe our approach as a performance-led research design approach [2]. Importantly, we documented our process though notes and discussions between the researcher/musician and dancers. In particular, we recorded interviews of the dancer after each experimentation session and after the performance.



(a) Overview of our motion-sound interactive system



(b) Latent space exploration methods

Figure 1: Overall workflow of our motion-sound interactive system (a), where we propose several latent space exploration methods (b) in order to control the generation of a deep model through motion analysis of a dancer in real-time.

4 MOVEMENT-SOUND INTERACTIVE SYSTEM AND INTERACTION METHODS

Here, we describe the general strategies and implementation of our motion-sound interactive system integrating deep audio generative models. First, we introduce three interaction methods to perform embodied exploration of deep audio latent spaces (see Fig. 1). The proposed interaction methods are generic and can be applied to VAE-based model. Second, we detail the implementation including our choices of motion sensing, analysis and deep synthesis models.

4.1 Movement-sound interaction design

As depicted in Fig. 1, the dancer and musician both interact with the system, performing therefore what we call live *co-exploration*. The dancer is equipped with wireless motion sensors composed of accelerometers and gyroscopes. The dancer's movements are analysed in real-time to compute a set of motion descriptors, which are mapped to the latent space, depending on the different *interaction methods* described below. The resulting trajectories in the latent space are then processed in real-time by the pre-trained decoder, which generates the final audio outcome.

I1 - Interaction 1: "direct motion exploration". A first 'blind' approach corresponds to directly mapping motion descriptors to parameters of the audio latent space. We called such an approach as 'blind' since the dancer has *a priori* no information about how certain movements could generate specific sounds, and thus she must discover it through her own movement. Nevertheless, the musician can guide the dancer in such an embodied exploration process by choosing 1) specific area of the latent space, i.e. such as specific latent space parameters to explore, 2) adapting practical ranges of variations for each latent dimension.

Importantly, the latent space should be controlled using specific dynamic parameters variations to produce specific learned sounds. As human gestures are generally quite different to such variations, this approach typically produces sounds that differ substantially from those in the original database. Specifically, static poses by the dancer result in sound textures with a granular synthesis effect.

I2 - Interaction 2: "local exploration around existing latent trajectories". In this approach, the idea is to gesturally control the variation of a given sound that establishes the temporal structure. For this, a specific audio sample is selected by the musician, and the pre-trained encoder produces the corresponding latent trajectory. The dancer can then modify the sound textures of the original sound with her movements by exploring locally around the inferred latent representation. Similarly to **I1**, the musician can manually manage the exploration ranges. As the audio analysis and synthesis processes are managed independently, we could also mix different sound models to perform more hybrid synthesis.

I3 - Interaction 3: "implicit mapping between motion and latent trajectories". This approach was develop in order to predicatively choose a given sound as well as a given movement. For

this, we adapted the *Mapping-by-Demonstration* IML framework [22, 24] to implicitly map a motion descriptors trajectory to a specific latent trajectory. Contrary to **I1** and **I2**, which focuses on sound textures, this method allows to directly generate musical phrases with pre-choreographed movements. These movements, are set based on recording performed by the dancer directly, while listening to a specific original sound. Technically, the mapping is done offline by training a Hidden Markov Regression (HMR) model on paired examples [22]. The regression model then implicitly maps the temporally-aligned motion and latent signals features. Once trained, the dancer can reproduce the recorded movement, entirely or partially, to generate the associated latent trajectory that is then processed by the decoder in real-time. Similarly to **I2**, it can be combined with live timbre transfer and hybrid analysis-synthesis.

4.2 Interactive system implementation

We implemented our motion-sound interactive system in Max/MSP using real-time motion analysis and deep audio processing objects. Max/MSP patches are available on our github[4].

Motion capture and analysis. We used 4 R-IoT[5] IMU motion sensors composed of accelerometers and gyroscopes. We relied on the *Gestural toolkit*[6] with `mubu` to monitor each sensor live stream and compute motion descriptors in real-time. The motion frame rate was set to 10ms. Our interface allowed to compute 5 motion descriptors: *accelerometers*, *gyroscopes*, *bandpass* (biquad filter on accelerometers data), *orientation* (combining accelerometers and gyroscopes data with a complementary filter) and *movement intensity* (based on low-pass filtering of the without-gravity acceleration norm). The signal is 3-dimensional (1-dimensional for *intensity*). We used two sensors simultaneously for **I1** and **I2**, and one for **I3**.

Deep audio synthesis. Among state-of-the-art deep synthesis models, we decided to rely on RAVE [8] which enables fast and high-quality raw waveform synthesis in real-time setups. The models are first trained off-line and then integrated into the system. We trained RAVE models on custom datasets that we built by manually selecting Youtube links with the Common Creative licence (see Sec. 5). Each dataset must contain approximately 10 hours of audio with a sufficient recording quality to ensure high-quality synthesis. We worked with audio sampled at 44100Hz and used the official RAVE implementation[7] to process the data, train and export the models with the "v1"[8], "causal" and "streaming" configuration with a compression ratio of 2048 leading to a latent space sampling rate of around 21.5Hz. A reduced set of latent dimensions are obtained with the post-training analysis from [8]. As it relies on Principal Component Analysis (PCA), the latents are sorted in descending order of explained variance, leading to imbalanced dimensions with the first entry corresponding to the most informative one. The fidelity value was set to 0.95 resulting in 8 latent dimensions for each model. We imported our models in Max/MSP using `nn~`[9] for live audio processing, and manually set the buffer size to 1024.

Motion-sound interaction mappings. The motion analysis and mappings were designed using the *MuBu* library[10] [40].

For **I1** and **I2**, we used an interactive matrix object so that the motion-sound explicit mappings could be easily changed. The musician exploration parameters correspond to sliders and knobs mapped to a MIDI controller. We used two motion sensors simultaneously to control the 6 most informative latent dimensions.

For **I3**, we used only one motion sensor with the *orientation* descriptor to capture the movement. First, we selected audio samples, then, we synchronously recorded the input motion-sound paired example in Max/MSP using `mubu`. The dancer could listen to the audio synthesized by the pre-trained decoder while performing the gesture. We worked with short movement segments of 4 seconds. We used all the informative latent dimensions for the mapping, 8 in our case. As the sampling rate of the RAVE latent space is set to 21.5Hz, we also manually resampled the latent trajectory to align with the motion frame rate set to 10ms (i.e. 100Hz). After that, we trained a different HMR model for each mapping directly in Max/MSP using the `mubu` implementation XMM library. We manually parameterized the HMM and used 10 states per gesture and a relative regularization of 0.2 based on informal iterative testing.

5 BUILDING THE PERFORMANCE

The co-design of the system was at the core of the creation process. Here, we describe the performance and the working methodology.

5.1 The performance *PRELUDE*

Figure 2: *PART 1 - "Awakening"*: The dancer, (falsely) wired to electrical connections, realizes the power of her body over the music and completely loses control of it. Her movements gradually distort the music until it is completely erased, giving way to the sounds of the body liberated from the wires.

PRELUDE is a 20-minutes live dance/music performance where the dancer M.B. produces sounds in real-time through her movements. The piece unfolds in three parts illustrated in Fig. 2, Fig. 3 and Fig. 4. Supplementary materials and videos are available online[11].

The piece unfolds a metaphoric 'liberation' of the dancer's body. Connected at the beginning of the piece with fake cables, the dancer progressively embraces a new 'musical body'. It stages diverse

[4]https://github.com/ircam-ismm/embodied-latent-exploration
[5]https://ismm.ircam.fr/riot/
[6]https://github.com/ircam-ismm/Gestural-Sound-Toolkit
[7]https://github.com/acids-ircam/RAVE
[8]We did not use the "v2" as it suffered from posterior collapse at that time.
[9]https://github.com/acids-ircam/nn_tilde
[10]https://ismm.ircam.fr/mubu/
[11]https://ircam-ismm.github.io/embodied-latent-exploration/

Figure 3: *PART 2 - "Introspection"*: The dancer discovers and explores the sounds produced by her movements. Carefully listening to each part of her body, she embraces this new musical body and begins to build her instrument.

Figure 4: *PART 3 - "The musical body"*: The dancer becomes a living synthesizer. She regains control of her body and uses it to create live music.

qualities of embodied exploration of sound spaces as she navigates them through her movements under the guidance of the musician.

5.2 Creative process and working methodology

We organized working sessions and rehearsals into three phases.

1) Co-designing the movement-sound interaction based on existing platforms: First, we experimented with existing computational platforms for both motion capture and analysis (i.e. `mubu`) and deep audio synthesis (i.e. `nn~`). The goal was to establish the technical feasibility of different methods in movement-sound interaction and decide the overall trajectory of the performance.

We implemented an initial prototype of the interactive system in Max/MSP that combined real-time movement sensing and analysis connected to deep audio synthesis. As our goal at this point was solely to explore different movement-sound interaction possibilities with this system, we directly used publicly available pre-trained RAVE models provided by Antoine Caillon[12] . We implemented and tested different types of interactions with the audio latent space using one, and then two, motion sensors, placed on various parts of the dancer's body. We tested each motion descriptors with different mappings and models. We decided to keep only the *orientation* and *intensity* descriptors for the performance with motion sensors attached on top of the hands and at ankle level, as it produced results that the dancer felt comfortable with. This phase took us approximately three weeks, that includes the system development, the 3 days of co-exploration with the dancer in a studio and the collective discussions and decisions regarding the performance. At the end, we had defined the three modalities of interaction introduced in Sec. 4, disregarding some other attempts. We decided to structure the performance into three parts, one for each exploration method, in the following order (**I2**, **I1**, **I3**).

2) Aligning technology with artistic choices about sound material: Secondly, we focused on iteratively crafting the sound spaces while refining the interaction approach in order to align both technical constraints and artistic decisions regarding the performance story.

We wanted to have a sufficiently wide range of sounds to work with. Hence, we targeted percussive, ambient and melodic sounds from both electronic textures and acoustic instruments. We faced the challenge of building datasets to train our RAVE models on, which involved two main issues. First, we needed to collect approximately 10 hours of audio samples per model. To address this, we manually selected and downloaded audio recordings sampled at 44100Hz from Youtube links with the Common Creative licence. After several attempts with mixed audio databases, we realized that RAVE was not able to handle simultaneously a wide variety of polyphonic sounds into one latent space, which would require to model different levels of complexity. Hence, we finally decided to work with three distinct audio models. The first one was trained on techno live sets to create percussive sounds, the second on synthwave ambient sounds and the third on violin, alto and cello solo recordings of Bach music. Although this clear separation eased the creation of the latent spaces, we had to iteratively refine each model several times before it suited the performance setup. As the model training phase requires approximately one week, time constraints was very challenging. It was important that both the musician/researcher and the dancer could test and assess together the synthesis models using the interactive system, and make both artistic and technical adjustments.

After six weeks, while not totally finished , we decided to stop this phase considering enough material was gathered.

3) Creating the final piece form and rehearsing toward live performance: During the last three weeks, we focused on developing and rehearsing each part of the performance using the interactive system and the pre-trained RAVE models we had designed. We refined our artistic choices regarding the plot and stage design. Although structured with the choice of specific interaction method and audio spaces, each part contained a varying degree of improvisation for both dancer and musician. Hence, such live co-adaptations was practiced to build confidence for the musician and dancer to interact together *through* and *with* the system.

For part 1, which involved distorting an existing music based on **I2**, we designed 4 audio loops with our *techno* and *solo-strings* models using timber transfer. A simple track was composed using these samples and we agreed on the order of display so that the dancer

[12]https://acids-ircam.github.io/rave_models_download

could gradually modify the music while the musician expanded the exploration ranges. For part 2 based on **I1**, as the mappings were already set for each model, we agreed on the global timeline. While practicing co-improvisation, we found interesting latent areas and decided to save some presets. The part 3 based on **I3** was more challenging as the dancer had to perform specific movement in order to activate the pre-learned gesture-sound mappings. There were fewer possibilities of connection between the dancer and the musician, who was only selecting specific mappings, replacing the classifier from the original approach [22] that failed to recognize the right gesture in real-time with low latency. Hence, the dancer could only improvise in a single sequence per sensor at a time.

6 INTERVIEWS AND DISCUSSION

Finally, we report feedback from the dancer's interviews and discuss different important points we found during this research process. We applied *Reflective Thematic Analysis* [7] on the last interviews conducted after the performance.

6.1 The need to understand, but to what extent?

When starting the project, the dancer initially *"thought that everything was already prepared and that [we] just had to click on buttons"* so she *"could create music while moving"*. After the first working sessions, she understood that we had to co-design the interactive system together but *"did not realize right away the work that needed to be done behind, especially from a technical perspective"*. When testing the initial prototype, she directly felt the system complexity: *"at first, I felt like it was something completely beyond me, so, for me, it was inevitably complicated"*. Without scientific and technical knowledge, for her the system was difficult to understand and control. Hence, she struggled to find her place: *"I had a lot of difficulty understanding what was expected of me, what I needed to do"*. For designing the system, we had to reconcile both artistic and technical skills. Hence, the collaboration was difficult at first as *"we were not speaking the same language"*. This led to disparities in roles and responsibilities in the creation process: *"You [researcher/musician] had to think about a lot of things on your side, whereas I [dancer] honestly had much less to do"*.

To alleviate this, the dancer needed to gain a sufficient level of understanding not only about the system but also regarding the roles of each performer in the live interaction: *"It was also crucial for me to understand, which took me some time to realize, that you were managing the exploration ranges and you were taking me to places"*. Still, the key point was to clarify the ambiguity surrounding the concept of *space*. We actually had to cope with three different spaces namely 1) the *gestures space* perceived by the dancer, 2) the *motion descriptor space* that captures the movement dynamics and finally 3) the *latent space* containing the synthesis parameters that was to be discovered from her own perspective through embodied exploration. The dancer embraced the system when she managed to *"visualize and imagine herself dancing in the latent space as if [she] were inside the computer"*. She even made a distinction between the three interaction methods. For **I1**, she was *"**inside** the machine"* like she was part of a musical instrument, for **I2** she was *"**along** the machine"* interacting with it, while for **I3** she was *"**with** the machine"* that became kind of a partner. Hence, the latent space was *"essential to understand"*: *"I wouldn't have been able to dance properly if I hadn't understood what a latent space was"*.

Interestingly, it appears it was not necessary, for the dancer, to understand all the technical aspects but rather to build one's own interpretation to enable her to be creative. Although she was *"just starting to understand what a model is"* after the 3-month collaboration, she *"had time to adapt, understand what was happening, and find it enjoyable, allowing [her] to dance with it easily in front of people"*.

6.2 A constraint fostering interdisciplinarity

The co-design of the system was a core constraint that strongly influenced the creation process: *"the way of building the performance was extremely different because when we create a dance show, there's a phase of discovery and experimentation, but it stops after a while. [...] here, it was exploration from the beginning to the end."* We constantly had to alternate steps of discovery and creation while refining technical and artistic choices. Hence, this constraint strengthened the link between dance and music by fostering interdisciplinarity. While the dancer usually *"only have to focus on the dance, here there was another crucial aspect, that of music and 'computing,' which turned out to be a constraint to consider equally"*. The system compelled us to collaborate and integrate our various expertise both in the design and performance phases. The decisions were to be made collectively *off* and *on* stage. To co-discover and co-create, the system constantly *"forced [us] to adapt"*: *"It's really interacting with the system that gives us ideas"*. Although this involved some difficulties as mentioned, the complementarity of both profiles allowed pushing the system limits. Without the scientific understanding of the system, the dancer *"didn't know what was possible and what wasn't, so, [she] allowed herself to try things that might have been impossible"*. As a dancer, she *"controls the dynamics of [her] body"* which was also an *"advantage"* in the embodied exploration process.

6.3 Trade off between control and expectations

A particularity of this project was the absence of strong *a priori* expectations regarding the aesthetics of movements and sounds. The shared mindset was to take inspiration directly from the system. The dancer's goal was for the audience to perceive that she produces the sound through her movements while the musician/researcher aimed to investigate the creative potential and limits of the latent-based interactive system through embodied exploration. Therefore, the focus was on the interaction design and not the aesthetics criteria. Hence, the lack of control of deep audio generative models was not a real problem, at least for the dancer: *"sometimes we didn't get the result we wanted, but that's part of the thing, so for me, it's not really a disadvantage"*. While making the performance, we did not try to find a solution to an engineering problem, which aligns with the anti-solutionism design approach highlighted by Fdili Alaoui in the *SKIN* project [19] and enabled us to avoid some tensions and frustrations from an artistic viewpoint.

However, this does not mean there was no control. There were actually a gradual complexity in the interaction methods as we tried to impose artistic constraints regarding the sound and movement in **I2** and **I3**. Interestingly, the more we added input control, the

more we lost the link between motion and sound as it became less intuitive for the dancer while also highlighting current technical limitations. The dancer preferred **I1** as *"there was something reassuring because it worked"*. The movement-sound relationship was more obvious while the co-improvisation setup with the musician modifying the areas to explore enabled to preserve the amount of *"surprise for everyone"* both performers and audience.

Although the system was technically deterministic (RAVE implementation is deterministic and no use of random function in the MAX programming), some unpredictability arose in its use. We can distinguish three levels of *randomness* preventing reproducibility. First, technically, it is due to small precision variations in the live capture of motion signals. As the latent space can be very sensitive to small variations , it also impacts the latent trajectory used to generate the sound. Second, it is difficult for human to be self-consistent in reproducing exactly the same gesture. Third, the interaction was designed to perform co-improvisation. Both the dancer and musician were interacting with system. Hence, reproducing exactly the same sound outcome would require to have exactly the same inputs combination from both performers which is very unlikely to happen. Improvising *with* and *through* the system was considered as the strong point of the performance as *"none of the people in the room, including us, will really know what's going to happen"*, *"we're not in a monotonous thing where we know exactly what we have to do, we're not in control at all, and anything can happen. It's action"*.

6.4 Questioning and influencing dance practice

Interacting with the system has strengthened the link between dance and music and challenged existing practices. For the dancer, *"the technology was the center of the piece, it allowed [her] to embrace the constraint, pushing [her] to explore further"*. It encouraged her to step out of her comfort zone and to actively *"be in the search for sounds"*. Freed from the choreography considered *"too independent from the music to be used in this context"*, she experienced more freedom: *"I can do absolutely whatever I want, the only thing that matters to me is you and what will come out of the speakers"*.

By physically augmenting the dancer's body (*"instead of seeing my movement, I can listen to it, I couldn't do it without it"*), the system interaction seems to favor learning sensorimotor skills [26]. She learned to *"be and dance in the present moment"* and to *"adapt [her] movement to the music"*. Hence, it strongly influenced the movement-sound relationship and helped her improving her kinaesthetic awareness [28]: *"being truly in the present with what's happening in my ear and what's happening in my body"*.

At the end, she truly became a *hybrid* performer mastering both dance movement and sound, such that people from the audience *"were surprised to find out that [she] had never practice music before"*. Yet, although it had questioned her perception of the motion-sound relationship while improved her improvisation skills, without the system she *"returns to [her] usual patterns because putting on music and dancing to it doesn't constrain [her] at all"*.

6.5 Creative potential and ethical impacts

This project also had an impact beyond the stage. According to the dancer, the live co-creation of sounds by both performers *"strengthened the connection with the audience in the sense that we're not just doing dance or just music. We question many things and take people into a universe, which is not easy to do"*. The audience got more involved questioning the motion-sound relationship and the role of each performer, *"coming out saying 'at moments, I wasn't sure if she was controlling the music or if the music was controlling her'"*. Despite the complexity of the system the gestural interaction was perceived by the dancer as *"intuitive"* and *"accessible"*. This encourages us to further investigate embodied exploration for designing intuitive interaction with latent-based generative models. Although controlling these models remains challenging as the latent representation is very abstract and high-dimensional, embodied interaction seems to be stimulating for creative endeavors.

On a technical side, we believe that temporal regularization of the RAVE latent space could alleviate some issues and improve the sampling procedure through movements. These limitations will be investigated in future works. Still, some broader technical limitations need to be addressed in order to fully deploy generative AI in music/dance context. Training these models is very time-consuming and requires large dataset. This slows the artistic workflow, and might be cumbersome to reach a pre-conceived aesthetics intent. Nevertheless, such an approach can fit a more open exploration process as described in this paper. While some frustrations appeared about the current complexity of building audio models, the dancer found overall the process sufficiently stimulating artistically and wish to pursue this endeavor. For now, it seems that we are far from reaching the creative limits of the system.

7 CONCLUSION

In this project, we studied the use of deep generative models in movement-based sound/music interaction system. We carried on this research through a performance-led research design approach, by collaborating with a professional dancer towards a live performance. A central issue with these models concerned the mapping of movement descriptors to parameters of the latent space derived from the *machine learning*. By considering an exploration approach of this abstract parametric space through movement, we called *embodied exploration*, we designed three interaction methods, documented with open-source patches. They are based on different approaches to parameterize the latent space and represent a contribution to the MOCO community.

The documentation of the project, including several interviews of the dancer allowed us to outline the major strengths and drawbacks from our approach. These discussions highlight the potential of deep models for the interactions between sound and movement, while several technical limitations remains to be further investigated. Finally, we hope that this research will open new questions about the integration and development of AI in dance-music practices.

ACKNOWLEDGMENTS

This work has been supported by the Paris Ile-de-France Région in the framework of DIM AI4IDF, and by *Nuit Blanche-Ville de Paris*. We extend our heartfelt thanks to Marie Bruand without which this study would not have been possible. We are also deeply grateful to our friends and colleagues from the STMS-IRCAM lab, particularly Victor Paredes, Antoine Caillon and Victor Bigand.

REFERENCES

[1] Sarah Fdili Alaoui and Jean-Marc Matos. 2021. RCO: Investigating social and technological constraints through interactive dance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] Steve Benford, Chris Greenhalgh, Andy Crabtree, Martin Flintham, Brendan Walker, Joe Marshall, Boriana Koleva, Stefan Rennick Egglestone, Gabriella Giannachi, Matt Adams, et al. 2013. Performance-led research in the wild. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 3 (2013), 1–22.

[3] Daniel Bisig. 2022. Generative dance-a taxonomy and survey. In *Proceedings of the 8th International Conference on Movement and Computing*. 1–10.

[4] Daniel Bisig and Pablo Palacio. 2016. Neural narratives: Dance with virtual body extensions. In *Proceedings of the 3rd International Symposium on Movement and Computing*. 1–8.

[5] Maaike Bleeker. 2016. *Transmission in motion: The technologizing of dance*. Taylor & Francis.

[6] Susanne Bødker. 2006. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*. 1–8.

[7] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[8] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011* (2021).

[9] Antonio Camurri, Shuji Hashimoto, Matteo Ricchetti, Andrea Ricci, Kenji Suzuki, Riccardo Trocca, and Gualtiero Volpe. 2000. Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. *Computer Music Journal* 24, 1 (2000), 57–69.

[10] Baptiste Caramiaux and Marco Donnarumma. 2021. Artificial intelligence in music and performance: a subjective art-research inquiry. *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity* (2021), 75–95.

[11] Baptiste Caramiaux and Sarah Fdili Alaoui. 2022. " Explorers of Unknown Planets" Practices and Politics of Artificial Intelligence in Visual Arts. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.

[12] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438* (2022).

[13] Ninon Devis, Nils Demerlé, Sarah Nabi, David Genova, and Philippe Esling. 2023. Continuous descriptor-based control for deep audio synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[14] Constance Douwes, Giovanni Bindi, Antoine Caillon, Philippe Esling, and Jean-Pierre Briot. 2023. Is Quality Enough*f* Integrating Energy Consumption in a Large-Scale Evaluation of Neural Audio Synthesis Models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[15] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2018. GANSynth: Adversarial Neural Audio Synthesis. In *International Conference on Learning Representations*.

[16] Jesse Engel, Chenjie Gu, Adam Roberts, et al. 2019. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*.

[17] Philippe Esling and Ninon Devis. 2020. Creativity in the era of artificial intelligence. *arXiv preprint arXiv:2008.05959* (2020).

[18] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[19] Sarah Fdili Alaoui. 2019. Making an interactive dance piece: Tensions in integrating technology in art. In *Proceedings of the 2019 on designing interactive systems conference*. 1195–1208.

[20] Rebecca Fiebrink, Perry R Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 147–156.

[21] Rebecca Fiebrink and Laetitia Sonami. 2020. Reflections on eight years of instrument creation with machine learning. (2020).

[22] Jules Françoise and Frederic Bevilacqua. 2018. Motion-sound mapping through interaction: An approach to user-centered design of auditory feedback using machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–30.

[23] Jules Françoise, Sarah Fdili Alaoui, and Yves Candau. 2022. CO/DA: Live-Coding Movement-Sound Interactions for Dance Improvisation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[24] Jules Françoise, Norbert Schnell, and Frédéric Bevilacqua. 2013. A multimodal probabilistic model for gesture–based control of sound synthesis. In *Proceedings of the 21st ACM international conference on Multimedia*. 705–708.

[25] Karmen Franinovic and Stefania Serafin. 2013. *Sonic interaction design*. Mit Press.

[26] Andrea Giomi. 2020. Somatic sonification in dance performances. From the Artistic to the Perceptual and Back. In *Proceedings of the 7th International Conference on Movement and Computing*. 1–8.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[28] Stacy Hsueh, Sarah Fdili Alaoui, and Wendy E Mackay. 2019. Understanding kinaesthetic creativity in dance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Andy Hunt, Marcelo M Wanderley, and Ross Kirk. 2000. Towards a model for instrumental mapping in expert musical interaction. In *ICMC*.

[30] Théo Jourdan and Baptiste Caramiaux. 2023. Machine Learning for Musical Expression: A Systematic Literature Review. (2023).

[31] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. arXiv:http://arxiv.org/abs/1312.6114v10 [stat.ML]

[32] Micheline Lesaffre, Pieter-Jan Maes, and Marc Leman. 2017. *The Routledge companion to embodied music interaction*. Taylor & Francis.

[33] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th ICML*, Vol. 97. PMLR, 4114–4124. https://proceedings.mlr.press/v97/locatello19a.html

[34] Fabio Morreale. 2021. Where Does the Buck Stop? Ethical and Political Issues with AI in Music Creation. *Transactions of the International Society for Music Information Retrieval* 4, 1 (2021), 105–114.

[35] Tim Murray-Browne and Panagiotis Tigas. 2021. Latent mappings: Generating open-ended expressive mappings using variational autoencoders. In *NIME 2021*. PubPub.

[36] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. (2016). http://arxiv.org/abs/1609.03499 cite arxiv:1609.03499.

[37] Ashis Pati and Alexander Lerch. 2021. Is disentanglement enough? On latent representations for controllable music generation. *arXiv preprint arXiv:2108.01450* (2021).

[38] Joseph Butch Rovan, Marcelo M Wanderley, Shlomo Dubnov, and Philippe Depalle. 1997. Instrumental gestural mapping strategies as expressivity determinants in computer music performance. In *Kansei, The Technology of Emotion. Proceedings of the AIMI International Workshop*. Citeseer, 68–73.

[39] Jan C Schacher. 2010. Motion To Gesture To Sound: Mapping For Interactive Dance.. In *NIME*, Vol. 2010. 250–254.

[40] Norbert Schnell, Axel Röbel, Diemo Schwarz, Geoffroy Peeters, Riccardo Borghesi, et al. 2009. MuBu and friends–assembling tools for content based real-time interactive audio processing in Max/MSP. In *ICMC*.

[41] Diemo Schwarz. 2007. Corpus-based concatenative synthesis. *IEEE signal processing magazine* 24, 2 (2007), 92–104.

[42] Hugo Scurto and Ludmila Postel. 2023. Soundwalking deep latent spaces. In *Proceedings of the 23rd International Conference on New Interfaces for Musical Expression (NIME'23)*.

[43] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[44] Jonathan Sterne and Elena Razlogova. 2021. Tuning sound for infrastructures: artificial intelligence, automation, and the cultural politics of audio mastering. *Cultural Studies* 35, 4-5 (2021), 750–770.

[45] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2021. Ai-terity 2.0: An autonomous nime featuring ganspacesynth deep learning model. (2021).

[46] Kıvanç Tatar, Kelsey Cotton, and Daniel Bisig. 2023. Sound Design Strategies for Latent Audio Space Explorations Using Deep Learning Architectures. *arXiv preprint arXiv:2305.15571* (2023).

[47] Gabriel Vigliensoni, Rebecca Fiebrink, et al. 2023. Steering latent audio models through interactive machine learning. (2023).

[48] Federico Ghelli Visi and Atau Tanaka. 2021. Interactive machine learning of musical gesture. *Handbook of artificial intelligence for music: Foundations, advanced approaches, and developments for creativity* (2021), 771–798.

[49] Marcelo M Wanderley and Philippe Depalle. 2004. Gestural control of sound synthesis. *Proc. IEEE* 92, 4 (2004), 632–644.

[50] Qiushi Zhou, Cheng Cheng Chua, Jarrod Knibbe, Jorge Goncalves, and Eduardo Velloso. 2021. Dance and choreography in HCI: a two-decade retrospective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.