Performing with the inclusive machine: Roadmap for the design of AI collaborative musical instrument

Pablo Mollenhauer*
pablo.mollenhauer@posteo.de
Independent artist and researcher
Valparaíso, Chile

Alejandra Pérez Núñez*
alejandra.perez@uv.cl
Universidad de Valparaíso
Valparaíso, Chile

Abstract

Computer musical instruments that use traditional programming paradigms have leaned towards predictability and direct causality based on deterministic mappings, rule-based synthesis, and explicit parameter control. Many studies have implemented instruments using machine learning for mapping and sound synthesis, but yet, few studies have explored the implications of these technologies on agency, causality, and power relations in the design process and performance practice. This paper explores the political-agential balances and aesthetics that emerge when machine learning technologies are integrated into the design process of computer music instruments. The hypothesis is that control and agency is distributed by machine learning's algorithms, in which designer and performer, have to navigate through the feedback system made of gestures, latent space and spatialized sound. The methodology comprises several chained machine learning stages: a single pose tracking system driven by pre-trained models, a supervised neural network for mapping parameters of the latent space of models for sound synthesis. The design stage is informed by first-person accounts of experience using micro-phenomenology.

Keywords

Agency, Relational Interfaces, Machine Learning, micro-phenomenological interview, research-through-design

ACM Reference Format:
Pablo Mollenhauer and Alejandra Pérez Núñez. 2026. Performing with the inclusive machine: Roadmap for the design of AI collaborative musical instrument. In *Proceedings of International Conference on New Interfaces for Musical Expression (NIME '26).* ACM, New York, NY, USA, 7 pages.

1 Introduction

Academic research in the field of computer music instruments and human computer interaction increasingly focuses on agency. Tom Mudd [24] problematized agency in musical tools and interactions, examining the different perspectives that have dominated computer musical instruments. He proposes the perspective of the "entangled whole" taken from Karen Barad's notion of intra-action, to think about the complex web of material and social concerns surrounding both the design and the use of creative technologies. With the advent of intelligent instruments and Machine Learning , cultural and political implications capture the attention of the community [17] yet, the primary concerns are environmental, under representation, exclusion and colonialism, which makes the practice with ML essentially political (Ibid.).

This article describes practice-based research on musical instruments using machine learning as we expect to gain theoretical insight emerging from the process of design and play with the system. We can assert that the technical aspects of the implementation are not the focus; however, the system's development is the playground from which a theoretical and politically informed discourse is built about how the gestural computer musical instrument can be understood as a relational agency that constitutes the foundation for thinking of technical assemblages as interfaces for human-machine collaboration.

2 Background

2.1 Gesture, Material oriented musical expression

Agency has been a particular subject of interest in the field of computer music instruments. The emergence of machine learning and artificial intelligence has widened the discussion from material-oriented interactions that involve control, communication, and processes [24] towards questions of "intelligence," assemblages, and ecology [1]. Despite the significant importance these studies have for our work, we would like to take a critical approach regarding the question of agency in computer musical instruments that use machine learning. The approach is an attempt to open the discussion about agency in MCI. which has somehow been devoid of a politics of power, particularly in the increasing use of machine learning in MCI. In defining musical expression, it may be relevant to consider synthesizing the functional and communicative roles of performer gesture, the subjective 'felt sense' of kinesthetic awareness , and the material-oriented perspective of instrumental interaction.

We should specifically address how expression is differentiated between sound-producing (functional) and ancillary (non-sound-producing) movements. Furthermore, our definition should explore how expression is viewed not as an inherent quality of a tool, but as an emergent property of the interaction between a human and a system, often requiring 'control intimacy' and an 'investment of play' to achieve nuance [19, p7,16].

Finally, it may be necessary to contrast the communication-oriented view, where the instrument is a transparent conduit for ideas, with the material-oriented view, where the instrument is a collaborator that 'kicks back' to shape the expressive outcome." Experience in this context, is an emergent property of the "ergodynamic" interaction, where the performer's skills and the instrument's potential co-evolve over time (Ibid., p20,30).

A core component of expression is kinaesthesia, the internal sensation of bodily movement and position that offers a "felt sense" of dynamics such as smoothness, intensity, and swiftness (Ibid. p. 4, p. 5, p. 26-27). This internal guidance allows musicians to integrate formalized training with intuitive interpretation,

*Both authors contributed equally to this research.

NIME '26, June 23–26, 2026, London, UK

creating a mental "mind map" of associations between gesture and sound. Mastery of an instrument occurs when the tool is no longer perceived as a distinct entity but is felt from within, integrated into the somatic knowledge of the musician (Ibid., p.5, p.89).

The practices that use AI have brought new possibilities not only for mapping and controlling computer musical instruments but also for tracking body gestures without the need to build devices or wear sensors to obtain the gestural data. For instance, in this group is the research team from the University of Arts London [22], "that has implemented interactive sonification of human movement through unsupervised machine learning, creating maps between latent spaces that link a pose estimator to a neural audio generator to enable sonification of human bodies." This has brought a new set of questions around the materiality of the instrument, as the gesture has become increasingly free of cables and devices that restrict the movement of the performer. This feature opens up new insights about the agency of the technical system and its role in the musical execution of the performer in the system.

The notion of assemblage appears useful as we recall it in the context of 21st-century musical practice. We should consider the theoretical frameworks proposed by Gilles Deleuze and Felix Guattari [p.71-73, 503-505][12] and their re-interpretation by Paul Théberge and Deniz Peters [in 5, p.60], specifically, examining how an instrument is transformed from a 'singular technical object' into a 'multiplicity of heterogeneous terms' that establishes liaisons between objects, practices, and social discourses. Such a design addresses how these systems can acquire their own self-agential agency, as well as distributed and interpersonal instrumentality within an 'installatory assemblage.'

2.2 Interactivity and ideology

To think about the relationship between agency and power politics in MCI, it is useful to review theoretical accounts of interactive art, an art form that has historically intersected with MCI yet possesses a wide critical literature on the politics of interaction.

Roberto Simanovsky [p.123][34] has stated that interactive systems create instances of dialogue; rather than presenting messages to be deciphered, they create events. This type of dialogue takes three forms: a physical dialogue with oneself, between interactors, and with the machine. The first two dialogue forms constitute a reflective surface for the interactors as individuals and as part of a group by producing space-times of personal and inter-human experiences (ibid.). Furthermore, by using deep learning tools, the dialogue is expanded to different scales and dimensions that no longer focus uniquely on the inter-human experience but also on the experiences in relation to the latent spaces and the open responses of the neural networks.

The interactive space of dialogue described by Roberto Simanovski 2011 coincides with the relational space described by Nicolas Bourriaud 2002, who defines it as "spaces where we can elaborate alternative forms of sociability, critical models, and moments of constructed conviviality" [4, p.44]. This conceptual similarity transforms the interactive space from a space to be seated in, to a space to stand or walk through, into a period of time in which one lives with other human beings (Ibid.:15) Even though Simanovski sees interactive art as a relational experience, he problematizes the political and social implications, labeling the participants of the interactive event as "collaborationists of the society of spectacle". Considering his critique, interactivity, and particularly sound interactive experiences should be a strategy to activate political and ethical physical dialogs by intervening in the daily sensorial framework of the community in such a way as to make explicit that "interactivity is cultural industry in camouflage" [34, p157].

2.3 Interface and ideology.

From the Marxist theories of the cinematographic apparatus, a parallel can be traced between the cinematic theory of suture and ideology in interactive new media. Jean-Pierre Oudar's theory of the suture deals with the psychological mechanisms of representation in cinema that create transparency. Oudar suggested that the cinematic succession of images deconstructs the representational system because it raises the question, who is watching this? [8, p.186]. The system of suture works by negating the question of "who is watching" with another opposite image: the reverse shot model [15]. Thus, the film discourse presents itself as a product without a producer, a discourse without an origin. Jean-Pierre Oudart refers to the spectator who occupies the missing field as the "Absent One" (ibid.:188). The suture theory explains how identification is attained, how immediacy is achieved, and finally how ideology is concealed in representational regimes. Similarly, we can think that interactive systems and human computer interfaces in new media art conceal ideology. However, not through cinematic grammar, but through interactive structural rules.

An interactive grammar is constituted by an interface. An interface is a device or program that enables a user to communicate with the interactive system or software. Simanovski 2011 assumes that an interface defines a way of sensing and a way of acting in an interactive system. Also, authors like David Rokeby and Lev Manovich remark that the interface is the content [31] and it also acts as a representational system [20]. Interactive media art can conceal ideology through representational mechanisms constituted by the interface.

An interface conceals ideology in the design's strategy of the interactive system that makes the user react in a pre-determined way without being aware of it. This is produced by a process of "interpellation" (Louis Althusser)[10], where the interactor is asked to mistake the structure of somebody else's mind, an absent one, for his own [20]. The interface provides its own model of the world, its own logical system, or ideology; subsequent cultural messages or whole languages created using this code will be limited by this model, system, or ideology (ibid.:76). This discussion raises the question of whether HCI and MCI avoid this burden, or whether, as interactive art, they conceal their ideology.

Supervised and un-supervised Neural networks and the control of the models' latent space have the potential to present designs in which the interaction rules are not written in stone. Instead, the pre-determined rules are rather loosely defined instructions for a system that operates ambiguously; despite this, it is more prepared for the unexpected than the systems in which the interface is fixed with galvanized mappings.

Despite the impossibility of total control by the designer of the interactive system, the creation of an interactive grammar using a chain of artificially intelligent pre-trained models also involves ethical considerations for both the creator and the interactors. It is the responsibility of the creator because they create a kind of belief system that permeates the response of the models, carrying and reinforcing our assumptions about the way things are, and therefore shaping and modifying the user's subjective point of

view [31]. It is the responsibility of the audience because the designer is also passing the responsibility onto the interactor to represent the world and the human condition within it [39]. It is necessary to explore the ways in which the interactive system carries these beliefs as hidden content and to consider the ethical problems that arise as a representational system.

2.4 The emancipation of the interactor

In an interactive work, in most cases, particularly with gestural systems, the performers are kept in a state of ignorance about the process of producing the interactive rules and the logical structure they conceal, which is somehow a similar process of subjectivization for the audience. For Jacques Rancière, emancipation is the confirmation of equality [11, p.90]. He remarks that the essence of equality lies in the acts of subjectivization that undo the natural order of the sensible. The order of the sensible is governed by "the distribution of the sensible", which refers to forms of inclusion and forms of exclusion. In the political domain, subjectivization disturbs the organization of bodies based on a communal distribution of the sensible [28, p.89]. The interactive structural rules define what can be done, the modes of being, and the ways of communication within the system. Consequently, it defines a "distribution of the sensible", where emancipation means the undoing of the dominant "distribution of the sensible".

From Rancière's definition of "emancipation," it can be said that a modification of the interactive rules is produced when the interactors do something that the creator of the rules does not know herself. In this context, the use of a chain of pre-trained artificial intelligence models re-balances the equality between the creator and the interactors in the interactive system, which can serve, to some extent, as confirmation of the equality of each member of the community within the music listening situation.

An approach for emancipation can be the capacity of interactors to modify the elements of the interface or the rules that define the systems in real-time. Hence, there is a blurring between the roles of the user and the designer. Manovich [20] defines this kind of interactivity as open. Considering deep learning systems, Manovich's notion of openness concerning the rules suggests potential systems in which the neural networks could be trained just-in-time, consequently changing the behaviours of the system accordingly. This process can be thought of as a way to divert one of the problems that Manovich sees in this openness in traditional paradigms of computer-human interfaces, which is a subset of the variability principle because it is based on some structure or prototype that remains unchanged. This kind of interactivity conceals an ideological problem because it pretends to give unique choices rather than being pre-programmed and shared with others. Hence, it apparently allows the interactor to perform acts of subjectivization within the system; however, these acts are unique in the system's measure, and this creates a simulation of subjectivization.

2.5 Relational space

In order to understand more deeply how and when such users' acts of subjectivization are produced, it is necessary to explore how users utilize the interactive system. Manovich (Ibid.) remarks that while it is relatively easy to specify the different interactive structures used in new media objects, it is much more difficult to deal theoretically with users' experiences of these structures. This remains one of the most difficult theoretical questions raised by new media.

Spatial studies provide us with significant concepts that enable us to explore what users make and do with the structure of social space, which can be applied to interactive systems. Michael de Certeau [9] argues that a system is based on rules that ensure its production, repetition, and verification. He remarks that these rules render an order that is constructed by others and that redistributes time and space (ibid.:43). This definition corresponds to the definition of an interactive system. However, there is always a certain margin that enables the user to manoeuver within the spatial order of unequal forces (ibid.:43). As in productive social systems, the users of interactive systems are not passive or constantly guided by established rules or pre-determined choices; they constantly adapt and use the system to their own ends as a collective. De Certeau's model assumes, paralleling Rancière's logic of emancipation, that users create something with and within the system that its designer did not expect or envisage. In other words, they perform acts of appropriation by transforming the system's rules in order to adapt them to their own interests.

Rebecca Fiebrink 2016 argues that using supervised neural networks in the process of instrument design allows access to surprise and discovery, which fundamentally changes the relationship between the computer and the instrument that is built (Ibid.). We argue that the use of a chain of supervised and non-supervised machine learning models can provide instances for the unexpected during the instrument's design, and also in the performative event, in which the performer as being part of the instrument, produces relationships between gesture-sound that were not intended by the designer. As Fiebrink (ibid.) notes regarding the supervised neural network mapping system Wekinator, "it allows them to move away from a paradigm of control over a computer into one where the computer is a collaborator." This shift from an apparatus that allows us to mappping complex data towards the apparatus as a "collaborator", assigns agency to the instrument, or at least to a part of the design, which has the potential to radically change the politics and power of the interactive systems.

2.6 Relational Sonic Agency

In its basic form, albeit not its most, the live theatrical contract in the context of film is an interactive system. This form of presentation can also be considered as a sonic and physical responsive system that puts into action a set of structural rules that define the form in which the interaction takes place. In other words, the aesthetic live experience can be thought as devices that define the experience by enabling or disabling certain behaviours or actions that define specific roles in specific places. In fact, most of the time these "interfaces" of art experience intend to be transparent, to establish implicit environment-rules outside of the representational system.

In the context of the proposed work, a sonic gestural event constitutes a relational experience on the basis of which the performer explores a social configuration that is structured through mediated feedback, in which she affects the sound, which is more or less loosely mapped by the neural networks to control a less direct manipulation of the latent space of the model for sound synthesis, ultimately affecting the performer through sound spatialization and sound structure. The collaborative process of deciphering the interactive mechanism somehow demystifies the system, exposing the mechanisms of the spatial intersections that are represented by the interface. Thus, the interaction not

only articulates sonic and physical forms of dialogue, but also addresses the other less immediate dimensions of the system.

Interactive systems function as encounters with relational agencies when they move beyond being neutral tools to become dynamic ecosystems where performers, audiences, and technology are mutually entangled [19]. In *Future Perfect*(2018), using a high order ambisonic sound field, a VR film being projected around users and their personal cell phone become a speaker in the work, "the work itself becomes this dynamic ecosystem."[Paine G., in 19, p136]

2.7 Intra-action and the Inclusive Instrument

The concept of intra-action posits that agency is not an inherent property located in individuals but is an emergent property of mutual entanglement [Mudd, T. in 19, p124]. In this paradigm, the interactive environment is viewed as an inclusive instrument that the ensemble plays collectively, lacking complete individual control.

Works like Garth Paine's *Future Perfect* engage the audience as an active part of the ecosystem, utilizing networked smartphones to turn spectators into sound-producing agents. This removes the binary distinction between performer and audience. The techno-somatic dimension describes the porous, embodied relationship between a player and an instrument that incorporates feedback and listening into a cognitive map. The design approach proposed by Garth Paine's 2015 named techno-somatic, refers to "the 'feel' of an instrument, formed through both somatosensory feedback and listening, representing the cognitive map a performer develops in order to play an instrument, the technique, and how the instrument responds under different circumstances' [26, p84].

Paine's model views performance as a multidimensional embodied space comprised of enmeshed sets of relationships between the performer's exertion (breath, gesture) and the resultant sound.

The quality of these encounters is shaped by the depth of engagement required to master the system. For performers like Laetitia Sonami, building a custom instrument is a method of creative renewal, allowing her to "be different" and explore "dreams of chaos" rather than just predictable control [Sonami in 19, p43].

To attain coordination between micro-gestural movements and the instrument's capacity to evoke affective qualities, such as a tight somatic feel, often require low-latency systems and appropriate control metaphors [40].

2.8 Machine Learning, Misuse and Arenas

The culture emerging from the community using Machine Learning for the creation of instruments for musical expression has been recognised as political [17] in its inception, as it rises ethical issues related to energy consumption, gender and diversity, under representation of minorities colonial power relations, amongst others. Hence, a focus on plurality and heterogeneity will characterize an attempt to counterbalance these dispositions. Following this thread of thought and considering the black-boxed condition of the interaction with deep learning models, practitioners have considered surpassing standard design in favour of divergent views centred on hacking, probing, tinkering [35], misuse and unexpected uses of the technology. The implementation of such a system should take into account this considerations and as Sally Norman et al. [25] formulate the need of "maximising exposure to possible failure; shifting from interfaces to interfacing, to create arenas for action rather than tools for purposes" (Ibid.).

3 System Overview

The system [1] is built using existing technology that is at hand. It consists mainly of three stages: (1) a gesture tracking stage, (2) a neural network, (3) and a deep generative model for neural audio synthesis.

The first stage is software written in Python programming language that use MediaPipe Framework[18], an opensource BlazePose[3], GHUM[41] 3D estimation Convolutional Neural Network similar to MobileNetV2 [32], which estimates the full 3D body pose of an individual. The model returns 33 keypoints named "landmarks" describing the approximate location of body's joints. Each keypoint is a vector of three elements, a x, y and z coordinate. The z value is fitted into a 2D point projection. In the system, we use two sets of points: a version that filters the z coordinate of each of the 33 landmarks, which results in 66 values, and a reduced version that takes only 8 values in total. The Python script sends the tracking data through OSC (Open Sound Control protocol) messages using the python-osc [2] library.

The OSC message containing the data is received by the platform for audio synthesis and algorithmic composition, Supercollider[21]. In this platform, a supervised multi-layer perceptron neural network from FluCoMa (FluidCorpusManipulation Project) [36] is used[2]. This neural network is used to perform a regression to map the space of 66 and 8 dimensions received from MediaPipe Python script to the space of 77 dimensions that control the latent values of the 6 pre-trained models to generate audio, and 12 xy coordinates pairs to control the sound's place in a quadraphonic system. The input data can change accordingly to the inputs and outputs methods of the model. The neural network is trained by storing data points relating the data of a particular pose to a specific set of parameters of the synthesizer to produce a particular sound. Once the neural network is trained, it predicts based on the continuous input of numerical values.

The reason to experiment with the two configurations, the one with 66 and 8 input values, was to test the different artistic consequences when the Neural Network reduces or increases the dimensionality. This process can be scalable and easily redefined to any number of dimensions for the input and output. Moreover, it is possible to easily change the tracking machine learning model to one that tracks the features of the hands or the face.

Finally, the output from the neural network modulates the latent space of a neural audio synthesis model that uses a realtime Audio Variational autoEncoder (RAVE)[7]. The system uses the Rave-Supercollider implementation developed by Victor Shepardon [33]. This object allows us to access the latent values of pre-trained models to produce high-quality audio waveform synthesis. Ultimately, the sound produced is spatialized in a quadrophonic system, which is also controlled by the neural network. The system works in real-time on a relatively low end computer depending of the amount of pre-trained models used.

The system has many features that make it prone to producing relational agencies and serves as a probe and arena for contemplating the roles of the designer, performer, and audience in the process of making music. In the design stage, it allows for a participatory design process that surveys, in a similar fashion to Rebecca Fiebrink's Wekinator [14].

Like Wekinator, the system has the capacity to save data points and trained models. This allows for the iterative development of

[1]Repository: https://github.com/PaoloMolinari/Performing-with-the-inclusive-machine

[2]The system is based and adapted fron the the tutorial of Ted Moore *Controlling a Synth using a Neural Network* [23]

different variants of an instrument. Hence, it does not only serve to build an instrument from a pre-conceived idea but rather to refine the problem (Ibid.). This configuration converts the process of design into a process of discovering the unexpected rather than "testing" whether the mapping corresponds to the training examples (Ibid.). The use of machine learning also allows designers to engage their bodies in the design process. This is important because, as Fiebrink (Ibid.) argues, it makes the designer open to being influenced by the affordances of the tool, which brings aesthetic and philosophical considerations pertaining to the role of computers not only in music performance but also in the creation of instruments.

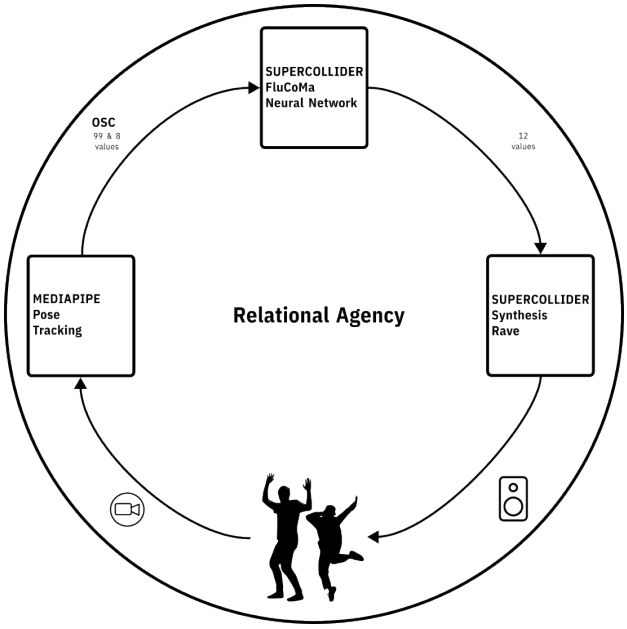

Figure 1: Diagram of the feedback system.

4 Methodology, Instrumental Encounters

The hypothesis is that by using artificial intelligence models for gesture recognition and body tracking alongside supervised neural networks for the mapping of control of the sound generator, it is possible to create interfaces that produce relational, more-than-human agencies as forms of collaborative and emancipated experience. The study argues that the sonically related motion event driven by AI constitutes a relational experience in which not only do the different dimensions dialogue but they also mediate the relation within agencies. From these agential intersections arises a set of questions that inquire about the relational agency that the system promotes: is this relational sonic agency a collaborative process? are the participants, musicians or not, attempting to decipher the instrument?. Relational agencies not only articulate sonic and physical forms of dialogue amongst the audience-performers, between performers and machines, but also between machines.

The research design considers microphenomenology to tackle the experience of a heterogeneous instrument, such as the described installatory assemblage, and the affordances that the instrument presents. Performative self-study may integrate the evaluative aspect of the described system. In this path, Courtney Reed et al.'s application of micro-phenomenology to NIME research [30] represents a significant methodological innovation [29]. New research on neural audio systems has suggested the use of microphenomenology [1] as a useful research tool that produces fine grained accounts and has been used to describe experiences in Human Computer Interaction research [16].

Micro-phenomenology, developed by Claire Petitmengin and colleagues, employs structured interviews [27, 38] to elicit detailed descriptions of lived experience at fine temporal scales. Reed et al. demonstrate how this method can reveal subtle aspects of performer experience—including pre-reflective bodily sensations, temporal dynamics of attention, and affective qualities that are often inaccessible through conventional user study methods. These approaches privilege subjective, embodied knowledge and recognize the researcher-performer's body-mind as a legitimate site of inquiry.

5 Discussion

This article describes the design of an instrument for musical expression informed by research-through-design methodology, similar to rehearsal-as-research and Software Development as Research (SoDaR)[6]. In the context of musical expression, the development of an interactive expressive sound instrument-system that uses artificial intelligence technology already available attempts to explore gestural interfaces to open questions about the causality of interactive systems from a critical perspective.

The agential elements present in the different stages can be identified as the performer's gestures, the neural network, and the sound itself. The agency of these actants manifests on different dimensions and forms. The gestures produce figures that change over time at different velocities. Their dimensions are spatial; subtle movements produce small changes on the system, while quick movements greatly alter its character. During the use of the instrument, the performer was always keen to reach the limits of the dataset, or surpass them, because it was on these instances that the performer found herself exploring the system, modulating her agency with something that was neither foreseen nor controlled. This exploration cannot be seen as an imbalance; but rather as a necessary act that evaluates the possibilities of the system. The system is an assemblage of multiple elements; its exploratory character cannot be uniquely attributed to the performer, the neural network, or the sound generator individually. As a result, the design of the musical instrument has been informed by micro-phenomenological insights, as the performer's first person account recalls,

> "I begin at the point where there is no sound and then start to move, first repeating the movements I used during the training (of the model) and quickly directing myself toward the limits of the situation. It seems as if I am moving in a space without light, first watching the lines on the screen and then disengaging from the screen, although I look back to check whether I am still within the frame. I am testing the differences: I execute quick movements and the system lags in responding, and when it does respond, it does so with delay—it feels like a sonic echo of my movements."

The neural network and the training's process were particularly important for promoting or diminishing the speculative system's nature. However, the training's process was in itself a task full of indeterminacy. The attempt was to associate a posture

with sounds in such a way that apparently opposite postures are associated with sounds that could be recognized as similar or opposite in some aspect. For instance, if the performer lifts the left arm and the sound A, the sound B corresponding to the lift of the right arm could be the opposite of the sound A. Yet, what is the opposite of a sound?. The choice was not technical, in which for instance, a sound with a lower spectral mean is the opposite of a sound with a high spectral mean. Rather, the choice was perceptual and aesthetic, in which the pose landmarks were associated with what were considered more expressive and musical sounds in the training process.

The behaviour of the neural network changes radically with different amount of data, the characteristics of the data selected and the relation between them, and the amount of training errors. While small datasets resulted in small training error, the resulting sound were more predictably and similar between each other. This can be changed by increasing the data points' amount. However, this brought about another set of problems by increasing the chances of creating inconsistent training data. For instance, when similar postures corresponded to very different sound figures, or when two completely different postures were associated with very similar sounds the neural network struggles to "learn". In these cases, the error of training did not get low enough for the neural network to infer unless the postures were way out of the limits of the system. As result, the system broke.

From these experiences, the training process was a balancing act, which consequently determined the agential balance of the dialogue between the different system's parts. It was important to systematize the training process to create a consistent dataset without over-fitting it. In this process, the neural network could be seen as a navigation chart, whose map does not represent accurately or completely. Rather, it renders an open field of action. This field has not clear limits, particularly when the performer did not have visual input of her or his position through the the camera's image and MediaPipe's tracking points.

> "I feel as if I am moving in a narrow space, almost like an oversized suit. It does not truly feel like a place; it resembles a bare tunnel under construction, in which I move with very little room, as if I cannot reach the limits of the space itself, only the limits of some kind of suit I am wearing."

While the neural network and the training's process determined the character and agential structure of the system, the sound reveals itself equally or even more important in the system's agential structure. The question that naturally turned up was, how does the sound's agency manifest? In our first person observations, it emerged that the sound agency manifests itself in its allure, in a sort of sonic locale to explore, that incited the performer to move in certain form and speed, to occupy specific distances, to adopt gestures, and also, to surpass the spatial and postural limits of the training stage. Despite the recognition of its importance, the sound produced by the instrument used in the system was limited. The timbral variety and sonic figures available were rather poor. The restricted range of sounds seemed to be more important than the positive results in executing control in the training process.

Whether the system controls the performer or the performer controls the system was relegated to the backdrop. This brings us to something discussed in the subsection 2.5. While the system could offer more or less control in its design, it is not only the system's design what guides the performer to act one way or another, but rather what the performer makes out of the musical flow that has the capacity to relate to her expression.

The micro-phenomenological interview, has been used to provide fine grained descriptions of the "lived experience"[37] of training and performing with the instrument. In conjunction with this primary results and apropos of the literature review, we present the following roadmap to study relational agencies at play when performing the aforementioned musical instruments. The focus should be on agency collisions, or the loss of agency, to specifically singularize moments of loss of control; the lived experience of the unexpected and surprise and the lived experience of sense making during design and use of the instrument.

6 Future Works

A complement for the qualitative method described above is the replacement of the sound source of the instrument: from the abstract sound and low-level mapping that require the latent variables of RAVE models, to the high-level control necessary to navigate through a corpus of sound recordings classified by a neural network. This stage opens up a new set of questions because the sounds can have significance beyond the signifier, shifting the assemblage and the instrumental affordances. This aspect opens the performance and the instrumental exploration to representational aspects of sound.

7 Conclusion

A composited algorithmic methodology based on machine learning is presented to inquire the balance in agency in the design of new instruments for musical expression using machine learning. In this design, control is seen as the distribution of control within a chain of feedback interfaces, two performers and a neural audio music instrument. The use of AI as a navigator is one of the metaphors used in the interaction although as navigation may be related to exploration, some implicit power relations like colonialism or the conquer of the unknown should be expected as to be haunting the installatory assemblage. Other possible epistemological frameworks may reframe relational agencies such as the proxy or representative. In this sense, the use of complex audio synthesis in addition to the described setting may be considered a proxy to AI models. This approach has been considered to increase the embodiment of the interaction with the ML models and to singularize relational agencies for inquiring the lived experience of sense making.

8 Ethical Statement

The following ethical considerations have been addressed in accordance with the NIME Principles and Code of Practice.

This work makes use of pretrained models, avoiding the need to train a new deep learning model from scratch and, therefore, significantly reducing the associated energy consumption and carbon footprint. Cloud computing resources were not used during the development of this project. SuperCollider, FluCoMa and Rave are open source tools that run efficiently on standard and low end consumer hardware, encouraging to reuse old hardware that can end up as electronic waste. Hence, the project attempts to encourage the use of technologies that allow a certain environmental sustainability despite the use of highly contaminating technologies.

All software used in this project is freely available, in alignment with FLOSS principles. Despite of the project seeks to support a democratic and inclusive approach to tool/instrument

making, in which people with limited income or access to resources can replicate the project, we acknowledge, however, that some technical knowledge of programming environments may limit accessibility for non-specialist users. We hope that communities can benefit from the technology developed and participate actively in the research and artistic practices proposed in the article.

No personally identifiable data was collected or used in this work. The audio corpus used for analysis and synthesis was composed of sounds created from openly licenced repositories, with appropriate attribution to the original authors.

It is acknowledged that the use of machine learning models trained on an existing audio corpus whose cultural scope may not be fully representative of global musical traditions is a limitation of current research.

In this submission, no AI tools were used to generate text, images, or code without explicit acknowledgement. However, large language models were used to assist with grammar checking and editing of this manuscript. No AI generated text was included without review and modification by the authors.

References

[1] Jack Armitage and Thor Magnusson. 2023. Agential Scores: Exploring Emergent, Self-Organising and Entangled Music Notation. In *Proceedings of the 8th International Conference on Technologies for Music Notation and Representation (Northeastern University, Boston, Massachusetts, USA, 2023)*.

[2] Timothé attwad. 2026. *python-osc*. Timothé attwad. https://github.com/attwad/python-osc

[3] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. arXiv:2006.10204 [cs.CV] https://arxiv.org/abs/2006.10204

[4] N. Bourriaud. 2002. *Relational Aesthetics*. Les Presses du réel. https://books.google.cl/books?id=GAxhQgAACAAJ

[5] Till Bovermann, Alberto de Campo, Hauke Egermann, Sarah-Indriyati Hardjowirogo, and Stefan Weinzierl. 2017. Musical instruments in the 21st century. *Berlin: Springer* 10 (2017), 978–981.

[6] Andrew R Brown. 2007. Software development as music education research. *International Journal of Education & the Arts* 8, 6 (2007), 1–14.

[7] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv:2111.05011 [cs.LG] https://arxiv.org/abs/2111.05011

[8] Daniel Dayan. 1974. The tutor-code of classical cinema. *Film Quarterly* 28, 1 (1974), 22–31.

[9] Michel de Certeau and Steven Rendall. 1984. *The Practice of Everyday Life*. Number v. 1 in Sociology, Anthropology, History, Literature. University of California Press. https://books.google.cl/books?id=WVn1XMEO168C

[10] Paul Du Gay, Jessica Evans, and Peter Redman. 2000. *Identity: a reader*. Sage.

[11] Jacques Ranci ere. 2009. *The Emancipated Spectator*. Verso.

[12] Guattari Felix and D Guattari. 1987. A thousand plateaus: Capitalism and schizophrenia. *Trans. by Massumi, B.), University of Minnesota, Minneapolis* (1987).

[13] Rebecca Fiebrink. 2016. Machine learning as meta-instrument: Human-machine partnerships shaping expressive instrumental creation. In *Musical Instruments in the 21st Century: Identities, Configurations, Practices*. Springer, 137–151.

[14] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, Vol. 3. 2–1.

[15] Stephen Heath. 1977. Dossier suture: Notes on suture. *Screen* 18, 4 (1977), 48–76.

[16] Katrin Heimann, Minke Nouwens, Suneetha Saggurthi, and Peter Dalsgaard. 2025. Micro-Phenomenology as a Method for Studying User Experience in Human-Computer Interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[17] Théo Jourdan and Baptiste Caramiaux. 2023. Culture and politics of machine learning in nime: A preliminary qualitative inquiry. In *New interfaces for musical expression (NIME)*.

[18] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC] https://arxiv.org/abs/1906.08172

[19] Mary Mainsbridge. 2023. *Body as Instrument: Performing with Gestural Systems in Live Electronic Music*. Bloomsbury Publishing Inc. http://dx.doi.org/10.5040/9781501368578

[20] L. Manovich. 2002. *The Language of New Media*. MIT Press. https://books.google.cl/books?id=7m1GhPKuN3cC

[21] James McCartney. 2026. *SuperCollider*. SuperCollider Community. https://supercollider.github.io/

[22] Joseph Meyer, Nick Bryan-Kinns, Sarah Fdili Alaoui, Mick Grierson, and Rebecca Fiebrink. 2025. Interactive Movement-to-Audio with Pre-Trained Neural Networks. In *Proceedings of the 2025 Conference on Creativity and Cognition*. 491–493.

[23] Ted Moore. 2026. Controlling a Synth using a Neural Network. https://learn.flucoma.org/learn/regression-neural-network/

[24] Tom Mudd. 2019. Material-Oriented Musical Interactions. Springer.

[25] Sally Jane Norman, Paul Stapleton, and John Bowers. 2025. NIME: A Mis-User's Manual. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 304–311.

[26] Garth Paine. 2015. Interaction as Material: The techno-somatic dimension. *Organised Sound* 20, 1 (2015), 82–89.

[27] Claire Petitmengin, Anne Remillieux, and Camila Valenzuela-Moguillansky. 2019. Discovering the structures of lived experience: Towards a micro-phenomenological analysis method. *Phenomenology and the Cognitive Sciences* 18, 4 (2019), 691–730.

[28] Jacques Rancière. 2013. *The politics of aesthetics*. A&C Black.

[29] Courtney N Reed and Andrew P McPherson. 2023. The body as sound: Unpacking vocal embodiment through auditory biofeedback. In *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–15.

[30] Courtney N Reed, Charlotte Nordmoen, Andrea Martelloni, Giacomo Lepri, Nicole Robson, Eevee Zayas-Garin, Kelsey Cotton, and Andrew McPherson. 2022. Exploring experiences with new musical instruments through micro-phenomenology. In *NIME 2022*. PubPub.

[31] David Rokeby. [n. d.]. The construction of experience: interface as content. In *Digital Illusion: Entertaining the Future with High Technology*.

[32] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381 [cs.CV] https://arxiv.org/abs/1801.04381

[33] Victor Shepardson. 2026. *Rave-SuperCollider*. Victor Shepardson. https://github.com/victor-shepardson/rave-supercollider/tree/master

[34] Roberto Simanowski. 2011. *Digital art and meaning: Reading kinetic poetry, text machines, mapping art, and interactive installations*. Vol. 35. U of Minnesota Press.

[35] Halla Steinunn Stefánsdóttir and Thor Magnusson. 2025. Of altered instrumental relations: a practice-led inquiry into agency through musical performance with neural audio synthesis and violin. *Frontiers in Computer Science* 7 (2025), 1578595.

[36] Pierre Alexandre Tremblay, Owen Green, Gerard Roma, Jacob Hart, James Bradbury, Ted Moore, and Alex Harker. 2024. The Fluid Corpus Manipulation Learn Platform. (2024). https://zenodo.org/doi/10.5281/zenodo.14930640

[37] Camila Valenzuela-Moguillansky, Ema Demšar, and Alexander Riegler. 2021. An introduction to the enactive scientific study of experience. *Constructivist Foundations* 16, 2 (2021), 133–140.

[38] Camila Valenzuela-Moguillansky and Alejandra Vásquez-Rosati. 2019. An analysis procedure for the micro-phenomenological interview. *Constructivist Foundations* 14, 2 (2019), 123–145.

[39] Graame Weinbren. 1995. In the ocean of streams of story. *Millenium Film Journal* 28 (1995), 15–30.

[40] David Wessel, Matthew Wright, and John Schott. 2002. Intimate Musical Control of Computers with a Variety of Controllers and Gesture Mapping Metaphors.. In *NIME*, Vol. 2. 1–3.

[41] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM GHUML: Generative 3D Human Shape and Articulated Pose Models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6183–6192. https://doi.org/10.1109/CVPR42600.2020.00622