Audience Experience of Visual Augmentations for Human-AI Co-Improvisations

Florent Berthaut
Univ. Lille, CNRS, Centrale Lille,
UMR 9189 CRIStAL, F-59000 Lille, France
florent.berthaut@univ-lille.fr

Ken Déguernel
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL, F-59000 Lille, France
ken.deguernel@cnrs.fr

Mathieu Giraud
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL, F-59000 Lille, France
mathieu.giraud@cnrs.fr

Barbara Dang
Independent artist
chechwoi@gmail.com

Matteo Plichon
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL, F-59000 Lille, France
plichon.matteo@gmail.com

Olivier Capra
ETHICS, Univ. Catho. Lille
olivier.capra@univ-catholille.fr

Abstract

Human-AI co-improvisation on digital musical instruments opens many expressive opportunities. However, it is not yet clear what the audience perceives and understands of such performances, for example regarding the respective contributions of musicians and artificial agents. In this paper, we investigate the use of mixed-reality visual augmentations to represent the interactions between a pianist and an artificial musical agent, with the aim of enriching the audience experience. We present the implementation of the system and a study involving 23 participants and augmentations with different levels of detail. Our results suggest that visual augmentations could increase the confidence of audience members in what they perceive of the respective contributions of human and artificial agents and the dialogue between them. They also provide insights on preferences for personalised levels of detail in the augmentations. We believe that this research will help inform the design of co-improvisation systems to benefit the audience.

1 INTRODUCTION

Co-improvisation between musicians and artificial musical agents (Tatar and Pasquier, 2019) creates new expressive possibilities, ranging from agents as semi-autonomous audio effects on a single instrument, to musician-artificial-agent duos, to large ensembles with musicians and artificial agents responding to each other.

While the degree of autonomy and contribution of these agents to the performance constitutes a parameter that musicians can explore creatively, it may be confusing for the audience watching such performances. Indeed, performing with an artificial musical agent constitutes an extreme example of a loss of *attributed agency* in Digital Musical Instruments, defined here as the degree of control over the sound attributed by the audience to a musician. More precisely, artificial musical agents might lead to a disruption of both the exclusivity criterion, *i.e.,* potential origin for a change in the sound (Berthaut et al., 2015), and the gesture-to-sound "association challenge" (Capra et al., 2020b).

To the audience, human-AI co-improvisation therefore raises questions such as: Does it disrupt objective and/or subjective understanding of the performances or trust in musicians? Can it benefit from adapting instruments/gestures or making use of spectator experience augmentation techniques (Capra et al., 2020a)? These questions echo recent research on explainable AI for the Arts (Bryan-Kinns, 2024) and transparency and trust in algorithmic interfaces (Kizilcec, 2016).

In this paper, we investigate how the collaboration between artificial agents and musicians is perceived by the audience, how it impacts their experience, and whether this experience can be enriched.

We deliberately focus on note-level attributed agency, *i.e.,* perceiving whether the artificial agent or the musician played a given note, as it contributes to the understanding of higher levels of human-AI interactions, *e.g.,* initiating ideas, responding, synchronising. In turn, this comprehension constitutes one important aspect of the overall audience experience.

To do so, we propose visual augmentations providing various levels of detail on the activity of the artificial agent, and we study how these levels influence the objective (actual) and subjective (perceived) understanding of the respective contributions, and which augmentation strategies are preferred by the audience.

2 RELATED WORK

This paper builds upon research on Human-AI co-creativity, co-improvisation, the audience experience and visualisations in mixed reality.

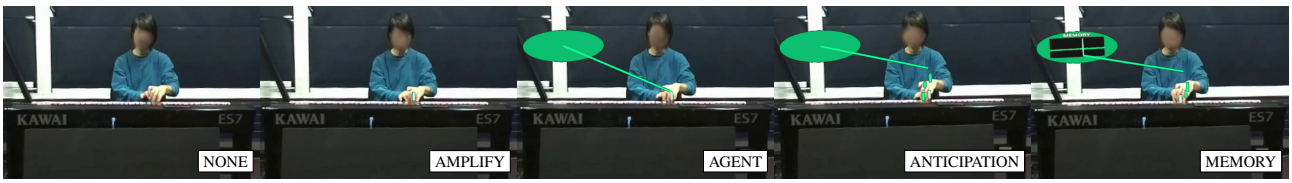

Figure 1: Five levels of detail in visual augmentations representing the artificial agent contribution, designed with the goal of improving the audience experience of human-AI co-improvisations.

2.1 Human-AI musical co-improvisation

Co-improvisation systems are designed for live musical interaction between a machine and one or more musicians, reacting to their performance and generating stylistically appropriate responses. As such, these systems represent a subset of *music generation systems* (Herremans et al., 2017) and of *music performance systems* (Jung, 2023). One of the first significant co-improvisation systems is George Lewis' *Voyager* (Lewis, 2000), developed in the 1980s, using rule-based strategies to generate, in a live setting, complex orchestral responses and accompaniment to a human solo improviser. A key characteristic of co-improvisation systems is that the interaction takes place through *musical agents* (Tatar and Pasquier, 2019), *i.e.,* "autonomous systems that initiate actions to respond to their environment in a timely fashion" (Wooldridge, 2009).

In this paper, we focus on Dicy2 (Nika et al., 2017b), a library based on prior work such as OMax (Assayag et al., 2006; Déguernel et al., 2018), Improtek (Nika et al., 2017a) and SoMax (Bonnasse-Gahot, 2014). Dicy2 uses a Factor-Oracle-based memory representation, combined with active-listening and anticipation mechanisms, allowing for real-time learning and generation, in free-form or idiomatic music. According to Tatar and Pasquier (2019)'s agent taxonomy, Dicy2 can be classified as a hybrid agent, *i.e.,* combining cognitive (knowledge representation through sequence modelling) and reactive components. It can be configured in either single-agent or multi-agent mode. It interacts through either audio or symbolic data, relying solely on music for communication through the real-world environment, and can learn from humans, be controlled by humans (through different parameters), and play with humans.

2.2 Augmenting the Audience Experience

Audience experience has been an important focus in NIME research notably in the past decade, as either a way to design or to evaluate instruments (Barbosa et al., 2012; Fyans et al., 2010). Recognising the disruption that Digital Musical Instruments may cause on perceived gesture-sound causality compared with acoustic instruments (Berthaut et al., 2015), researchers have proposed providing additional haptic or visual feedback for the audience. In particular, Capra et al. (2020b) have shown how dynamic visual augmentations that display the mechanisms (sensor values, mappings, sound processes) of instruments can improve the trust in musicians compared to no or verbal explanations. Moreover, they report on the importance of levels of detail in the information provided to the audience (Capra et al., 2021), which we draw upon in our proposed system and study. However, most of this research has focused on a single musician. Work on multiple musicians, *i.e.,* digital ensembles, has also been conducted with visualisations that highlight the contributions of each musician (Perrotin and d'Alessandro, 2014). To the best of our knowledge no research has yet focused on live visualisation of human-AI collaboration for the audience.

2.3 Visualisations of musical agents

Artificial musical agents are often visualised for interaction purposes, either on 2D graphical interfaces or as virtual elements displayed in the physical space (Lucas and Fasciani, 2023). In non-graphical musical-agent systems, additional visual cues have also been shown to increase the engagement of musicians when improvising with AI agents (McCormack et al., 2019). Closer to our contribution, Wang and Martin (2024) propose to visualise musical agents' activity through a virtual replica of a physical controller, and the system Andante uses projection of avatars of artificial agents on a Disklavier (Xiao et al., 2014).

Only a few visualisations have been explicitly designed for the audience, a notable exception being CAVI (Erdem et al., 2022a,b), which provides a visual representation of a musical agent for guitar improvisation. We believe that visualisations of musical agents may be beneficial for the audience and should be thoroughly studied.

3 VISUAL AUGMENTATIONS FOR THE AUDIENCE OF HUMAN-AI PIANO CO-IMPROVISATIONS

In this paper we focus on the case of an agent as a partner in a duo context, where a single AI musical agent co-improvises with a musician on a shared piano. We collaborated with the pianist Barbara Dang for both the performance and experiment design phases.

Our system combines the Dicy2 musical agent system (Nika et al., 2017b) and visual augmentations displayed using a mixed-reality display (either an optical combiner or mixed-reality headset). The instrument used during the project was either a digital piano, or a "MIDIfied" and actuated harpsichord (more specifically, a *virginal*).

3.1 Dicy2 agent

In order to generate music and interact with a musician, Dicy2 builds a "memory" from an audio or symbolic corpus, either offline or online (during the time of the performance), using a Factor Oracle automaton, akin to a variable-order Markov model (Assayag et al., 2006). This memory consists of events associating a symbolic label (based on selected musical features) used for the construction of the Factor Oracle with musical content that will be produced during the generative performance. Different strategies and generative heuristics based on free generation, active listening and anticipation can then be used to navigate the memory (Nika et al., 2017b).

Figure 2: Performance implementation of our system. Left: Seen from the musician's side, visual augmentations are projected on black foam board panels on the other side of a large glass panel. Center: Seen from the audience, the reflection of the projection in the glass overlaps with the piano. Right: With correct lighting, the augmentations appear to be floating above the piano and are correctly perceived by all audience members (here with the MEMORY LOD).

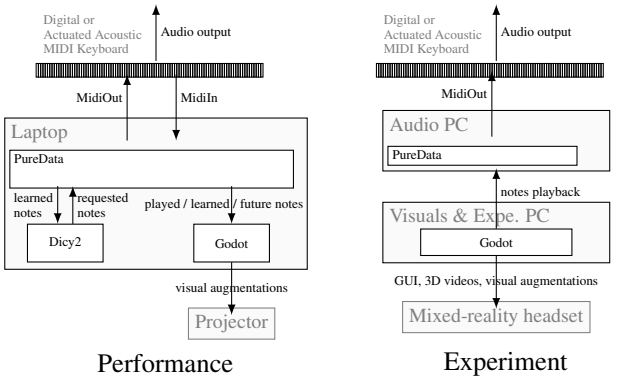

Figure 3: Hardware, software and data flow for the implementation of our co-improvisation system and visual augmentations in performance and experiment contexts.

In the scope of this work, the Dicy2 agent's memory is built online: notes played by the musician are captured in a PureData patch, and sent to the agent (running in a Python script[1]) as events for learning. Each event is labelled with a MIDI note-pitch value and associated with the musical content which includes pitch, velocity, time interval since the previous note, and note duration.

The interaction with the agent is controlled by the musician using a keyboard pedal as a toggle. Whilst pressed, the patch sends a request to the Dicy2 agent for the free generation of 10 musical events, and repeats the request as soon as the events have all been played. The interaction stops when the pedal is released and after the final request has been completed.

When receiving a request, the agent navigates its memory and replies with the appropriate musical content. Moreover, for each musical event, the agent also indicates the position of the event in its memory. While this information does not affect the music generation, it is necessary for the MEMORY visual augmentation described below. The agent's replies are then sent as MIDI data for sound synthesis in the digital piano or for actuation of the *virginal* keys, and are sent to Godot to inform the visual augmentations.

3.2 Visual augmentations with Levels of Detail

Following Capra et al. (2021), we designed five levels of detail (LODs) for the information provided by the visual augmentations, which are depicted in Figure 1.

- NONE does not display any augmentation.
- AMPLIFY displays notes played by the artificial agent as small sticks pushing on the keys, revealing the agent's contribution.
- AGENT adds a visual representation of the agent as an ellipse floating above the keyboard and arrows from this ellipse to the sticks pressing the keys, in effect giving a body to the artificial agent.
- ANTICIPATION adds a visual representation of the preparation of key presses by the artificial agent, with arrows stopping above the keyboard and upcoming notes falling to the keys, in the style of a piano roll.
- MEMORY provides a representation of the artificial agent's memory, with the chronology of learned events and a cursor that indicates where played notes originate from in the agent's memory.

While the second and third levels provide information resembling what can be perceived by the audience from a musician in a performance, the last two levels reveal information beyond that, providing cues to analyse and anticipate the artificial agent's actions.

We implement these visual augmentations differently for performances and experiments, as shown in Figure 3. In both contexts, the augmentations are dynamically generated using the Godot game engine,[2] which communicates through OpenSoundControl messages with a PureData patch, in order to receive: 1) notes as they are being played by the artificial agent and the musician; 2) groups of future notes sent by Dicy2, used to display upcoming notes in ANTICIPATION and MEMORY.

In public performance contexts, as shown in Figure 2, visual augmentations are displayed overlapping the keyboard using an optical combiner (*i.e.,* a semi-transparent mirror, glass or acrylic panel) and a projector. This way, they appear above the keyboard for all audience members looking through the mirror, regardless of their point of view.

For individual controlled experiments, such as the one described below, and as shown in Figure 4, 3D videos and notes from the musician and artificial agent can be played back with a mixed-reality headset. This allows visual augmentations to be controlled dynamically during the experiment and displaying the performances at the original (physical) scale with correct visual depth, in contrast to displaying the performance on a screen. In effect, it aims

[1] https://github.com/DYCI2/Dicy2-python

[2] https://godotengine.org/

to recreate the experience of having the musician and the augmentations in the room.

4 STUDY

To study the effect of the visual augmentations on the audience experience, we investigated the following dimensions:

1. **Objective comprehension** by assessing the ability of the participant to correctly evaluate who played the most notes between the musician and the artificial agent (note-level attributed agency).
2. **Subjective comprehension** by asking participants if they were able to perceive both note-level contributions and higher-level human-AI interactions, through questionnaires and interviews.
3. **Strategies in augmentations selection** adopted by audience members when watching such a performance, evaluated through interface logging, gaze tracking and interviews.

While we recognise that objective comprehension at the note production level (*i.e.,* who played which notes) is only a fraction of what audience members might understand in a co-improvisation performance, we believe that it supports higher-level understanding. Similarly, the aspects of subjective comprehension addressed here (responding, initiating ideas ...), are only a subset of the components of the audience experience, that we try to approach through the interviews.

To conduct these investigations, we presented short videos of co-improvisations between a musician and a Dicy2 agent on a digital piano, with either fixed or controllable augmentations.

4.1 Stimuli

We recorded the 3D videos (with a Zed-mini stereoscopic camera) and MIDI data from a series of co-improvisation performances on a digital piano, in 5 different playing styles. From these, we extracted 15 30-second videos, selecting 5 videos for three ratios of musician/artificial agent contributions (*i.e.,* number of played notes): between 40%/60% and 50%/50% for the equal-contribution condition, and between 30%/70% and 80%/20% for the unequal-contribution condition (more human or more artificial agent). We also extracted three 60-second videos from the same performance that we use for the FREE_AUG block described below.

4.2 Experimental procedure

As shown in Figure 4, participants sat on a chair and wore a Varjo XR3 mixed-reality headset, which displayed the physical room, questionnaires on virtual panels and stereoscopic videos of the performances at their original scale with integrated visual augmentations. The headset also had gaze-tracking capabilities. They held a Valve Knuckles controller in their right hand, used to complete questionnaires with a virtual ray and the trigger, and to select the LOD with the joystick during the training and FREE_AUG blocks. The virtual screen was displayed at 1.3 meters in front of the participants, and the 3D video panel at around 2 meters.

As shown in Figure 5, the experiment consisted of three blocks: a training block with a first video extract with

Figure 4: Participant during the study and view from the mixed-reality headset.

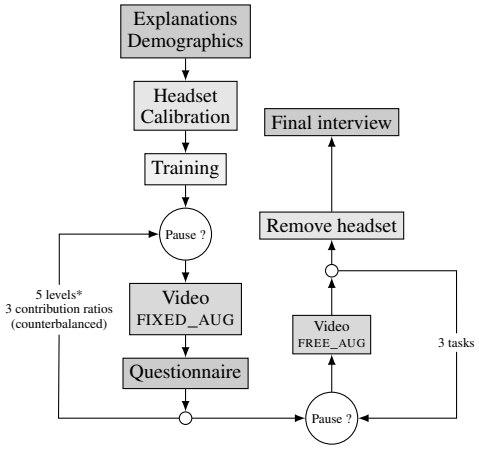

Figure 5: Experiment procedure for each participant in our study.

controllable LOD and the questionnaire, a FIXED_AUG block with 15 videos (3 human–AI contribution ratios × 5 LODs) and a questionnaire, and a FREE_AUG block with 3 videos during which participants could change the LOD dynamically. The experiment ended with a short interview. Overall, the experiment lasted between 35 and 45 minutes.

The questionnaire at the end of each video of the FIXED_AUG block was composed of the following questions (and answers):

- *"I could distinguish what was played by Agent/Musician"* (0/Not at all → 100/Completely),
- *"Who played the most notes?"* (Musician/Agent/Both equally) with a confidence rating (0 → 100),
- *"I enjoyed the performance"* (0 → 100),
- *"I could perceive the dialog between musician and agent"* (0→100),
- *"Who replied most to the other?"* (Musician/Agent/Both equally) with confidence rating (0 → 100),
- *"Who contributed the most new ideas?"* (Musician/Agent/Both equally) with confidence rating (0 → 100).

The instructions during the three tasks of the FREE_AUG block were to select the LOD that allowed them, at all times, to: 1) understand best; 2) enjoy the performance the most; 3) best appreciate the musician's contribution.

During the interview, participants were asked whether visual augmentations were useful to understand and enjoy the performances, which components of the augmentations were useful and which were not, and why, and if

Figure 6: Results for three of the subjective comprehension questions, for all levels of detail. Each question was rated with a score from 0 to 100.

they would like to have such augmentations in human-AI performances in a live context in the future.

4.3 Participants and Analysis

A total of 23 participants took part in the experiment (14 M, 9 F, mean age = 32 years, SD = 9.13). Participants 1–12 had good knowledge about XR and 3D interactive visualisations (as developers or researchers in the field), while participants 13–23 did not. Almost half had no musical practice (N=11), and almost a third did not listen to musical improvisations (N=7).

Quantitative data were analysed using Bayesian statistics with JASP 0.95.4 (JASP Team, 2025). This choice was made in order to allow for a finer discussion of differences or similarities between experimental conditions (here, the LODs) (Kay et al., 2016). We report the natural log value of Bayes factors $LogBF_{10} = X$, which quantifies the relative evidence for the model including the independent variable compared with the null model. We use the interpretation scale proposed by Jeffreys (1961) and implemented in the JASP software: negative values indicate evidence for an absence of difference/effect, positive values indicate evidence for a difference/effect; an absolute $|LogBF_{10}| >$ 4.6 means *decisive evidence*; 4.6 > $|LogBF_{10}|$> 3.4 means *very strong evidence*; 3.4 > $|LogBF_{10}|$>2.3 means *strong evidence*, 2.3 > $|LogBF_{10}| >$ 1.1 means *moderate evidence*, 1.1 > $|LogBF_{10}| >$ 0 means *anecdotal evidence*.

4.4 Subjective measures

We present here the analysis of the subjective measures of the audience experience. These results are plotted in Figure 6.

I can distinguish what was played by the agent and musician: A Bayesian RM ANOVA (Rouder et al., 2017) showed *decisive evidence* for an effect of the visual augmentations $LogBF_{10} = 21.173$. Post hoc tests show *decisive evidence* for differences between NONE and all the levels with augmentations (AMPLIFY: $LogBF_{10} = 8.745$, AGENT: $LogBF_{10} = 8.373$, ANTICIPATION: $LogBF_{10} = 8.167$, MEMORY: $LogBF_{10} = 7.046$), but also *moderate* evidence for an absence of difference between the levels with augmentations (all between $LogBF_{10} = -1.2$ and $LogBF_{10} = -1.5$).

I appreciated the performance: A Bayesian RM ANOVA showed *strong evidence* for an absence of effect of the LOD $LogBF_{10} = -2.894$.

I could perceive the dialog between human and artificial agent: A Bayesian ANOVA showed *strong evidence* for an effect of the visual augmentations $LogBF_{10} = 2.413$.

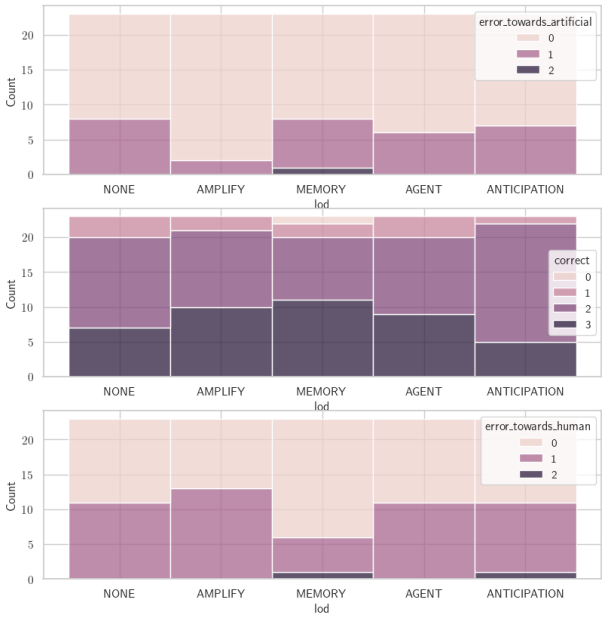

Figure 7: Distribution of answers per participant for the objective question ("Who contributed the most?") for each LOD. Top: Errors in favour of a greater contribution of the artificial agent. Middle: Correct answers. Bottom: Errors in favour of the musician.

Post-hoc tests revealed: *strong evidence* for differences between NONE and AMPLIFY ($LogBF_{10} = 2.024$) and between NONE and MEMORY ($LogBF_{10} = 2.594$); *moderate evidence* for a difference between NONE and ANTICIPATION ($LogBF_{10} = 1.117$); *moderate evidence* for an absence of difference between AMPLIFY and MEMORY ($LogBF_{10} = -1.275$), between AMPLIFY and AGENT ($LogBF_{10} = -1.139$), between AMPLIFY and ANTICIPATION ($LogBF_{10} = -1.230$), and between AGENT and ANTICIPATION ($LogBF_{10} = -1.418$).

Confidence in who replied most to the other: A Bayesian ANOVA showed only *anecdotal evidence* for an effect of the visual augmentations ($LogBF_{10} = 0.675$).

Confidence in who proposed most new ideas: A Bayesian ANOVA showed only *anecdotal evidence* for an effect of the visual augmentations ($LogBF_{10} = 0.964$).

4.5 Objective comprehension

We compared answers to the question *"Who played the most notes?"* with the actual ratio of the level of contribution of the video (see Section 4.1).

Figure 7 depicts the distribution per participant and for each LOD of errors in favour of the artificial agent, of correct answers and of errors towards the musician. Bayesian Contingency Tables (Gunel and Dickey, 1974) showed *decisive evidence* for an absence of effect of the LODs on the proportion of correct answers ($LogBF_{10} = -11.35$), but also on errors in favour of the musician, *i.e.,* overestimating the musician's contribution ($LogBF_{10} = -9.470$), and on errors in favour of the artificial musical agent, *i.e.,* overestimating the artificial agent's contribution ($LogBF_{10} = -7.931$).

4.6 Gaze analysis

Figure 8 depicts the heatmap for each level during the FIXED_AUG block. The main regions which participants looked at appear to shift between levels. While for the first

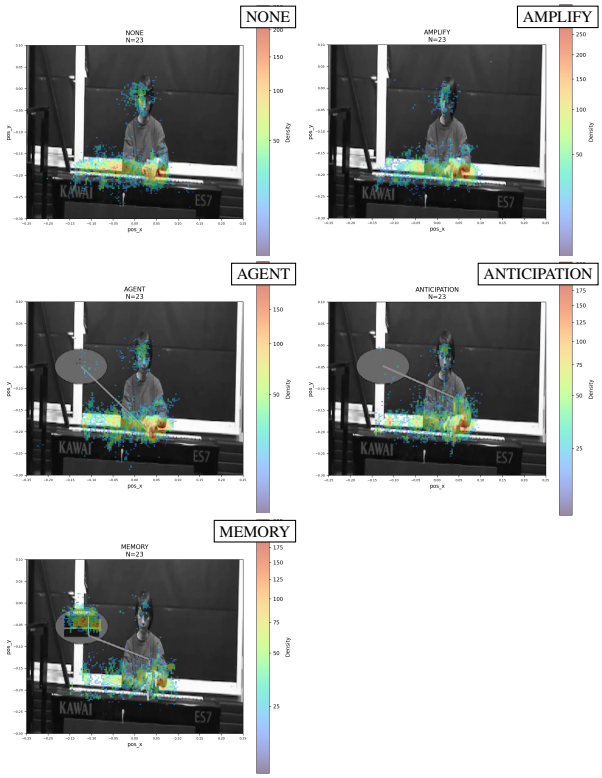

Figure 8: Gaze: Heatmap for each level of augmentation.

two levels (NONE and AMPLIFY) the distribution remains on the face and hands/keyboard, AGENT moves some of the focus to the origin of the rays (*i.e.,* inside the represented agent). ANTICIPATION seems to bring the focus back to the musician and above the keyboard (due to "falling" notes). Finally, MEMORY lowers the focus on the musician's head and shares it between the visualised memory, the hands and the notes. More precisely, Figure 9 presents the ratios of gaze in each region. Due to the noisiness of the data (which explains the proportion of time spent outside an ROI) and overlap between content (*e.g.,* hands and visual augmentations), we only provide a qualitative analysis of these ratios. However, data suggest consistent ratios within one LOD and between LODs. One may observe a large focus on the AI agent in the MEMORY level at the expense of attention to the musician's head in levels with augmentations.

4.7 Strategies in selection of levels

Here, we report on selected LODs during the three tasks of the FREE_AUG block, *i.e., Choose the level that allows you to best understand* (Understanding task), *to best enjoy* (Enjoying task), *to highlight the musician's contribution* (Musician task). Due to a technical issue, the data on selected levels for the first four participants were not recorded, so these results only cover 19 participants.

Figure 10 depicts the accumulated temporal distribution of levels for each task. An overall difference can be observed in the selection of levels, with a relatively more frequent use of higher levels in the *Understanding* task and relatively more frequent use of lower levels for the *Musician* task. However, the plots also show that there is a strong variability between participants, with almost all levels being used across time.

Figure 11 depicts the average ratios of time spent on each level. A Bayesian ANOVA (Rouder et al., 2017) for the *Understanding* task revealed *moderate evidence* of a difference between levels ($LogBF_{10} = 1.889$). Post-hoc tests revealed *moderate evidence* for a difference between NONE and ANTICIPATION ($LogBF_{10} = 1.582$) and *moderate evidence* for an absence of difference between AMPLIFY and MEMORY ($LogBF_{10} = -1.328$). A Bayesian ANOVA for the *Enjoying* task only indicated *anecdotal evidence* for an absence of differences between the levels ($LogBF_{10} = -0.645$). A Bayesian ANOVA for the *Musician contribution* task revealed *decisive evidence* of a difference between levels ($LogBF_{10} = 5.223$). Post hoc tests showed the following differences: *strong evidence* between NONE and AMPLIFY ($LogBF_{10} = 2.033$); *decisive evidence* between AMPLIFY and MEMORY ($LogBF_{10} = 4.541$); *strong evidence* between AMPLIFY and ANTICIPATION ($LogBF_{10} = 2.118$). They also showed *moderate evidence* for an absence of difference between NONE and AGENT ($LogBF_{10} = -1.265$), NONE and ANTICIPATION ($LogBF_{10} = -1.349$), ANTICIPATION and MEMORY ($LogBF_{10} = -1.323$).

4.8 Reflexive Thematic Analysis of Interviews

Two authors conducted a reflexive thematic analysis of the interviews (Braun and Clarke, 2021). One author listened to and transcribed the interviews (each of the 23 interviews lasted 5–10 min, the total transcription amounts to approximately 6,900 words), and the second author coded them. Then the first author reviewed the codes and interviews and proposed a first set of themes. The second and first author finally discussed and refined the themes. We present here the final themes that we extracted.

***Levels of detail had diverse impacts on understanding and appreciation*:** Overall participants highlighted the importance of *"having more information"*(P7) to be able to *"distinguish between agents"*(P22) and perceive their relation and dialogue, especially during fast or subtle gestures of the pianist.

The simplest LOD (AMPLIFY) was deemed the most useful to distinguish the contributions by many participants, especially as the visual augmentations *"remain in the same area as the pianist's hands"*(P18). This choice seemed to be prominent among participants without expertise in XR/3D graphics, which could be explained by the fact that XR experts might tolerate more complex visualisations better. However, this should be investigated in more depth with larger groups of experts and non-experts.

While some participants believed that the representation adds a *"physical presence"*(P4) and would even *"use a human face [...] to create empathy"*(P3), most deemed the representation not useful, especially because the agent was virtual it is fine *"seeing it [only] through the manifestation of notes"*(P18).

Many participants highlighted the relevance of visualising upcoming notes in the ANTICIPATION level, which provided more information on what was going to be played, the notes' position, density and articulation of *"what was going to be played in relation to what was being played"*(P11). In that sense it also helped perceive the interaction and dialog between agents. Some, however,

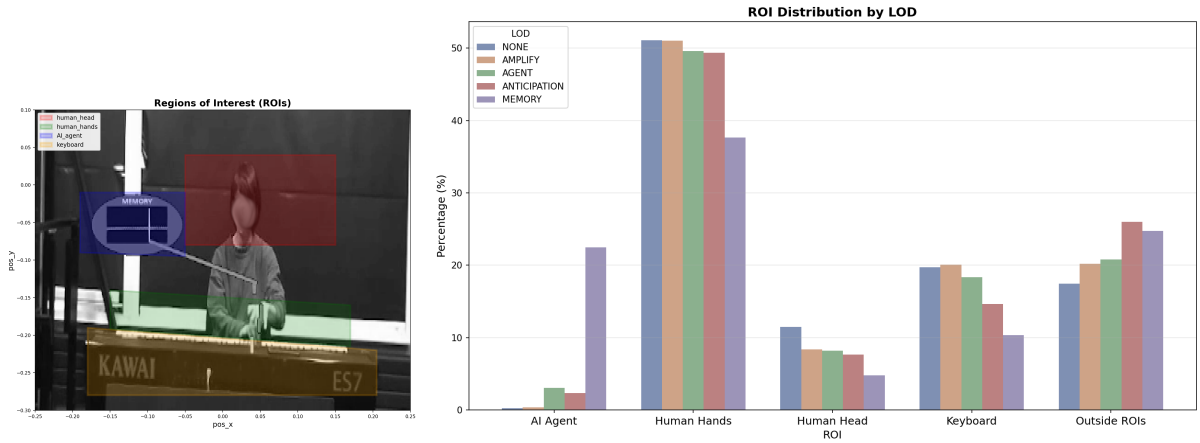

Figure 9: Gaze analysis according to Regions of Interest (ROI). Left: ROIs defined on the stimuli. Right: Ratio of gaze in ROIs depending on the level of augmentation during the FIXED_AUG block.

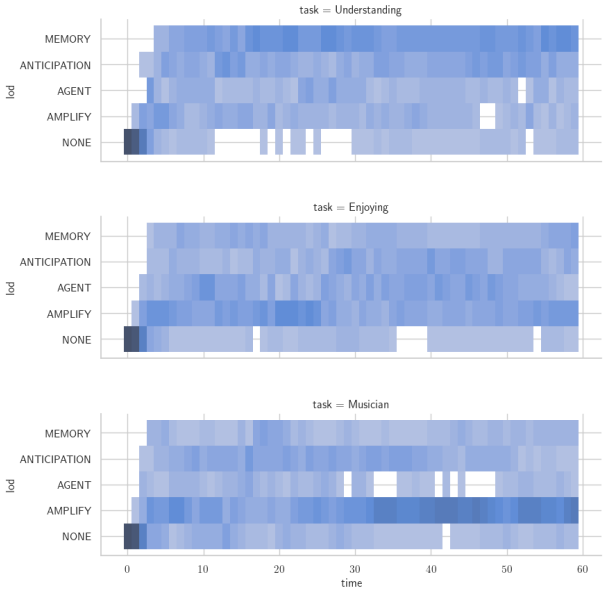

Figure 10: Temporal distribution of the use of LODs in the three tasks of FREE_AUG block.

pointed out that these would be more useful for experts and if one is not *"focused on the moment"*(P22), while for some novices, it was *"enjoyable because visually fluid, without bringing more information, but still poetic"*(P18).

Finally, the MEMORY LOD helped participants both to perceive what the musician had played more clearly and to better understand the agent (*"to see that it draws inspiration from what the musician plays"*(P4)). However, some participants did not understand what was being displayed, others understood but did not judge it useful (*"I feel I know what it is [...] but I did not find it useful"*(P17)).

Visual augmentations might visually overload the performance and mask the musician's contribution: Many participants commented that some of the levels impaired their perception and required more focus because of the amount of visual information. For example with the AGENT level *"the lines [...] hid the hands of the person who was playing"*(P1), they were *"too instantaneous"*(P6) compared to falling notes in ANTICIPATION. They also commented that representing the agent *"puts the musician in the back,*

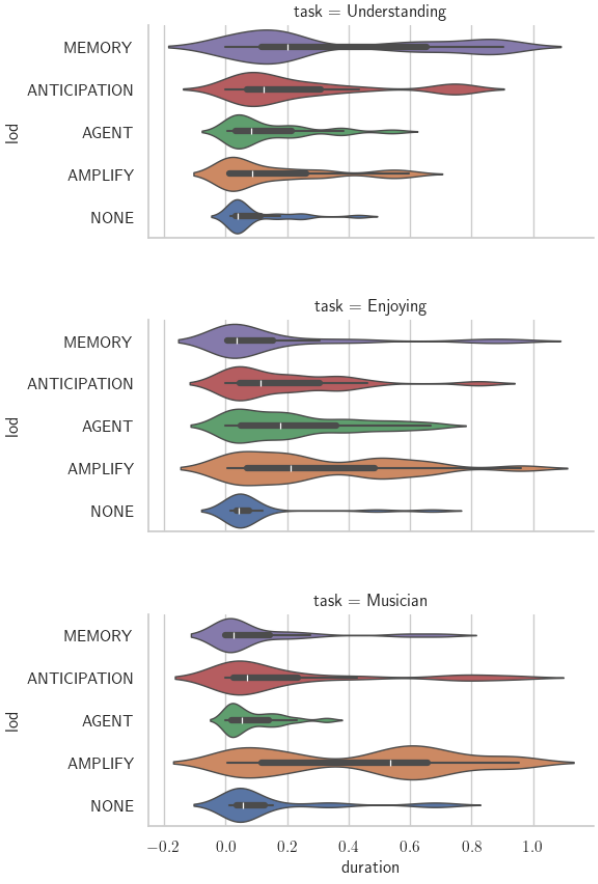

Figure 11: Ratio of time spent in each LOD during the three tasks of FREE_AUG block.

so you see more the agent than the collaboration"(P14), and that the first level (AMPLIFY) *"polluted less the visual aspect of the physical person"*(P18). This also impacted their temporal perception, due to the simultaneous visualisation of *"what has been played, what is being played and what will be played next"*(P15). Some recommended moving the visualisations so that they do not overlap with the musician (*"the notes should come from below"*(P8), *"lines higher do not hide the performance"*(P12)). Others proposed adding visualisations to also highlight the musician's activity, or changing the visualisations depending on the musician's activity (*"memory is interesting when the musician is not playing"*(P16), *"if the pianist did not play I would choose level 3 when the notes were falling"*(P20)).

Different visualisations correspond to different listening strategies so being able to customize the visual augmentations is essential: Participants expressed a variety of strategies, such as using mostly the first level (*"I preferred the levels where one only sees the small key presses"*(P10)), changing depending on what the musician was playing (*"to understand better what was being played I really navigated [levels]"*(P9)), or in order to better perceive what the agent was doing (*"when I had trouble seeing what the agent was playing I usually put more information on that side"*(P18)). Some even suggested additional levels that combine elements such as memory and notes on keys, or "falling" notes and notes on keys (*i.e.,* without representation of the agent). Participants insisted that visual augmentations might need to be removed in order to *"preserve the uncertainty of improvisation"*(P15), but also that *"knowing that the AI is playing is important ethically"*(P15), and that experts might even want to have additional details (*"maybe have the full tablature"*(P23)).

5 DISCUSSION

From the questionnaires, gaze tracking, analysis of changes in levels and interviews, we derive the following insights and guidelines on visual augmentations of co-improvisation for the audience.

5.1 Effect of visual augmentations on the audience experience

Our results suggest that the presence of visual augmentations had no impact on objective comprehension, *i.e.,* understanding the relative contribution (played notes) of the musician and the musical agent. In other words, this comprehension was not better with the augmentations, nor was it worse without them.

However, results on subjective comprehension provide evidence that visual augmentations improve the audience *belief* that they could perceive the respective contributions of human and AI and the dialogue between them. It would therefore appear that visual augmentations had a positive impact, not on the actual understanding of the human-AI co-improvisation but rather on the confidence in being able to perceive at least some of the components of this co-improvisation. Interestingly, this differs from the results of Capra et al. (2020b) which showed that visual augmentations led to a higher agency attributed to the musician (in addition to a higher confidence) in Digital Musical Instruments. Here the augmentations did not alter the perceived contribution, but only the confidence of the audience. Although our questionnaires did not provide evidence for an influence of the augmentations on the enjoyment, thematic analysis of the interviews suggests that many participants favoured at least the first level of augmentation (AMPLIFY) for appreciating or enjoying the performances.

5.2 Audience choice of visual augmentations

Analysis of the selection of LODs, gaze and interviews in the FREE_AUG block allows us to derive a number of insights on the audience experience.

First, many participants favoured the lower LOD (AMPLIFY) which according to the thematic analysis was deemed sufficient to perceive the contribution of the artificial musical agent without hindering the appreciation of the musician's contribution.

Second, exploration of higher levels often depended on the activity of both the musician and the agent, so that they would be used when the musician was less active and/or the artificial agent more active. These levels were however used at the expense of focus on the musician's head and hands, as illustrated by gaze heatmaps and distribution ratios.

Finally individual choices seemed to strongly depend on personal preferences, such as ethical stance towards AI, ways of listening to music and attending performances, but also on expertise regarding musical practice and XR/3D visualisation. For instance some preferred avoiding this additional information in order to preserve the "surprising" and "listening" aspects of performances.

5.3 Guidelines on the design of visual augmentations

From the variety of strategies and preferences described above, an essential first guideline is to either allow audience members to choose their own level of detail on visual augmentations (through personal mixed-reality displays or multiple shared augmented views on the performance). Another possibility is to dynamically change the level displayed for all audience members, for example depending on the musician's and the agent's activity (as it was a recurring strategy of our participants), or as an artistic choice in the performance, with segments that either emphasise or hide the contributions. A second guideline would be to avoid visual clutter by spatially (or graphically) separating the musician's hands and head and the visualisations of the musical agent. One possibility is to have the visual augmentations originating from another part of the instrument (for instance here, below the keyboard), or using different colours to represent the activity of both. It is also essential to remove unnecessary elements (*e.g.,* the representation of the agent) especially if they move the focus away from the instrument and gestures of the musician, for example because of frequent changes.

5.4 Limitations and Future Work

A first limitation of our research pertains to the experimental design of our study. We aimed at measuring the influence on the objective comprehension by asking the ratio of notes played by the musician and the musical agent. Participants made very few errors in stimuli with a stronger contribution of one of them, and more errors where the contribution was equal by design. Although the results suggest an absence of difference between levels of detail in all contribution ratios, the task might not have been sufficiently demanding to perfectly assess the understanding. An alternative would be to have participants identify

who contributed a specific musical phrase heard during the performance, as was done by Capra et al. (2020b).

The musical style chosen for the study stimuli (free improvisation) may also have had a strong impact. This suggests a need to replicate the study with different music styles and different cultural contexts (Déguernel and Sturm, 2023) in order to control their effect on the audience experience. Moreover, our protocol relies on a mixed-reality headset and 3D filmed performances, in order to obtain fine-grained data on gaze and precisely controlled presentation of the stimuli, at the same scale as a physical musician. We believe that this is an important improvement over previous experiments on the subject which were conducted on small 2D screens. However, the conditions remain very different from those of a live performance, and a complementary *in situ* experiment with a live musician and multiple audience members might bring further insights.

Finally, as noted multiple times in this paper, the objective comprehension of note-level contribution by musicians and artificial agents represents only a small fraction of the audience experience, which however supports higher levels of comprehension of musician-AI interactions. These should also be investigated, both on the subjective and objective sides, together with their impact on the overall audience experience.

6 CONCLUSION

In this paper, we explored the design of visual augmentations of human-AI piano co-improvisations with multiple levels of detail, and studied how these augmentations influence the audience experience. While visual augmentations did not improve the objective comprehension of the respective contributions of the artificial agent and of the musician, our results indicate that adding any visual augmentations may improve subjective comprehension, *i.e.,* how well the audience believes they can perceive the contribution. Furthermore, gaze tracking, usage of levels of augmentations and interviews suggest that different levels of information should be offered to audience members, in order to let them choose based on personal preferences and expertise. We believe that these results can inform the design of future co-improvisation systems and contribute to the understanding of the audience experience with digital musical instruments.

AUTHOR DECLARATIONS

This research was approved by the Ethics committee of the University of Lille under number *AVAALOD:2024-818-S132*.

REFERENCES

Assayag, G., Bloch, G., Chemillier, M., Cont, A., and Dubnov, S. (2006). OMax brothers: A dynamic topology of agents for improvization learning. In *Proceedings of the ACM Workshop on Audio and Music Computing for Multimedia*, pages 125–132. https://doi.org/10.1145/1178723.1178742.

Barbosa, J., Calegario, F., Teichrieb, V., Ramalho, G., and McGlynn, P. (2012). Considering audience's view towards an evaluation methodology for digital musical instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. https://doi.org/10.5281/zenodo.1178209.

Berthaut, F., Coyle, D., Moore, J., and Limerick, H. (2015). Liveness through the lens of agency and causality. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. https://hal.archives-ouvertes.fr/hal-01170032.

Bonnasse-Gahot, L. (2014). An update on the SoMax project. Technical report, Ircam-STMS, http://architexte.ircam.fr/textes/BonnasseGahot14a/index.pdf.

Braun, V. and Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, https://doi.org/10.1080/14780887.2020.1769238.

Bryan-Kinns, N. (2024). Reflections on explainable AI for the arts (XAIxArts). *Interactions*, 31(1):43–47, https://doi.org/10.1145/3636457.

Capra, O., Berthaut, F., and Grisoni, L. (2020a). A Taxonomy of spectator experience augmentation techniques. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. https://hal.science/hal-02560931.

Capra, O., Berthaut, F., and Grisoni, L. (2020b). Have a SEAT on stage : Restoring trust with spectator experience augmentation techniques. In *Proceedings of the ACM Designing Interactive Systems Conference*, pages 695–707. https://doi.org/10.1145/3357236.3395492.

Capra, O., Berthaut, F., and Grisoni, L. (2021). Levels of Detail in Visual Augmentations for the Novice and Expert Audiences. *Computer Music Journal*, pages 92–107, https://hal.science/hal-03106626.

Déguernel, K. and Sturm, B. L. T. (2023). Bias in favour or against computational creativity: A survey and reflection on the importance of socio-cultural context in its evaluation. In *Proceedings of the International Conference on Computational Creativity*. https://hal.science/hal-04109783/.

Déguernel, K., Vincent, E., and Assayag, G. (2018). Probabilistic factor oracles for multidimensional machine improvisation. *Computer Music Journal*, 42(2):52–66, https://doi.org/10.1162/comj_a_00460.

Erdem, Ç., Wallace, B., Glette, K., and Jensenius, A. R. (2022a). Tool or actor? expert improvisers' evaluation of a musical AI "toddler". *Computer Music Journal*, 46(4):26–42, https://doi.org/10.1162/comj_a_00657.

Erdem, Ç., Wallace, B., and Refsum Jensenius, A. (2022b). CAVI: A coadaptive audiovisual instrument–composition. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. https://doi.org/10.21428/92fbeb44.803c24dd.

Fyans, A. C., Gurevich, M., and Stapleton, P. (2010). Examining the spectator experience. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 451–454. https://doi.org/10.5281/zenodo.1177775.

Gunel, E. and Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557, https://doi.org/10.1093/biomet/61.3.545.

Herremans, D., Chuan, C.-H., and Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Computing Surveys*, 50(5):1–30, https://doi.org/10.1145/3108242.

JASP Team (2025). JASP (Version 0.95.4)[Computer software]. https://jasp-stats.org/.

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press, ISBN: 9780198503682.

Jung, M. (2023). Intelligent music performance systems: Towards a design framework. *Studia Musicologica Norvegica*, 49(1):28–44, https://doi.org/10.18261/smn.49.1.3.

Kay, M., Nelson, G. L., and Hekler, E. B. (2016). Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 4521–4532. https://doi.org/10.1145/2858036.2858465.

Kizilcec, R. F. (2016). How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 2390–2395. https://doi.org/10.1145/2858036.2858402.

Lewis, G. E. (2000). Too many notes: Computers, complexity and culture in Voyager. *Leonardo*, 10:33–39, https://doi.org/10.1162/096112100570585.

Lucas, P. P. and Fasciani, S. (2023). A human-agents music performance system in an extended reality environment. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 10–20. https://doi.org/10.5281/zenodo.11189090.

McCormack, J., Gifford, T., Hutchings, P., Llano Rodriguez, M. T., Yee-King, M., and d'Inverno, M. (2019). In a silent way: Communication between AI and improvising musicians beyond sound. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300268.

Nika, J., Chemillier, M., and Assayag, G. (2017a). ImproteK: Introducing scenarios into human-computer music improvisation. *Computers in Entertainment*, 14(2):1–27, https://doi.org/10.1145/3022635.

Nika, J., Déguernel, K., Chemla-Romeu-Santos, A., Vincent, E., and Assayag, G. (2017b). DYCI2 agents: merging the "free", "reactive", and "scenario-based" music generation paradigms. In *Proceedings of the International Computer Music Conference*. https://hal.science/hal-01583089/.

Perrotin, O. and d'Alessandro, C. (2014). Visualizing gestures in the control of a digital musical instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 605–608. https://doi.org/10.5281/zenodo.1178901.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., and Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, 22(2):304–321, https://doi.org/10.1037/met0000057.

Tatar, K. and Pasquier, P. (2019). Musical agents: A typology and state of the art towards musical metacreation. *Journal of New Music Research*, 48(1):56–105, https://doi.org/10.1080/09298215.2018.1511736.

Wang, Y. and Martin, C. (2024). Off-the-shelf: Improvising with a minimal intelligent musical instrument in mixed reality. In *Proceedings of the AI Music Creativity Conference*. https://aimc2024.pubpub.org/pub/ll85912p.

Wooldridge, M. (2009). *An Introduction to Multiagent Systems*. John Wiley & Sons, ISBN: 978-0-470-51946-2.

Xiao, X., Tome, B., and Ishii, H. (2014). Andante: Walking figures on the piano keyboard to visualize musical motion. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 629–632. https://doi.org/10.5281/zenodo.1178987.