In a Silent Way

Communication between AI and improvising musicians beyond sound

Jon McCormack
SensiLab, Monash University
Caulfield East, Australia
Jon.McCormack@monash.edu

Toby Gifford
SensiLab, Monash University
Caulfield East, Australia
Toby.Gifford@monash.edu

Patrick Hutchings
SensiLab, Monash University
Caulfield East, Australia
Patrick.Hutchings@monash.edu

Maria Teresa Llano Rodriguez
Goldsmiths, University of London
London, United Kingdom
m.llano@gold.ac.uk

Matthew Yee-King
Goldsmiths, University of London
London, United Kingdom
m.yee-king@gold.ac.uk

Mark d'Inverno
Goldsmiths, University of London
London, United Kingdom
dinverno@gold.ac.uk

Figure 1: Training with percussionist and instrumentalist improvisation (left); performance with the AI Improviser (right)

ABSTRACT

Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn't exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the collaboration. We developed a collaborative improvising AI drummer that communicates its confidence through an emoticon-based visualisation. The AI was trained on musical performance data, as well as real-time skin conductance, of musicians improvising with professional drummers, exposing both musical and extra-musical cues to inform its generative process. Uni- and bi-directional extra-musical communication with real and false values were tested by experienced improvising musicians. Each condition was evaluated using the FSS-2 questionnaire, as a proxy for musical engagement. The results show a positive correlation between extra-musical communication of machine internal state and human musical engagement.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300268

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Sound and music computing*.

KEYWORDS

AI Systems, Improvisation, Extra-musical Communication

ACM Reference Format:
Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. 2019. In a

Silent Way: Communication between AI and improvising musicians beyond sound. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300268

1 INTRODUCTION

In this paper we investigate the benefits of extra-musical interaction in real time music improvisation and co-creation with an artificially intelligent creative partner. Recent advances in Artificial Intelligence (AI) techniques, coupled with increasingly powerful computer resources, make it feasible to engage in artistic collaborations with a machine intelligence, because it exhibits degrees of creative agency [3] and autonomy [2, Chapter 9] which traditional tools or instruments – either digital or analogue – do not possess [12, 30].

Interactions between musical improvisers largely occur through the music itself, although visual and other extra-verbal cues play an important role [41]. We explore the role of extra-musical cues and feedback between human and AI performers to achieve better performance outcomes with a longer term goal of establishing trust between human and machine – widely acknowledged as an important factor in successful human group improvisations [42, 45]. We are also interested in understanding how human-AI improvisations with and without extra-musical cues are perceived by audiences; our research also evaluated this aspect of resultant performances between human and AI.

Improvisation and Extra-musical Communication

Live performance and improvisation are amongst the most challenging creative activities undertaken by humans. To do them successfully requires a great deal of proficiency and virtuosity which typically takes many years of practice and experience before one can claim anything close to mastery [14, 17, 35]. In musical improvisation the ability to communicate allows participants to be aware of and understand the behaviour and intentions of those involved. Working within an improvised performance setting without some level of understanding of what others are experiencing makes it very difficult for performers to create settings that inspire each other and take the performance forward.

When it comes to improvising with a non-human performer, important traditional cues and indicators may be missing: body language and movement, eye contact, visual cues, etc., do not exist. For the performer this means that all of the AI's intentions must be inferred through the musical output. This makes it more difficult to build trust and, therefore, to take risk during a performance; both trust and risk-taking are widely considered important aspects of successful group collaboration and teamwork [23, 45].

Musical Engagement and Flow

Theories of human musical engagement typically posit a phenomenological state known as *flow* [7] as both an underlying driver and fundamental metric of engagement [10, 22, 48]. A flow state is characterised by a number of cognitive, affective and psychophysiological indicators. Cognitive factors include a sense of effortless control and complete focus [43]. Affective factors include loss of self-consciousness and high intrinsic motivation [26]. Psychophsyiological markers of flow include salivary cortisol, blood pressure, and heart rate variability [10] as well as skin conductance [32].

Flow is theorised to be an important aspect of successful group improvisations [21, 38, 41] and has also been discussed as an engagement metric for human-machine creative partnerships [16, 34].

Biometrics During Music Performance

Biometrics can provide real-time quantification of human psychophysiological state with relatively minimal distraction, and thus may be useful for communicating human mental state to an AI during live performance. Given the importance of flow states to group musical improvisation, the biometric markers discussed above for flow suggest themselves as relevant variables to be exposed. In this experiment we selected skin conductance as an indicator of human internal state.

Skin conductance (SC) has been studied in the context of music performance by Dean & Bailes who note that "SC measures are frequently interpreted as an index of not only emotional response ... but also task effort and attention" [11]. They find that time-series analysis of real-time SC can predict musically salient features of an improvisation. The relationship between SC and human internal state is, however, not straightforward, perhaps suggest Namakura & Roberts due to the "inherent complexity of flow experiences" [32]. Of interest in this study then was whether or not a machine learning system could extract useful information from a real-time SC measure.

Aims and Contributions

This study tests two related hypotheses in human-machine musical improvisation. We hypothesise that extra-musical communication of 'internal state':

(1) ***of* the human musician *to* the machine** can enhance the machine's capacity to generate appropriate and complementary improvised output;
(2) ***of* the machine *to* the human** can facilitate more engaging human-machine musical interactions.

Our study employs a factorial design, where a series of human-machine musical improvisations are evaluated under combinations of conditions: with/without human-to-machine extra-musical communication, and with/without machine-to-human extra-musical communication.

Results of both detailed evaluations from improvising musicians using the system and a listening study of 100 external observers conducted on-line support the use of machine-to-human extra-musical communication in the form of machine confidence visualisations. When truthful, such visualisations were found to produce a higher reported flow state [8] in performers than reported when using deceptive or absent visualisations, on average. Significantly more listeners perceived a better musical balance between machine- and human-performed instrumental parts in recordings made when truthful, rather than deceptive, confidence visualisations were used during recording.

Human-to-machine extra-musical communication, in the form of skin conductance measurements, was not found to have a significant effect on either training the AI system or on the flow state of performers, probably due to confounding noise from muscle movements.

2 RELATED WORK

Many different factors have been identified as significant to increasing trust in human-machine collaborations, including reliability, predictability, utility, provability, transparency and explainability [27, 28, 31]. We are particularly interested in how revealing intrinsic aspects of the workings and purpose of both humans and machines can influence trust and promote engagement. In other words, we are interested in revealing the state of a human-machine collaboration in a way that helps both parties understand the interaction taking place. Sawyer [40] suggests that improvisational creativity in a collaborative performance is achieved through ephemeral signs; thus, the way an AI improviser communicates must be simple and precise, yet the communication itself needs to be meaningful so that the new information helps progress the performance.

Extra-musical Communication

Research addressing issues in extra-musical communication between human and AI improvisers is currently in its infancy.

Weinberg, Hoffman and Bretan developed a series of expressive robotic improvising musicians [5, 19], most notably their improvising jazz marimba robot *Shimon*. The physical embodiment of *Shimon* is an important aspect of their research, both in terms of its extra-musical communication through movement, which affords human-machine temporal co-ordination through anticipation, and its visually animated appearance, lending it the impression of musical personality [4]. Bretan suggests that an obvious next step for this area of research is to incorporate "social cues" to convey musical emotion and "lead to more convincing performances by the robot in which the system looks truly expressive" [5].

Ravimukar et al. [36] describe a research proposal centred on the question: "will the addition of two-way extra-musical notational communication enhance the human's experience of coordinating musical transitions with AI music partners?" which bears similarity to our study, although focusing on notational communication for temporal anticipation of musical changes. Their study is in progress, and does not appear to have reported results at the time of writing.

Skin conductance (also known as *galvanic skin response*) has been used to train generative music systems with the goal of producing controlled, affective output. Kim and André [25] utilised galvanic skin response, along with electrocardiogram, electromyogram and respiration data, as input for a generative music system that used genetic algorithms. While the input is provided by a human listener they are not presented as a collaborative partner in producing music together. Hamilton [18] developed a real-time composition system that records galvanic skin response of a human performer and uses this measure as input for a software composition system that generates notes on a score for the performer to play in real-time.

AI generative systems and confidence

Neural network-based systems have been utilised for symbolic music generation for decades [13] and have seen a resurgence along with the deep learning boom of the last five years. Increased model complexity has seen improvements in effective memory, expressive range and consistency of generated outputs, making neural networks a good candidate for collaborative AI music systems.

Recurrent Neural Networks (RNNs) have been the predominant neural network architecture for music generation tasks [6], including drum sequence generators [20, 29]. Recent research has shown however that Temporal Convolutional Network (TCN) models [37] can perform just as well or better in the analysis of sequential data [1]. TCN models also have a greater number of parallel pathways that allow for models to be trained significantly faster on GPUs.

In sequential data models, a softmax layer can be used to predict upcoming events and post-assessed in terms of an entropy metric, or an accumulated improbability score. Softmax layers output real values between 0 and 1 such that all values in a specified dimension sum to unity, and as such can be used to represent a probability distribution.

When creating a generative system from a predictive model, probability distributions can be sampled to generate an output. By doing so, the distribution is collapsed into a single choice and the majority of data output from the network itself is discarded.

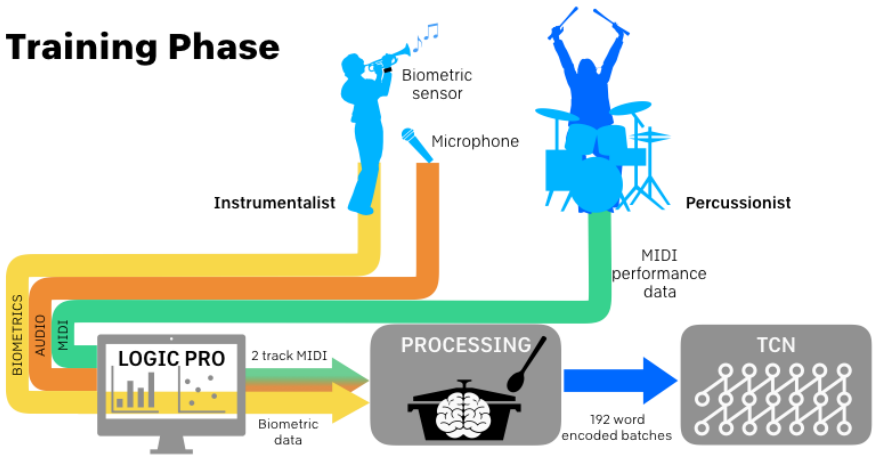

Figure 2: Training the machine improviser with improvisations on the electronic drum-kit and melodic instrument.

3 SYSTEM IMPLEMENTATION

We implemented a machine improviser and studied the experience of musical co-creation with it. The machine improviser is software utilising a Temporal Convolutional Neural Network (TCN) [37] – a machine learning system to drive real-time algorithmic generation of percussive accompaniment in a musical duo. In particular, the study probes the efficacy of bi-directional communication of 'internal state' via extra-musical communicative channels between improvisers.

The research design included data gathering from human musicians in two distinct stages: (i) training of a TCN software drummer based on human-human improvisational duets (Figure 1, left and Figure 2), and (ii) evaluation of the experience of co-creating with the trained TCN under various conditions of extra-musical communication (Figure 1, right and Figure 5).

The network was trained on 3 hours of human duet improvisation by experienced improvising musicians. Duets were performed between an instrumentalist playing melody on a monophonic instrument (variously saxophone and clarinet) and a drummer playing an electronic drum-kit.

Musical performance data collection

The drum-kit, a Roland TD-50, is an electronic kit designed to closely emulate the sound and feel of a standard acoustic drum-kit. It records and transmits performance data in extended precision (14 bit) MIDI format, whilst also synthesising emulated acoustic sounds via physical modelling [46].

Musical input from the human instrumentalist (clarinet, saxophone) was recorded as audio and algorithmically transcribed into MIDI format using Logic Pro's *Flex Audio*[1] feature.

The training data was gathered over four sessions, with two different drummers and two instrumentalists in all combinations. The sessions took place in a recording studio, and utilised a 'click-track' to define the underlying tempo and beat.

Each session had 9 exercises comprising combinations of 3 musical styles and 3 performance techniques in a factorial design. The styles were (i) Swing, (ii) Funk and (iii) Rock, all in common (4/4) metre, at a fixed tempo of 120 bpm. The performance techniques were (i) melodic lead with percussive accompaniment, (ii) trading groups of four measures between performers, and (iii) trading groups of two measures.

The use of a click-track facilitated symbolic transcription of the recorded duet improvisations as beat-relative note events, quantised to 12 time steps per beat, allowing triple and duple subdivisions down to the resolution of semiquaver triplets. After symbolic transcription, the musical data was tokenised for feeding into the TCN model. Each four-beat measure of the performance then comprised a 48 token string.

Tokens conveyed if a note was started (H) or sustained (s) by the melodic instrumentalist and the onset velocity, quantised to four bands (p,mp,mf and f). These tokens were concatenated with tokens for each drum hit encoded as midi pitch values and the same velocity metric to form longer tokens. This example phrase segment shows how sequences were tokenised with this method: *38mp o 36mf|38mf|44mf o*

[1]https://ask.audio/articles/logic-pro-x-tutorial-flex-pitch

38mp o o 36mp|38mp Hmp s s|38mp Hmp s|38mp s s s. A total of 1639 unique tokens were required to encode the training data. Of these, 1188 tokens were used less than 20 times and were replaced with the silent token 'o', leaving 451 tokens used in the corpus.

Biometric data collection

In addition to translating the musical performance data as input for the TCN model, the instrumentalist wore an Empatica E4 biometric wristband[2], which recorded real-time skin conductance (SC). Following Dean and Bailes [11] we utilised change in skin conductance (∂SC) as a real-time parameter containing information about the human musician's mental state and cognitive music processing.

The Empatica E4 wristband reports skin conductance in microSiemens at a sampling rate of 4 Hz. Baseline skin conductivity varies between people and its absolute level as measured by the wristband depends on how tightly the band is fitted. As such we used a relative measure of change in skin conductance defined by

$$\partial \mathrm{SC}_t = \frac{\mathrm{SC}_t - \mathrm{SC}_{t-1}}{\sigma_{\mathrm{SC}}}$$

where σ_{SC} is the standard deviation of the skin conductance over 2 minutes prior to commencing the improvisation session. A further 1 minute period (minimum) waiting period was used from the time the wristband was put on before commencing baseline calibration, to exclude the large changes related to fitting the wristband. The real-time value of ∂SC was quantised to 3 discrete levels:

$$Q_{\partial \mathrm{SC}} = \begin{cases} \text{High} & \partial \mathrm{SC} \geq 1 \\ \text{Med} & -1 < \partial \mathrm{SC} < 1 \\ \text{Low} & \partial \mathrm{SC} \leq -1 \end{cases}$$

[2]https://www.empatica.com/research/e4/

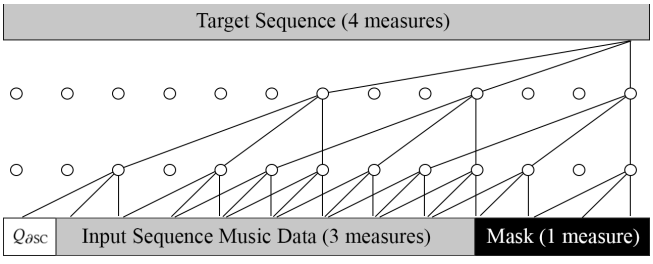

Figure 3: A simplified representation of the TCN model, showing how the final token of the target sequence is predicted with the use of skin conductance and quantised music data.

Across the performance sessions High, Medium and Low values were recorded for 15%, 65% and 20% of the measurements respectively. The quantised signal was then sampled at the start of each musical measure and communicated to the machine improviser via an OSC [47] request and response.

For communication to the human musician of the machine improviser's internal state, we expose a proxy index of the evolving 'confidence' reported by the neural net regarding its musical decisions, calculated from its internal distribution of probabilities. We elaborate on this later in this section.

Training the machine improviser

The machine improviser is a software system comprising a TCN that generates performance data for a drum synthesiser to perform over the next measure of the performance, run consecutively every measure. The TCN model [1] was trained to predict combinations of notes played in any given measure of performance based on musical and biometric data from the three measures directly preceding it. By sampling from these predictions, a generative system was produced.

The improvisation sessions used for training data provided a set of 4195 sequences of four measures which was divided into training (76%), validation (12%) and test sets (12%), which provided a natural split in the data.

During training, the network used four measures of performance data for input and target sequences. The fourth measure of the input sequence was masked to prevent the system from using any time step from the fourth measure to generate predictions. This allows the whole fourth measure to be predicted (and generated) at once, in parallel.

To facilitate an effective memory to cover the full 192 time steps in four measures, dilations were used through seven convolution layers. With a kernel size of 3 and 192 units per hidden layer, a receptive field of 257 time steps was achieved. The model used for training can be found at https://github.com/patHutchings/TCN/tree/Machine-Improviser

Two unique data preprocessing features were implemented for training and inference with the TCN model. In sequence-to-sequence tasks, where the next token in a sequence is predicted, the input sequence is typically offset from the target sequence by inserting a '<start>' token or similar at the beginning of the input sequence. $Q_{\partial \mathrm{SC}}$ from the last measure is used as the start token, making it visible at every step for training and inference (Figure 3).

An extensive hyper-parameter search was conducted to reduce model complexity without loss of predictive accuracy, as measured by per-token perplexity on a validation set of 500 four measure strings. Perplexity represents the average number of most probable tokens to appear next in the sequence. Complexity was reduced by using a small token embedding of only 20 dimensions and limiting effective

Figure 4: Still images of the visualiser at different levels of confidence ranging from 0.1 (low confidence) to 0.9 (highly confident). The expression changes continuously in response to the confidence of the machine improviser in addition to nodding in time with the current tempo. Sustained high levels of confidence result in a pulsating glow behind the emoticon in time with the beat.

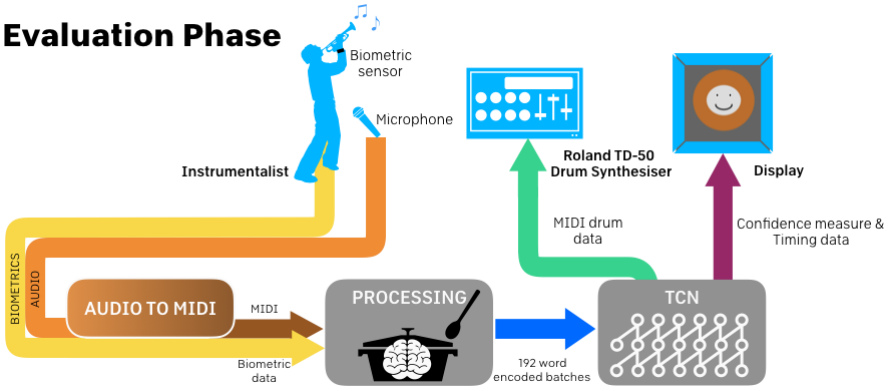

Figure 5: Evaluating the machine improviser with instrumentalists.

memory to four measures. The small model size allows inference of a measure of drum performance within 5ms on a consumer level 1080ti graphics card: less than a single time step in the quantised performance.

The silent 'o' token appeared an order of magnitude more frequently than any other token in the training data and the TCN quickly learned that predicting entirely silent sequences would still return a small loss. To counteract this effect a weighting of 0.1 was applied to 'o' token contributions to loss calculations.

A perplexity of 6.70 was achieved on the validation set after 69 epochs of training with batches of 16 sequences. This means the system finds an average of 6.70 tokens that are most likely to occur next at any point in the sequences, from the 451 tokens it knows. A perplexity of 6.68 was achieved when $Q_{\partial SC}$ was replaced with a standard '<start>' token, indicating that the inclusion of skin conductance data did not assist or impede the training of the network. The model with biometrics was used for the machine improviser and a final test set evaluation was performed, producing a perplexity of 7.34.

A softmax layer was used as the final layer of the TCN, so that outputs could be used as probability distributions for sampling in the generative system. The probability of any of the 451 tokens in the corpus dictionary being played at any timestep in the upcoming measure is used to make a weighted selection. The machine improviser uses only tokens observed in the training data. With this dictionary the system has 2.5×10^{127} possible outputs at each measure. Tokens are decoded to MIDI messages and sent to the drum synthesiser for performance.

Visualising Confidence

The AI's confidence was conveyed visually using an emoticon-style face (Figure 4) drawn with simple vector graphics, that bounces in time with the music with a refresh rate of 60hz. Confidence values were updated every 0.5 seconds and reflected on the display within 5ms. When system confidence is low, the face frowns and eyes shift in different directions, avoiding eye contact with the viewer. When confidence is high, the face smiles, its eyes widen, and it maintains eye contact with the viewer. Sustained high states produce a radiating glow behind the face that pulsates in time with the beat.

The use of a simplified, iconic representation of a face was chosen to minimise additional cognitive load on the musician, mimicking facial expressions at a high level of abstraction in ways similar to how human performers behave when they

lack confidence in performance. Iconic facial expressions are easy to process, but also avoid over-anthropomorphising the AI, which may potentially lead to false assumptions regarding the level of cognition, emotion, and behavioural characteristics if a more realistic or nuanced human face were used, for example.

4 PERFORMER EVALUATION

After developing and training the machine improviser, we then evaluated it musically through trials with improvising musicians. We describe the methodology and results below.

Method

We recruited seven experienced instrumentalists who each engaged in four improvised sessions for a total of 28 trials. Recruited participants were given a $15 department store voucher for their time.

Three female and four male instrumentalists aged between 18 and 56 took part in the evaluation. While all had experience with improvisation, different approaches to improvisation were reflected in the range of instruments (saxophone, clarinet, vocals and electronic keyboard), styles of familiarity (jazz, rock, funk, experimental, folk and classical) and roles (professional musicians, serious amateurs and skilled hobbyists).

In each session the instrumentalist improvised with the machine improviser for 3 minutes and then self-reported on the experience using the FSS-2 SHORT scale [8]. The FSS-2 SHORT scale comprises 9 items (i.e. questions) intended to capture the 9 dimensions of flow as described by Csikszentmihalyi [8].

Application of the FSS-2 suite of scales to musical performance has found that these items have differing correlation with other external flow metrics, and so a subset of the items may be more appropriate for measuring flow in musical contexts. Wrigley and Emerson [48] found that the "subscales of Sense of Control, Autotelic Experience, and Challenge-Skill Balance showed the strongest associations and explained the most variance" in live music performance.

Evaluations took place in the same recording studio used for data collection, with the addition of a portable screen, positioned at standing height, which displayed the confidence visualisation (see Figure 2, right and Figure 5).

We developed three different confidence visualisation conditions to differentiate effects driven by visualisation design choices from those driven by the content it is being used to communicate. Confidence was (i) truthfully communicated by the visualisation system, (ii) inverted to create a deceptive visualisation and (iii) not communicated by removing all facial features of the emoticon to create an 'absent' confidence visualisation. Inversion was used for the deceptive condition as an 'absent' condition was tested for and inverted patterns provided changes between the communication states with the same transition dynamics.

Instrumentalists were given a short verbal brief to communicate that the visualisations were communicating a metric of confidence of the improvising machine in predicting what to play next. It was emphasised that the visualisation is not a judgement on the instrumentalist's playing or overall quality of the musical performance. The 'intelligence' of the improvising machine was described as generating musical improvisations by repeating patterns learned to be useful in accurately predicting what might happen next in improvised duets we recorded.

Four of six possible conditions were tested with randomised order for each instrumentalist (see Table 1). Priority was placed on having longer sessions of improvisation without fatiguing instrumentalists physically and creatively.

Results and Discussion

We found that the **visualisation of machine confidence** noticeably **affected the tendency** of the instrumentalist **to achieve flow.** The biometric communication via the instrumentalist's skin conductance did not make any discernible difference to the experience of improvising with the system. These results emerged from comparison of an aggregate flow measure derived from the 9 survey items in the FSS-2 responses, compared between conditions across participants.

We performed a Principal Components Analysis on the FSS-2 responses (comprising 4 sessions x 7 participants = 28 responses for each of the 9 questions). The first principal component was consistent with the findings of Wrigley and Emerson in having substantial loadings for Sense of Control, Autotelic Experience, and Challenge-Skill Balance, and negligible loadings for Clarity of Goals and Transformation of Time. As such, we utilised an aggregate index of flow comprising the average of the questions relating to Sense of Control, Autotelic Experience, and Challenge-Skill Balance. For completeness we additionally ran all the analyses with an aggregate index constructed using the numeric weights contained in the first principal component of our data, which did not qualitatively change any of our conclusions.

The visualisation condition had a measurable impact on the aggregate flow index. The Truthful condition was more

Table 1: Test conditions with truthful (T) and deceptive (D) biometric data for each visualisation type. Random selections represented with (T/D).

	Confidence Visualisation		
	Truthful	Deceptive	Absent
Condition	Bio T/D	Bio T/D	Bio T and Bio D

Table 2: flow index *vs.* visualisation condition

Participant	Deceptive	Absent	Truthful
1	3.67	4.33	4.33
2	3.67	4.17	4.33
3	3.33	4.17	4.33
4	4.33	4.17	3.67
5	4.00	2.83	3.67
6	4.00	3.16	4.00
7	2.00	3.33	3.67
mean	**3.57**	**3.74**	**4.00**
s. d.	0.76	0.61	0.33

flow-inducing than the Absent condition, which in turn was more flow-inducing than the Deceptive condition. For 4 of the 7 participants (P1, P2, P3, P7) this relationship was monotonic across the three conditions, and for 5 of the 7 participants (P1, P2, P3, P6, P7) the Truthful condition showed equal or better flow induction than the Deceptive condition. Of the remaining 2 participants, 1 (P4) showed an inversely monotonic relationship, and the other (P5) strongly preferred either Truthful or Deceptive visualisation over Absent visualisation in terms of their flow index.

In order to assess the significance of these trends we performed matched-pair t-tests on each of the three between-condition combinations of the machine confidence visualisation: Absent *vs.* Deceptive, Truthful *vs.* Absent, and Truthful *vs.* Deceptive. The number of instrumentalists in the study is small (N = 7). Student's t-test was originally developed for statistical inference from small samples [49], and Cummings [9] recommends reporting of 95% confidence intervals derived from the t-distribution and the sample standard deviation for sample sizes between 5 and 30. Whilst some researchers promote non-parametric tests and much larger sample sizes for Likert scale analysis in HCI [39], Norman argues to the contrary that "parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of coming to the 'wrong conclusion' " [33], and recommends paired t-tests for comparing conditions via Likert scales for sample sizes of at least 5.

The average effects of the machine confidence visualisation conditions on the flow index are summarised in table 3. The Truthful visualisation condition was on average more flow-inducing than the Deceptive condition, with an effect size of approximately 1/2 out of a scale of 5, significant at the 99% level, and the Deceptive condition averaged 1/4 point lower than the Absent condition, significant at 95%.

Table 3: effect of visualisation

	Absent *vs.* Deceptive	Truthful *vs.* Absent	Truthful *vs.* Deceptive
upper 95%	0.49	0.62	0.83
mean	**0.26***	**0.16**	**0.42****
lower 95%	0.03	-0.28	0.03

The statistical tests described above allow us to make inferences about underlying effects in the presence of random noise, such as the variable behaviour of the algorithmic improviser. They are not, however, designed to allow generalisations to people beyond the study participants. In any research, the only reliable way to make quantitative generalisations to people beyond the study group is to recruit participants by randomly sampling from the entire population of interest. However, qualitative judgements regarding probable transferability of study results can often be argued from the diversity of the study participants [15]. In our study we did not have any particular population in mind, though we hope the results may be broadly indicative of the type and diversity of reactions that experienced improvising musicians would have had if included. We employed convenience sampling in recruiting instrumentalists to evaluate the system, and even this small group had quite diverse range of approaches and responses. As discussed in §5 the instrumentalists covered a broad range of musical styles, lending some confidence that our results may have relevance outside of the study group.

5 LISTENER EVALUATION

The experience study showed that the participants' tendency to achieve flow was enhanced by extra-musical communication of machine confidence. But what about the musical output? To see if effects observed through changes in the confidence visualisation conditions were perceivable only to a performer, or were also noticeable by external listeners, we conducted an additional on-line listener study.

Method

One of the authors, with tertiary qualifications and professional experience as an improvising saxophonist, participated in six improvisation sessions with the improvising machine to produce sixteen tracks. Although the use of an author-participant has the potential to introduce unintended bias, they had greater experience playing with the system over participants used in the performer study and their improvisation sessions were selected to best highlight the effect of the differing communication conditions. Prior research

suggests that layperson comparative evaluations of computer generated music are sharpest once the musical output has reached a reasonable level of mainstream musical plausibility [24, 44].

Randomised conditions and sampling were used to reduce any possible effects of unintended bias. Truthful and deceptive visualisation conditions were alternated between randomly throughout the improvisation sessions, such that the participant did not know which condition was in use. Six tracks were selected from the sixteen recordings by random stratified-sampling to balance the number of Truthful and Deceptive conditions and used for the listener study. The first minute of each of these tracks was paired into a series of A/B comparisons that were embedded into a web questionnaire. Audio files used in the questionnaire can be heard at https://dx.doi.org/10.6084/m9.figshare.7552235.v1

We used the Amazon Mechanical Turk platform to recruit 100 participants, who were asked to answer two questions for each of three A/B comparisons: 'Which performance was more interesting?' and 'Which performance had a better musical balance between drums and saxophone?'. Each participant was given $1 USD for participating.

Participants were not questioned on their musical interest or ability, but questions were included in the survey to identify participants who may have not understood the questions. An additional A/B comparison was added with one track not containing drums. This comparison also had an additional question asking which of the two recordings contained drums. We excluded participants who found a better balance in the track with only saxophone, stated the wrong track contained drums was excluded, or spent less than the 8 minutes required to listen to all tracks on the questionnaire. Of 100 participants, 96 met these requirements and their feedback was used for evaluation.

Results and Discussion

The results indicated a modest but significant tendency for the music produced in the Truthful visualisation condition to be perceived as more musically balanced than the Deceptive condition. Significance here is to be understood with respect to the number of participants (not the number of musical examples). A null hypothesis of random choice for the more musically balanced track is rejected by a test against the binomial distribution at the 95% for these particular musical examples. This suggests a noticeable effect for external listeners.

With on-line listener studies there are a number of environmental conditions that are not controlled for, that can have a significant influence on the experience of listening to music. Volume, speaker quality, speaker type and background noise can vary for each participant. By framing the questions as a series of A/B comparisons, most of these factors would likely stay consistent for compared tracks.

Table 4: Results of each A/B comparison in the on-line listener evaluation questionnaire.

	Truthful Condition	
Tracks	More interesting	Better musical balance
A *vs.* B	44%	51%
C *vs.* D	67%	65%
E *vs.* F	57%	60%
Total	**53%**	**55%***

Because live performance is a critical part of demonstrating improvisation, future evaluation of the machine improviser and effects of extra-musical communication is intended within live performance contexts.

6 CONCLUSION

As interaction with Artificial Intelligences – and in particular creative improvisers – becomes more commonplace, *how* we interact and collaborate with co-creational AI systems is an increasingly important area of research. In this paper we have investigated how extra-musical cues between human and AI improvisers can affect the achievement of flow states in an improvising duet. Our results demonstrate that, at least for the performers evaluated, communication of a confidence metric improves the human performer's ability to achieve flow states more readily. Additionally, we demonstrated that the resulting improvisational performances are more likely to be perceived positively by non-expert audiences in terms of musical balance between instruements.

These results support our hypothesis that the extra-musical communication of 'internal state' ***of* the machine *to* the human** can facilitate more engaging human-machine musical interactions.

Our experimentation with the use of biometric data (SC) as a proxy for the performer's musical engagement that could be communicated to the AI did not improve the performer's experience of musical flow, despite prior research showing this to be the case with pairs of improvising human musicians [11]. A difference between this previous study and our own was the location of the sensor: the Empatica E4 watch used in our experiments can only be worn on the wrist, whereas the previous study used a sensor attached to the ankle of the performer, to minimise inconsistencies due to movement which are common when improvising during performance.

That extra-musical communication of 'internal state'***of* the human musician *to* the machine** would also facilitate more engaging human-machine musical interactions was not supported in this study but is also not ruled-out. In further

research we aim to draw on other kinds of biometric sensors that are more resilient to the effects of sudden moment that is a necessary part of physical playing and explore other modes of extra-musical communication.

Conceptualising intelligent machines as creative partners rather than passive tools or instruments is relatively new. While we have a rich and well explored history of improvisation between human performers to draw upon, improvising with an alien, non-human yet active participant creates many exciting new possibilities for human-machine partnerships. The challenge, which we have begun to explore in this paper, is to maximise the creative benefits and possibilities for both performers and audiences.

REFERENCES

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[2] Margaret A. Boden. 2010. *Creativity and Art: Three Roads to Surprise.* Oxford University Press.

[3] Oliver Bown and Jon McCormack. 2009. Creative Agency: A Clearer Goal for Artificial Life in the Arts. In *ECAL (2) (Lecture Notes in Computer Science)*, George Kampis, István Karsai, and Eörs Szathmáry (Eds.), Vol. 5778. Springer, 254–261.

[4] Mason Bretan, Guy Hoffman, and Gil Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78 (2015), 1–16.

[5] Peter Mason Bretan. 2017. *Towards An Embodied Musical Mind: Generative Algorithms for Robotic Musicians.* Ph.D. Dissertation. Georgia Institute of Technology.

[6] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. 2017. Deep learning techniques for music generation-a survey. *arXiv preprint arXiv:1709.01620* (2017).

[7] Mihaly Csikszentmihalyi. 1990. Flow: The psychology of optimal performance.

[8] Mihaly Csikszentmihalyi. 1996. Flow and the psychology of discovery and invention. *New York: Harper Collins* (1996).

[9] Geoff Cumming. 2012. *Understanding the new statistics : effect sizes, confidence intervals, and meta-analysis.* Routledge, New York.

[10] Örjan de Manzano, Töres Theorell, László Harmat, and Fredrik Ullén. 2010. The psychophysiology of flow during piano playing. *Emotion* 10, 3 (2010), 301–311.

[11] Roger T. Dean and Freya Bailes. 2015. Using time series analysis to evaluate skin conductance during movement in piano improvisation. *Psychology of Music* 43, 1 (2015), 3–23.

[12] Mark d'Inverno and Jon McCormack. 2015. Heroic vs Collaborative AI for the Arts. In *Proceedings of IJCAI 2015.*

[13] Douglas Eck and Juergen Schmidhuber. 2002. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale* 103 (2002).

[14] K.A. Ericsson, R. Krampe, and C. Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 100 (1993), 363–406.

[15] B. Flyvbjerg. 2006. Five Misunderstandings About Case-Study Research. *Qualitative Inquiry* 12, 2 (April 2006), 219–245.

[16] Toby Gifford, Shelly Knotts, Stefano Kalonaris, and Jon McCormack. 2017. Evaluating Improvisational Interfaces. In *ICW2017: Proceedings of the Improvisational Creativity Worskshop*, Toby Gifford, Shelly Knotts, and Jon McCormack (Eds.).

[17] M. Gladwell. 2008. *Outliers, the Story of Success.* Allen Lane.

[18] Robert Hamilton. 2006. Bioinformatic feedback: performer bio-data as a driver for real-time composition. In *Proceedings of the 2006 conference on New interfaces for musical expression.* IRCAMâĂŤCentre Pompidou, 338–341.

[19] Guy Hoffman and Gil Weinberg. 2011. Interactive improvisation with a robotic marimba player. *Autonomous Robots* 31, 2 (Oct 2011), 133–153.

[20] P Hutchings. 2017. Talking Drums: Generating drum grooves with neural networks. *arXiv preprint arXiv:1706.09558* (2017).

[21] Elina Hytönen-Ng. 2016. *Experiencing'flow'in jazz performance.* Routledge.

[22] Helen Jackson. 2002. *Creative Evolutionary Systems.* Academic Press, London, Chapter Toward a Symbiotic Coevolutionary Approach to Architecture, 299–313.

[23] Duncan Jana and E. West Richard. 2018. Conceptualizing group flow: A framework. *Educational Research and Reviews* 13, 1 (Jan. 2018), 1–11.

[24] Anna Katerina Jordanous. 2013. *Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application.* Ph.D. Dissertation. University of Sussex.

[25] Sunjung Kim and Elisabeth André. 2004. Composing Affective Music with a Generate and Sense Approach.. In *FLAIRS Conference.* 38–43.

[26] Anne Landhäußer and Johannes Keller. 2012. Flow and its affective, cognitive, and performance-related consequences. In *Advances in flow research.* Springer, 65–85.

[27] J. D. Lee and N. Moray. 1994. Trust, self-confidence, and opera- tors' adaptation to automation. *International Journal of Human-Computer Studies* 40 (1994), 153–184.

[28] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80.

[29] Dimos Makris, Maximos Kaliakatsos-Papakostas, and Katia Lida Kermanidis. 2018. DeepDrum: An Adaptive Conditional Neural Network. *arXiv preprint arXiv:1809.06127* (2018).

[30] Jon McCormack and Mark d'Inverno. 2016. Designing Improvisational Interfaces. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016)*, François Pachet, Amílcar Cardoso, V. Corruble, and F. Ghedini (Eds.). 98–105.

[31] B. M. Muir. 1987. Trust between humans and machines, and the design of decision aides. *International Journal of Machine Studies* 27 (1987), 527–539.

[32] Jeanne Nakamura and Scott Roberts. 2016. The Hypo-egoic Component of Flow. In *The Oxford Handbook of Hypo-egoic Phenomena.* Oxford University Press, 133.

[33] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15, 5 (Dec. 2010), 625–632.

[34] François Pachet. 2006. 19 Enhancing individual creativity with interactive musical reflexive systems. *Musical creativity* (2006), 359.

[35] François Pachet. 2012. Musical Virtuosity and Creativity. In *Computers and Creativity*, Jon McCormack and Mark d'Inverno (Eds.). Springer, Berlin; Heidelberg, Chapter 5, 115–146. https://doi.org/10.1007/978-3-642-31727-9

[36] Prashanth Thattai Ravikumar. 2017. Notational Communication with Co-creative Systems: Studying Improvements to Musical Coordination. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition - C&C '17.* ACM Press, Singapore, Singapore, 518–523.

[37] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *IEEE International Conference on Computer Vision (ICCV).*

[38] Alison Robb and Matthew Davies. 2015. 'Being Inside the Story': a Phenomenology of Onstage Experience and the Implications of Flow. *About Performance* 13 (2015), 45–67.

[39] Judy Robertson. 2012. Likert-type scales, statistical methods, and effect sizes. *Commun. ACM* 55, 5 (May 2012), 6.

[40] R. Keith Sawyer. 2000. Improvisation and the Creative Process: Dewey, Collingwood, and the Aesthetics of Spontaneity. *The Journal of Aesthetics and Art Criticism* 58, 2 (2000), 149–161.

[41] R Keith Sawyer. 2003. *Group creativity: Music, theatre, collaboration*. Erlbaum, Mahwah, NJ.

[42] Frederick A. Seddon. 2005. Modes of communication during jazz improvisation. *British Journal of Music Education* 22, 1 (March 2005), 47–61.

[43] Milija Simlesa, Jerome Guegan, Edouard Blanchard, Franck Tarpin-Bernard, and Stephanie Buisine. 2018. The Flow Engine Framework: A Cognitive Model of Optimal Human Experience. *EuropeâĂŹs Journal of Psychology* 14, 1 (2018).

[44] Dan Stowell, Andrew Robertson, Nick Bryan-Kinns, and Mark D Plumbley. 2009. Evaluation of live human–computer music-making: Quantitative and qualitative approaches. *International journal of human-computer studies* 67, 11 (2009), 960–975.

[45] Ellen Waterman. 2015. Improvised trust: Opening Statements. In *The Improvisation Studies Reader: Spontaneous Acts*, Rebecca Caines and Ajay Heble (Eds.). Routledge, Oxon, UK, Chapter 8, 59–62.

[46] Craig Jonathan Webb. 2014. Parallel computation techniques for virtual acoustics and physical modelling synthesis. (2014).

[47] Matthew Wright. 2005. Open Sound Control: an enabling technology for musical networking. *Organised Sound* 10, 3 (2005), 193–200.

[48] William J. Wrigley and Stephen B. Emmerson. 2013. The experience of the flow state in live music performance. *Psychology of Music* 41, 3 (2013), 292–305.

[49] Stephen T Ziliak. 2008. Guinnessometrics: The Economic Foundation of Student's t. *Journal of Economic Perspectives* 22, 4 (2008), 199–216.