Proceedings of the Seventeenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2021)

Effects of Deep Neural Networks on the Perceived Creative Autonomy of a Generative Musical System

Jason Smith, Jason Freeman
Georgia Tech Center for Music Technology
840 McMillan St NW
Atlanta, Georgia 30318
jsmith775@gatech.edu

Abstract

Collaborative AI agents allow for human-computer collaboration in interactive software. In creative spaces such as musical performance, they are able to exhibit creative autonomy through independent actions and decision-making. These systems, called co-creative systems, autonomously control some aspects of the creative process while a human musician manages others. When users perceive a co-creative system to be more autonomous, they may be willing to cede more creative control to it, leading to an experience that users may find more expressive and engaging.

This paper describes the design and implementation of a co-creative musical system that captures gestural motion and uses that motion to filter pre-existing audio content. The system hosts two neural network architectures, enabling comparison of their use as a collaborative musical agent. This paper also presents a preliminary study in which subjects recorded short musical performances using this software while alternating between deep and shallow models. The analysis includes a comparison of users' perceptions of the two models' creative roles and the models' impact on the subjects' sense of self-expression.

Introduction

Deep learning has led to great improvements in computer vision and artificial intelligence, including the creation of collaborative agents (Szeliski 2010; Russell and Norvig 2009). These improvements in artificial agents have led to research in real-time systems that communicate with users to generate creative outputs, known as collaborative systems (McCormack et al. 2020; Elboushaki et al. 2020). Other artificial intelligence systems exhibit creativity through "emotional" or "unexpected" artistic generation, as perceived by its users (Salevati and DiPaola 2015). The ability for a collaborative agent to independently apply creative decisions is referred to as "creative autonomy" (Jennings 2010). This framework has been used in the discussion and evaluation of other collaborative systems (Augello et al. 2013; López-Ortega 2013),

In the field of music technology, deep learning is used in the creation of instruments with advanced gestural recognition controls, generative systems, and collaborative artificial intelligence agents that employ creative autonomy (Caramiaux and Tanaka 2013; Verdugo et al. 2020; Saffiotti et al. 2020). When combined with approaches in computational modeling, artificial intelligence can be used to quantify and support creativity (Boden 1998, 2009). Other dynamics of interaction between human and AI collaborators have been examined in the domains of pretend play and interactive visual art (Davis et al. 2017; Andrews 2019).

Applications that take advantage of deep learning for live music generation can either be characterized as an instrument that uses a machine learning model trained on inputs to create a musical result or as collaborative artificial intelligence agents that compose in collaboration with the user (Hantrakul and Kondak 2018; Xia 2016). These form the basis for Machine Musicianship, in which music theory and performance technique are used in the modeling of artificial intelligence behavior (Rowe 2001). Machine Musicianship has since given rise to the study of algorithmic improvisation and musical collaboration with improvisational agents (Fremont 2019; Brown 2018). Artificial intelligence can also shift between the roles of "tool" and an "actor" when used in music creation (Caramiaux and Donnarumma 2020). The systems using the role of a tool primarily respond to a musician's initiative and may function similarly to instruments. Other systems that take the roles of actors display their own initiative. They can be considered to employ a greater deal of creative autonomy, the independent application of creative decisions (Jennings 2010), in musical composition or performance. Communication of the state of the collaboration, both from a user to an AI and from the AI to the user, allows for AI systems to act in engaging human-machine improvisation (McCormack et al. 2019).

The research detailed in this paper aims to analyze AI's role in the musical process by measuring the effects of deep learning in human-agent musical interaction, focusing on answering two questions:

- To what extent can the inclusion of a deep neural network in a generative musical system affect its user's perception of it as a collaborative agent that exhibits creative autonomy rather than an instrument responding to their controls? How does this compare to a more shallow neural network in the same system?

- To what extent does a user's perception of the musical system's autonomy affect the level of control the user retains and gives up in the creative process, as well as on the user's self-reported expressive ability with the system?

To answer these questions, we have developed a system that generates music collaboratively with a user. The user collaborates by performing gestures in front of a camera, and the system collaborates by mapping their motions to musical parameters. It can be used with various neural networks architectures interchangeably, similarly to software instruments containing models and processes for live machine learning (Fiebrink, Trueman, and Cook 2009).

We have designed an exploratory experiment to compare users' experiences and perceptions of this system's autonomy across two models of contrasting complexity. Subjects performed with the system, alternating between deep and shallow neural networks. They rated their perceived autonomy of different versions of the system. They also rated the level of creative control they gave to the two versions, and reflected on their experiences after the performance.

Related Work

Designs for interactive music applications using gestural control have included sequential neural networks to detect drawings on a hardware peripheral, mapping motion data directly using hardware sensors, or using bodily movement sensors (Hantrakul and Kondak 2018; Ilsar and Hughes 2020; Verdugo et al. 2020). Our work only uses direct video feed and no external sensors, in order to maintain a portable and shareable system that also easily communicates the nature of its machine learning models to users.

Analysis of error in gestural musical interfaces leads to the conclusion that a system with increased error lowers the perception of the system's accuracy and responsiveness (Brown, Nash, and Mitchell 2020). Instead of analyzing a difference between the deep and shallow neural networks in terms of gestural recognition error, subjects in this study compare deep and shallow networks with similar performances. This paper focuses on deep learning's impact on a user's perceived autonomy when they do not have a clear understanding of the model's performance, unlike other works measuring performance increases in state-of-the-art models.

Previous research into creative autonomy in an interactive music system studied a system's ability to act autonomously in a musical setting without domain knowledge (Paolizzo and Johnson 2020). Our work expands on this concept through a user study comparing models of different depths, and the effects of model complexity on perception of autonomy. We also examine the effect that autonomy has on the relationship between a user and autonomous agent.

Previous human-AI collaboration research has measured a robot's ability to follow a human performer and also notes the feedback loop in which the human naturally adjusts to the model (Saffiotti et al. 2020). Our research adds to this by quantifying the relationship between the human and AI partners in terms of the perceived increase in creative control afforded to a deeper model, as well as by relating it to the level of recognition the user gives the model as an agent.

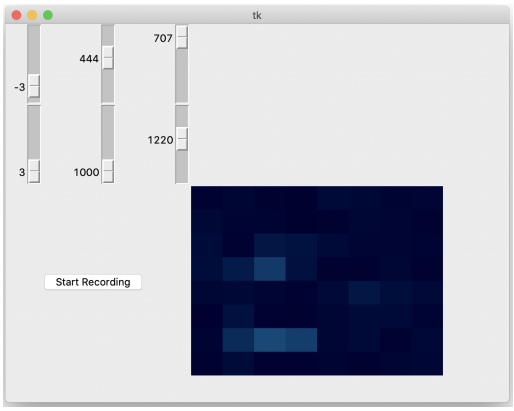

Figure 1: User view of the training version of the application, with sliders controlling the signal parameters (top), a recording button (left), and the downsampled camera input (right). The parameters include gain (left), center frequency (middle), and bandwidth (right) for two audio signals.

Other related work in co-creative systems discusses the behavior of human and AI collaborators during the process of improvisation (Magerko and Long 2020). Prior research to which we have contributed, focused on collaboration and creativity in the online music-making platform EarSketch, investigates how users engage with AI and human collaborators while formulating musical ideas (Sarwate, Tsuchiya, and Freeman 2018; Truesdell et al. 2021). Prior work into the effect of interface constraints on musical improvisation provides a comparison point for this study's use of intentionally constrained devices (Tez and Bryan-Kinns 2017). Other work in co-creative musical interaction dynamics explores how pairs of human collaborators respond to a third, AI collaborator while composing music (Suh et al. 2021).

System Design

Generative Music System

This generative musical system is designed to host neural networks that map motion to parameters controlling sounds. It receives a stream of video data and plays a stream of audio data simultaneously, and it synchronizes and records both to create a dataset for model training.

Sound is generated through looping of two ambient sound files. Ambient sound files were used in this study to prevent participants from confusing changes caused by their gestures with any rhythmic changes in the samples themselves. They also allow for the looping of an audio stream with minimal processing power, but with more complex and interesting timbres than raw waveforms from simple synthesizers. Data is recorded by manually setting audio parameters through sliders on the program interface and performing a repetitive motion while the "Start Recording" button has been activated. Figure 1 depicts the interface when training models.

When users are performing with the trained model, the system maps motion patterns to controlling gain, center frequency, and bandwidth of bandpass filters on two ambient, looping audio signals. To ensure that users base their ob-

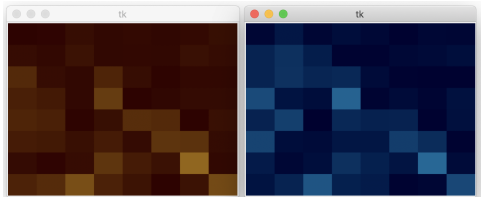

Figure 2: User view of sliderless testing versions of the application, with red (left) and blue (right) windows hosting the shallow and deep neural networks respectively.

servations of the model's behavior only on musical output, the amount of visual information identifying differences in mappings between the two models is minimized. The sliders themselves are hidden from the user when testing in order to obfuscate the mappings between motions and parameters. When in testing mode (Figure 2), the interface color changes depending on the model.

The software system was written in Python and developed using a Tkinter [1] interface. The camera readings of the laptop hosting the application are processed using OpenCV in Python [2]. The program downsamples camera images to an 8x8 grid. It stores the difference between successive frames to represent frame-by-frame motion as arrays for both data collection and visual representation. Downsampling is used to drastically reduce training time, standardize across different users and cameras, and to ensure that models generalize motion patterns to windowed regions of the screen.

Two ambient sound files, drones of varying frequencies, are loaded into asynchronous Pyaudio [3] streams. The parameters of the bandpass filters are updated and applied to the signal at each frame. Rather than storing audio data, this program stores the filter parameter values as they are represented in the GUI, creating a training set of parameters (1x6 arrays representing gain, center frequency, and bandwidth for two bandpass filters) that are synchronized with frame differences (8x8 arrays).

Neural Networks

Both models used in this experiment were Pytorch[4] neural networks. The shallow network uses a single 64x6 linear layer to predict 6 regression values (one for each signal's gain, center frequency, and bandwidth) given an 8x8 image. The deep network uses two two-dimensional Convolutional layers with 3x3 kernels and 10 and 20 filters alongside two linear layers to generate the same 6 regression values. Both models use dropout, ReLU, and identical hyperparameters with an Adam optimizer and a learning rate of 0.001. Additionally, both models were trained over 30 epochs, a batch size of 100 input-label combinations, and no GPU. As a result, the linear model trained over 11.511 seconds and the CNN trained over 28.140 seconds to achieve similar levels of Mean Square Error loss.

[1]https://docs.python.org/3/library/tkinter.html

[2]https://pypi.org/project/opencv-python/

[3]https://pypi.org/project/PyAudio/

[4]https://pytorch.org/

The models were trained on the same dataset of motion-parameter combinations, using a variety of gestures and distances from the camera. The motion dataset consisted of input gestures, performed by one of the authors, that corresponded to parameters set using the manual training version of the software (see Figure 1). Training data was collected through sessions of repeatedly performing a motion while changing a parameter, in order to create associations between variations in that motion and parameter changes. These were then combined as a dataset to train the models on multiple combined gestures. For example, one session included repeated circular motions while adjusting values of a gain slider to match the motions' size and speed. Differences between camera frames are captured and are used as input into both models to perform motion recognition with a point of data. Because the dataset pairs frame differences with single values for each parameter, inputting a motion over multiple frames causes the system to generate a series of parameter values over time. Through 17 data collection sessions, with an average length of 46 seconds, a dataset of 9423 samples (pairs of 8x8 arrays of camera frame-differences and 1x6 arrays of audio parameters) was created. 70% of these samples were used for model training, with the other 30% reserved for validation. This data generation and training was performed during development of the system, and before the user studies.

A convolutional neural network was chosen as opposed to sequential models such as recurrent neural networks, which are popular in music generation. This was done to compare two models that operate on the same temporal component of a single frame of input, rather than a series of frames. Also, features learned by a convolutional neural network are easier to extract and provide an interpretable framework that is more comparable to the mappings learned by a linear model (Selvaraju et al. 2017). Additionally, a CNN was used in order to compare a simple deep model to the most shallow neural network possible, despite the superior performance of RNNs. Because both a CNN and linear network are feed-forward networks that operate on a point of data rather than a sequence, the training and output pipelines were identical.

Methodology

Subjects were required to have a minimum basic understanding of music and computing concepts, as well as awareness of machine learning. These criteria ensured that they were able to understand what a collaborative musical agent is and to make informed commentary on their experiences with the two models. Subjects completed a pre-questionnaire, guided performance task, post-questionnaire, and interview as part of the one-hour experiment.

Pre-Questionnaire

Subjects began the study by answering a questionnaire regarding musical experience and preferences. They ranked their experience with live electronic music as well as their experience with and self-reported desire to perform music with artificial musicians by answering the following questions on a 7-point scale:

- I have listened/been exposed to live electronic music.
- I have interacted with live electronic music tools.
- I have performed live electronic music.
- I have performed live electronic music collaboratively.
- I am interested in performing live electronic music with an artificial agent collaboratively.

The survey also included questions used to establish the subject's familiarity with electronic music and machine learning concepts in order to characterize their prior experience. Experience with electronic instruments may influence them to prefer instrument-like behavior, or their machine learning experience may predispose them to show more interest in a more autonomous model. They reported their experience with machine learning through the following questions, derived from the Bloom's Taxonomy of Educational Objectives (Bloom et al. 1956), on a 7-point scale:

- I feel confident in my ability to remember **machine learning** concepts from my computer science education.
- I feel confident in my ability to describe **machine learning** concepts.
- I feel confident in my ability to compare **machine learning** models.
- I feel confident in my ability to evaluate a **machine learning** model.
- I feel confident in my ability to create a **machine learning** model.
- I am familiar with the use of **machine learning** in a live/interactive music context.

By answering the same questions for machine learning and deep learning, subjects reported their ability to remember, describe, compare, evaluate, and create machine and deep learning models respectively. They also reported their familiarity with the use of machine learning and deep learning in a musical context.

Following the survey questions, participants rated their musical tastes according to the genre categories of the Short Test of Music Preferences (STOMP) to provide a reference point for generalizing their musical preferences (Rentfrow and Gosling 2003). Subjects ranked their preferences among 15 genres, and their preferences were mapped to four dimensions: *Reflective & Complex*, *Intense & Rebellious*, *Upbeat & Conventional*, and *Energetic & Rhythmic*.

Performance

Subjects then completed a performance with a deep neural network-embedded version and a shallow neural network-embedded version of the interactive music software. Subjects were randomly selected to start with either the deep or shallow model, and rehearsed for 5 minutes with each.

During this time, they were encouraged to freely ask questions about the system and to describe their observations and impressions of the models. They then recorded 2-minute videos for both systems, containing the system's pixelated video capture and musical output without identifying the user. Subjects were not explicitly given compositional plans for these recordings, but were encouraged to form their own.

Post-Questionnaire

Following the performance task, subjects were guided through an analysis of the recorded audio and video to identify trends in model output. This was performed separately for the deep and shallow versions. Subjects identified start and end times for any number of notable musical interactions within the 2-minute performance period. They then rated the system in that window on two criteria: *autonomy*, or the amount to which the system felt like an autonomous agent as opposed to an instrument, and *control*, the amount of control that the system had over the current musical state through influencing the gestures and actions of the participant. For example, one subject reported a single musical interaction as a series of circular gestures they performed between *00:30* and *00:46* of their trial. They felt that they were "beginning to take control" of the system, so they rated it as a *3* out of 7 for control. They also felt that the system was generating musical output that "is based on what [they were] doing, but different that what [they] expected", so they gave it a high autonomy rating of *6* out of 7. These ratings are designed to evaluate a user's relationship with the system as a co-creator. They are derived from models of creative autonomy (Jennings 2010) and the interactions between musicians and AI agents. Each subject recorded between four and eight musical interactions for both of their trials. The subjects rated their musical output for the two versions on a scale defined by the Short Test of Music Preference's four music preference dimensions (see Post-Questionnaire), so that their original preferences in the pre-questionnaire can be compared to their perception of each version of the system's output (Rentfrow and Gosling 2003).

The remainder of the questionnaire included a usability and satisfaction questionnaire in the form of a modified version of the Creativity Support Index (CSI) for a collaborative musical agent (Cherry and Latulipe 2014). The modified creativity support questionnaire, listing the individual scales of the CSI, is one of the usage scenarios listed in (Cherry and Latulipe 2014). These questions were designed to gauge user experience relative to their perceived autonomy of the system, and subjects answered a set of questions for the deep and shallow versions of the system separately. Individual scores are recorded, rather than aggregating into a single metric, in order to evaluate and compare the two systems' enabling of expressive ability separately from user enjoyment and engagement. The comparison of CSI scale ratings is used alongside creative autonomy ratings as a measure of the effect that the inclusion of deep learning has on a user's perception of the system's creative autonomy. It is then used to measure whether or not it indicates increased satisfaction and self-reported expressive ability with the system.

Finally, subjects were asked a series of open-ended questions. For example, the question "Did this trial change your perception about performing with an artificial agent collaboratively?" is used for comparison against their originally-reported perception in the pre-questionnaire. The subjects were also able to freely comment on how they would approach future interactions with similarly unknown musical systems. Subjects were asked to verbally identify the kind of music they were trying to make and whether or not it

aligned with their musical preferences. Subjects with a well-defined compositional plan for their 2-minute performances commented on whether or not their musical ideas changed over time. Subjects who did experience changes attributed them to increased experience with the model output. Their time experimenting with the system led them to new ideas or to changes in their perceptions about the predictability of the output. While reflecting on their performances, subjects also reported any surprising decisions made by the model that changed their course of action and how their experience differed between the deep and shallow models.

Results

The five subjects for this experiment were selected from a pool of graduate-level Georgia Tech Music Technology students. These students (labeled A, B, C, D, and E) came from a variety of musical backgrounds and experience levels regarding computer science and machine learning concepts.

Discovery of Musical and Machine Learning Properties

Each subject commented on their impressions of the system as they practiced with it during the initial practice phase. For example, Subject E said "this one feels less responsive" when transferring from Blue [deep] to Red [shallow].

Each subject began by testing the limits of the system through a variety of extreme hand gestures before attempting to use smaller gestures. Some used the interface's visual indications of motion to determine that only motion triggered changes in the filter parameters, while others attempted to stand still in different positions to confirm it had no effect on the sounds. Additionally, each subject correctly identified that the system was manipulating filters on existing audio samples, by remarking aloud during the initial 5-minute practice period. Other common occurrences among the subjects were the use of traditional conducting patterns, inanimate objects, and beginning the practice for their second model with patterns they used in the first.

Perceptions of Creative Autonomy and Musical Output

Because both models affected the same set of musical parameters on the audio signals, the sounds the subjects were able to make with the two systems were largely the same. Additionally, they were trained on the same data and reached similar convergences. This was an intended result of the system design, as subjects instead observed differences in how the two models responded to subtle motion to achieve the same dimensions of musical variety.

As shown in Table 1, subjects rated the two models in their abilities as creativity support tools in a modified Creativity Support Index (Cherry and Latulipe 2014). They reported higher ability to collaborate and be expressive using the deep version, while they rated the musical quality of the two systems' outputs identically. Additionally, subjects reported higher enjoyment and willingness to re-use the software for the deep version.

Question	Shallow Avg. Score	Deep Avg. Score
Collaboration	3.8	**4.2**
Enjoyment	4.2	**4.6**
Expression	3.8	**4.2**
Would Use Again	3.2	**3.4**
High Standard of Output	3.6	3.6
Output Resembles a Human	2.6	2.6

Table 1: Creativity Support Index questionnaire scores, with average responses between subjects for the shallow and deep models on a 7-point scale.

Three of the five subjects reported that the deep neural network version felt "more responsive", with subject D commenting that "as long as I do something, it will do something in return". Subject C, with a strong preference for the deep version (see Table 2), stated that "Both versions brought me through the *narrative* of starting to work with it as an instrument, then getting confused when it acted unpredictably, then coming to recognize it as a collaborative partner. The blue [deep] version brought me through it in a much shorter time, so I was able to have more time to enjoy it". Conversely, subject B reported that "Red [shallow] felt more in control. Blue [deep] felt more responsive when practicing, but not when trying to do something specific", preferring the shallow version when enacting their compositional plans.

A set of music preference dimensions was calculated using each subject's answers to the Short Test of Music Preferences questions in the pre-questionnaire (see Pre-Questionnaire). In the post-questionnaire, subjects rated the output they were able to create on the same music preference dimensions for the deep and shallow models separately. For the deep model, subjects' scores differed from their initial musical preferences at an average of *1.92* out of 7 over the four dimensions. Scores for the shallow model differed by an average of *2.17* out of 7. This suggests that users were able to create music more akin to their tastes using the deep model, despite the two being the same in musical quality and audio content.

Subject D remarked that inconsistencies in model performance caused them to see the system as less autonomous and more random. This was exacerbated by a lack of depth in the shallow model, which to them was "unsuccessful at capturing beat-based movements". Both versions surprised the user with repeated use of the same gesture generating different audio parameters, but that the deep version had a "larger range of sounds" and was "more complex". Similarly, subject E reported that inconsistencies in the model output affected their confidence in mappings they had discovered while practicing. They stated that "Blue [deep] was better than red [shallow]" in this regard, but they experienced frustration when they "tried to produce contrast, but was limited except for going in and out of resting state".

Perceptions of Role in the Creative Process

Table 2 shows the average scores for control and autonomy ratings for each model, as each subject annotated between four and eight musical interactions for the shallow and deep versions respectively (see Post-Questionnaire). On average, subjects rated the deep version higher in both their perception of it as an autonomous agent and its ability to take control creatively. Participants rated the control of the deep model an average of *0.55* higher than the shallow model on a 7-point scale, with the autonomy rating *0.76* higher for the deep model as well.

Table 2 shows the average scores for autonomy and control for each subject, as well as their answer to the question "I was able to be expressive while using this system" in the Creativity Support Index post-questionnaire. Four out of the five subjects rated the deep learning version higher in terms of autonomy, and four out of the five stated that the deep learning version influenced their actions more, taking more control in the collaborative process. Additionally, three of the subjects reported higher expressive ability with the deeper model, with one preferring the shallow version.

Subject C, who rated the deep version much higher than the shallow version in terms of enabling expression, felt that the shallow version inhibited their ability to form a musical plan. Specifically, they intended to "start soft, have a first section, die down, then have a second section and climax", and they reported deviations from their original plan during the performance when discovering new mappings from the model. This user stated that they became "lost" while using the shallow version, and attempted to regain control by using familiar motions such as wide sweeps with their arm. They welcomed the lack of control in the deep version due to its responsiveness, stating that they "accepted its autonomy" and began to enjoy experimentation. They felt that alternating sequences of experimentation and relaxation allowed them to build a "different kind of control, trusting the system's autonomy and building a mutual relationship."

The two subjects that did not report a higher expressive ability with the deep learning version of the system cited a lack of confidence in the system's ability to interpret their musical goals. Subject A stated that human collaborators have "idiomatic conventions" to allow other performers to interpret their actions. The "black box" nature of this system allowed them to more easily follow their own compositional ideas with the less autonomous model. Even though they percieved the deep learning version as having more autonomy and afforded it more creative control, they enjoyed using the shallow version more and would be more likely to use it again. They stated that it was "more autonomous, but wasn't helping ME perform". Subject B had attempted to use traditional conducting as an input gesture, and felt that their ability to be expressive was limited due to the system's use of audio parameters on a constant audio loop. This subject stated "it feels more like someone mixing my music, rather than even collaboration like with another musician".

These relationships between subject ratings of system and their reported self-expression using the system indicate that the experience of a subject was largely determined by the nature of their compositional ideas while using the software.

	Autonomy		**Control**		**Expression**	
	S	**D**	**S**	**D**	**S**	**D**
A	3.4	**5.75**	3.2	**5.5**	**6**	2
B	5.0	**5.25**	4.83	**5.25**	3	3
C	4.375	**5.25**	**4.875**	4.125	1	**5**
D	**4.33**	4.2	3.16	**4.2**	5	**6**
E	3.25	**3.75**	3.25	**3.4**	4	**5**
Avg.	4.17	**4.93**	4.0	**4.55**	3.8	**4.2**

Table 2: Subjects' ratings for system autonomy, creative control, and user-reported expressive ability for the shallow (S) and deep (D) models respectively, on a 7-point scale.

The deep learning-enabled system improved the experience for subjects C, D, and E who all tested a variety of motions to find mappings, but came to the detriment of subjects A and B who were more reliant on traditional musical conventions.

Interest in Collaboration with AI

Subjects initially reported a high interest in performing using a collaborative agent, with an average of *6.2* out of *7* in their survey questionnaires about musical experience (see Pre-Questionnaire). When asked if the experience affected their interest, all subjects responded that it either stayed the same or increased, although many had new skepticism gained from the experience: Subject E stated that knowing details about the model's performance would affect their perception of it's quality and their "relationship with the app." Given sufficient time with a system, they would be able to balance consistent mappings with gestures resulting in machine-led parameter changes. Subject D stated that they would want to know the nature of future collaboration, as this experiment "didn't feel back and forth" due to the model manipulating audio parameters in response to gestures instead of simultaneously generating sound alongside the user. Lastly, subject B, who rated the deep version higher in terms of control and autonomy but not expressive ability, said that they would want a clear explanation of the parameters being manipulated. They spent more time learning how to use the deeper model, with less time to actualize a musical plan. As such, they would have appreciated more instruction or visual indicators in a future version of the system.

Discussion, Limitations, and Future Work

This research presents an examination of how the inclusion of a deep neural network in an interactive generative musical system might increase the user's sense of that system's creative autonomy and control. By comparing the experiences of users using the system outfitted with deep and shallow models, it also explores the relationship between the level of creative control a user gives to an agent and their perceived musical expression. Despite using identical audio samples and manipulating the same parameters, the deep model achieved higher ratings than the shallow model in the creative autonomy exhibited by the system and creative control it took. The users also had an increased sense

of expression and collaboration while interacting with the deep version of the system, but did not report a link between control and autonomy with expressive ability. These findings suggest that the perceived value of a system's autonomy is situationally dependent on the user's creative goals, and that a lack of understanding of the model's creative process can diminish a user's expressive ability. This study suggests an area for future work investigating the correlation between the interpretability of a model and users' experiences with and perceptions of their relationships with the models.

Sample size in this experiment was limited by unavailability of subjects due to the COVID-19 pandemic, with the subject pool taken exclusively from Georgia Tech Music Technology students. Future work can be generalized and increase statistical significance through a larger subject pool of more diverse perspectives and levels of self-reported familiarity with music and machine learning concepts. A larger subject pool would include non-expert musicians and general users without computer science knowledge.

Another limitation of this study was the lack of a third, random model. The decision to forego a control group was made due to limited time with each subject, factoring both subject availability constraints due to the COVID-19 pandemic and concern for subject fatigue with a third condition. Additionally, the display color for the two systems were not randomized between subjects. This may have influenced users' perceptions. Alternative visual representations, such as different shapes, will be used in future studies and accounted for in experimental design.

The example of subjects such as A and B (Table 2), who reported higher creativity support scores for the model with lower autonomy and control ratings, suggests that users perceiving a system's autonomy alone is not enough to enhance feelings of expression and collaboration on a per-user basis. These subjects experienced a negative effect on expressive ability due to a loss of control when they had strict compositional ideas, which manifested as a lack of confidence in the model. This indicates the need for collaborative musical systems to accommodate users with specific musical goals by following their directives. The system itself can be expanded to include audio generation so that it can better enable expression and act as a collaborative musician.

Additionally, after using the software, subjects were asked if their interest in collaborating with an AI musician had changed since the pre-questionnaire. Subjects were either more interested than at pre or had already been fully interested at pre (reporting a 7 out of 7 in the pre-questionnaire). However, users stated that they would appreciate more visual information to better understand the system's inner workings. Future work will include the addition of visual communication from the agent to the user, in order to ascertain their effects on perceived expressiveness and collaboration, and quality of interaction.

References

Andrews, C. 2019. An Embodied Approach to AI Art Collaboration. In *Proceedings of the 2019 on Creativity and Cognition*, 156–162.

Augello, A.; Infantino, I.; Pilato, G.; Rizzo, R.; and Vella, F. 2013. Introducing a creative process on a cognitive architecture. *Biologically Inspired Cognitive Architectures* 6: 131–139.

Bloom, B. S.; et al. 1956. Taxonomy of educational objectives. Vol. 1: Cognitive domain. *New York: McKay* 20: 24.

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial intelligence* 103(1-2): 347–356.

Boden, M. A. 2009. Computer models of creativity. *AI Magazine* 30(3): 23–23.

Brown, A. R. 2018. Creative improvisation with a reflexive musical bot. *Digital Creativity* 29(1): 5–18.

Brown, D.; Nash, C.; and Mitchell, T. J. 2020. Was that me? exploring the effects of error in gestural digital musical instruments. In *Proceedings of the 15th International Conference on Audio Mostly*, 168–174.

Caramiaux, B.; and Donnarumma, M. 2020. Artificial Intelligence in Music and Performance: A Subjective Art-Research Inquiry. *arXiv preprint arXiv:2007.15843* .

Caramiaux, B.; and Tanaka, A. 2013. Machine learning of musical gestures: Principles and review. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 513–518. Graduate School of Culture Technology, KAIST.

Cherry, E.; and Latulipe, C. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21(4): 1–25.

Davis, N.; Hsiao, C.-P.; Singh, K. Y.; Lin, B.; and Magerko, B. 2017. Creative sense-making: Quantifying interaction dynamics in co-creation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 356–366.

Elboushaki, A.; Hannane, R.; Afdel, K.; and Koutti, L. 2020. MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems with Applications* 139: 112829.

Fiebrink, R.; Trueman, D.; and Cook, P. R. 2009. A Meta-Instrument for Interactive, On-the-Fly Machine Learning. In *NIME*, 280–285.

Fremont, D. J. 2019. *Algorithmic Improvisation*. Ph.D. thesis, UC Berkeley.

Hantrakul, L.; and Kondak, Z. 2018. GestureRNN: A neural gesture system for the Roli Lightpad Block. In *NIME*, 132–137.

Ilsar, A.; and Hughes, M. 2020. A New Audio-Visual Gestural Instrument for Unlocking Creativity. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–4.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4): 489–501.

López-Ortega, O. 2013. Computer-assisted creativity: Emulation of cognitive processes on a multi-agent system. *Expert Systems with Applications* 40(9): 3459–3470.

Magerko, B.; and Long, D. 2020. Why Don't Computers Improvise With Us? *Extended abstract presented at the Workshop on Artificial Intelligence for HCI at the 2020 ACM Conference on Human Factors in Computing Systems* .

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 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.

McCormack, J.; Hutchings, P.; Gifford, T.; Yee-King, M.; Llano, M. T.; and d'Inverno, M. 2020. Design Considerations for Real-Time Collaboration with Creative Artificial Intelligence. *Organised Sound* 25(1): 41–52.

Paolizzo, F.; and Johnson, C. G. 2020. Creative autonomy in a simple interactive music system. *Journal of New Music Research* 49(2): 115–125.

Rentfrow, P. J.; and Gosling, S. D. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology* 84(6): 1236.

Rowe, R. 2001. *Machine musicianship*. MIT press.

Russell, S.; and Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press.

Saffiotti, A.; Fogel, P.; Knudsen, P.; de Miranda, L.; and Thörn, O. 2020. On human-AI collaboration in artistic performance. In *First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020*, 38–43. CEUR-WS.

Salevati, S.; and DiPaola, S. 2015. A creative artificial intelligence system to investigate user experience, affect, emotion and creativity. *Electronic Visualisation and the Arts (EVA 2015)* 140–147.

Sarwate, A.; Tsuchiya, T.; and Freeman, J. 2018. Collaborative coding with music: Two case studies with EarSketch. In *Proceedings of the Web Audio Conference*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Suh, M.; Youngblom, E.; Terry, M.; and Cai, C. J. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

Szeliski, R. 2010. *Computer vision: algorithms and applications*. Springer Science & Business Media.

Tez, H. E.; and Bryan-Kinns, N. 2017. Exploring the effect of interface constraints on live collaborative music improvisation. In *NIME*, 342–347.

Truesdell, E. J.; Smith, J. B.; Mathew, S.; Katuka, G. A.; Griffith, A.; McKlin, T.; Magerko, B.; Freeman, J.; and Boyer, K. E. 2021. Supporting Computational Music Remixing with a Co-Creative Learning Companion. In *Proceedings of the Twelfth International Conference on Computational Creativity*.

Verdugo, F.; Kokubu, S.; Wang, J.; and Wanderley, M. 2020. MappEMG: Supporting Musical Expression with Vibrotactile Feedback by Capturing Gestural Features through Electromyography. In *International Workshop on Haptic and Audio Interaction Design*.

Xia, G. 2016. Expressive Collaborative Music Performance via Machine Learning. *PhD Dissertation* .