6 Robotics

Fast and Curious: A CNN for Ethical Deep Learning Musical Generation

Richard Savery and Gil Weinberg

Introduction

Our work in robotic musicianship aims to facilitate meaningful and inspiring musical interactions between humans and robots. The motivation for our research is to discover how robotic collaborators can enhance and enrich musical experiences for humans. Robotics allows us to explore and achieve new musical possibilities, by combining computer generation with physical sound and embodied agents. In addition to rich acoustic sounds, robotic musicians can provide intuitive visual cues that improve musical interaction through expressive physical accompaniment to sonic generation. Our work is also driven by the artistic potential of mechanomorphic approaches, such as humanly impossible speed and precision, and the possibility to surprise and inspire human collaborators through artificial constructs and algorithms.

Over the last few years, developments in robotics and AI have led to new societal and ethical considerations that inform our work. Similar considerations have started to be explored in broader functional AI and robotic research, and we believe they have not been adequately addressed in creative work. In this paper, we examine a number of ethical considerations for the field of musical AI and robotic musicianship, where artificial intelligence and creativity are embodied in agents to create novel musical experiences. These ethical issues include bias and lack of diversity in data selection, prohibitive training requirements and subsequent environmental impacts, and the exclusion of artists without adequate computational resources. Other challenges include the impact of AI on human agency and employment, and the ownership of material and training data. Moreover, deep learning networks tend to have a distinct lack of transparency in system design, which decreases any chance of interpretability to developers, musicians, and listeners. We have considered these ethical consideration through our work in robotic musicianship and have developed new approaches to address ethical and societal concerns.

In the second half of the chapter, we present a new, human-focused deep learning system designed to address some of these ethical considerations by allowing participants agency over interaction and generation. The system is simple, easy to train, and allows for efficient, interactive real-time generation of musical

DOI: 10.4324/9780429356797-7

improvisations in performance with human musicians. It is comprised of a generative, convolutional neural network using a novel data format that appears to allow improved learning of nonlocal dependencies and repetitive structure across beats within musical phrases. We have observed that the system is able to learn to generate convincing and coherent improvisations from relatively small amounts of data. It can run effectively with limited computational resources, minimizing environmental impacts, and produces convincing musical interactions in a live performance setting.

Robotic Musicianship and Ethics

It is common for researchers in artificial intelligence and music to ignore the potential societal implications of their work (Briot et al., 2017). To address this lack of consideration, we have investigated our own work in robotic musicianship in an effort to identify potential ethical and societal challenges. In the passages that follow, we describe some of these challenges and our efforts to address and reconcile them.

Human Agency and Employment

One of the main recent societal concerns has been the replacement of human agency and employment by AI and robotics (Vochozka et al., 2018). In our own work on musical AI and robotic musicianship, we have constantly questioned whether our approaches might replace, rather than enhance, human musicians. To maximise enhancement over replacement potential, we centred our work on human-robot collaboration, rejecting project ideas that did not have a strong human presence in the loop. We design our robots to highlight their unique artificial advantages, such as novel, algorithmic-driven music generation and humanly impossibly mechanical abilities. Human collaborators, on their part, bring their unique human advantages to each interaction, such as emotion, expressivity, and creativity. Our ultimate goal is to facilitate musical experiences that would inspire and surprise human musicians, allowing them not only to explore new and exciting music but also to think about music in new ways. We believe that our human-centred design would not lend itself easily for replacing human agency and employment, not only by our team but also by others who may be building on our work in the future.

Bias, Diversity, and Accessibility

Significant concerns about bias and discrimination in AI and machine learning stem from inherited prejudices in dataset creation and selection as well as human algorithmic decisions (Gomez et al., 2018). It is difficult to dismantle such biases, as these systems' inner workings are often not transparent even to their creators (Barocas & Selbst, 2016). An added challenge to diversity in AI is the potential restriction of use and development of systems due to financial and technical

impositions. This might lead to broadening the gap between the have and the have-not, preventing a wide social adoption of AI and its benefits.

In our work in music and AI, we have explored several approaches to address these issues. For example, we have made an effort to collect and create datasets of works by underrepresented minorities addressing both gender and race, including datasets for music and lyrics in genres such as jazz and hip-hop. We also adapted and personalised our systems to allow for a wide range of users to participate in the interaction. Most of our systems can rapidly adapt to new datasets and allow for easy individual iteration when needed. Creating systems that are portable to a variety of operating systems is another effort we are making to allow users with limited hardware capabilities to engage with our systems. While our robots are not affordable for wide populations, we aim to include shareable software versions that can operate on many computer systems. We have also been as transparent as possible regarding the inner workings of our design, while acknowledging that some of the technical aspects might still be perceived as black boxes to our users, participants, and audiences.

Data Copyright and Consent

Copyright law around AI is rapidly evolving, addressing a variety of perspectives and stakeholders. For music generation, the key issues stem from potential ownership claims from music dataset creators, developers of the AI system, and potential users of the system. Sturm et al. (Sturm et al., 2019) address these copyright issues, current legal status, and the potential future legal implications, reaching the conclusion that a fundamental rethinking of these topics is needed. One of the main open questions in that regard is who owns the product of a creative AI system – the dataset creators, the system designers, the public, or maybe the machine itself? Leading AI companies such as OpenAI (a company with a mission statement built on AI benefiting all humanity) argue that IP should be free to use for AI, with training constituting fair use (O'Keefe et al., 2019).

In our own work, we have striven to receive explicit consent from the creators of the data we use. It is up for debate, however, whether in some cases, the goals of the system and the nature of creative development might not allow for receiving full consent from creators (Tinker & Coomber, 2004). We believe that consent should be asked for and received whenever possible. We have used this approach in multiple systems (Savery et al., 2021a, 2019b), where we manually created datasets. For some of our systems, which relied on extremely large datasets, we have not come up yet with a realistic way to ask and receive consent from all contributors.

Cultural Misuse

New developments in music and AI may be less susceptible to misuse by governments and corporations in comparison to technologies such as facial recognition

or behavioural data analytics. However, it is important to note that due to the strong cultural significance of arts and music, the unethical utilisation of AI in music might lead to serious societal consequences. Research in ethnomusicology offers many perspectives on approaches to ethical consideration of music as a cultural artefact (Shelemay, 2013). Philosopher Appiah extends that to say that the value of human life means 'valuing the practices and beliefs that lend them significance' (Appiah, 2008), such as music. The possibility and implication of devaluing a musical tradition has been explored by some research in AI and music, which, while subjective, is felt by many communities (Sturm et al., 2019).

In our own work, we have integrated culturally relevant datasets, such as an Australian Aboriginal language, with robotic voice (Savery et al., 2019a). These datasets were public domain and encouraged for use by the creator as a way to share the sound of the language. Even so, it is not clear that the creators of the dataset from the late nineties could predict this 'future use' case. Additionally, while the creator of the dataset gave permission, the language and substance of the dataset are a component and representation of the cultural identity of a larger population, which needs to be considered. Moreover, we recognise that our efforts to address bias and diversity by focusing on genres such as hip-hop and jazz should be done in collaboration with and consultation by members of the relevant communities to prevent cultural appropriation of these genres.

Music is a deeply personal medium central to the human experience, with implications beyond just commercial use. Clarke et al. demonstrate that even the act of passive listening to music can significantly change the cultural attitude of listeners (Clarke et al., 2015). We believe that it is crucial that future work in music and AI consider the outcome and possible influence of created systems.

Public Perception and Presentation

Musicians have been showing a wide range of responses to musical AI, some describing the integration of AI into music as a welcome collaborative development, while others address the combination as an existential threat (Knotts & Collins, 2020). In our own work, listeners have also questioned the use of AI as potentially reducing the essential quality of music, with some survey respondents providing quotes such as, 'It removes the inherent skill of a creator. To close ones eyes, and dig into one's own musical vocabulary, and come up with something original and tailor made' (Savery & Weinberg, 2018). We have also received informal feedback from audiences in our concerts and presentations, questioning a wide variety of topics – from whether our robots can indeed play in the style of humans to the possible ethical implication of even calling our robots 'musicians.' While it is inviting to dismiss these claims as common responses to the introduction of new technology, they should be given consideration in future work. We, therefore, make a deliberate effort to fine-tune our message to the public, avoiding overstatements and becoming extra sensitive to public concerns about our work.

Environmental Impact

One of the most important societal concerns for humanity today is climate change. Every action we take as a society and as individuals needs to address the potential environmental impact. The creation and development of AI systems has a significant environmental cost, especially in the training process. The training of one deep learning model has the carbon cost of 315 flights from New York to San Francisco (Strubell et al., 2019).

We believe one of the best ways to consider our impact is to factor efficiency as a key component in system design, as has been proposed by Schwartz et al. (Schwartz et al., 2019). This would mean that a bigger network that performs slightly better is worse than a more efficient network with slightly reduced performance. In the system presented in this paper, we extend this principle by placing efficiency as a primary goal and design constraint.

Overview of Robotic Musicianship at GTCMT

The Robotic Musicianship Lab at the Georgia Tech Center for Music Technology has developed multiple robotic platforms, including Shimon, Shimi, Haile, and multiple drumming prostheses. The first robotic musician was Haile, a percussionist robot designed to play a Native American powwow drum. Constructed from plywood, it used a solenoid to actuate one of the drumming arms and a linear motor to actuate the other. Shimon was the next robotic platform, designed to play the marimba and to provide visual cues with their social head. Shimon was also the first robotic musician to utilise artificial vision, and in 2019, it was transformed into a singing robot, recording and releasing an album of computer-generated songs and hip-hop freestyle. The third robot was Shimi, a table-top musical companion, designed to function as a musical social robot. We have also developed two primary robotic prostheses for amputees – a wearable drumming arm and a piano-playing arm, and a wearable drumming arm for general-purpose use. In addition to the hardware platforms, we also developed five key design principles to guide our research: listen like a human, play like a machine, be social, watch and learn, and wear it.

The first principle – listen like a human – relates to the way robots perceive music. This principle focuses on computational modelling of musical perception, with the goal of allowing robots to interpret music similarly to humans. Listening like a human requires the ability to recognise musical features, such as beat, similarity, tension, and release. Perceptual modelling of human input is crucial for meaningful interaction and collaboration as it allows robotic musicians to develop an internal model of human ensemble members' expressive, emotional, and musical creations.

Playing like a machine focuses on our mechanomorphic goal of developing novel musical outcomes not possible for human collaborators. We achieve this through hardware and software innovations focused on new techniques for musicianship. In terms of new hardware, this can involve the implementation of

brushless DC motors, allowing marimba playing at forty notes per second across eight mallets, creating new timbre, and allowing for new composition styles (Yang et al., 2020). We also incorporate software design that is built around mechanomorphic design and consider implementations of systems that offer nonhuman interactions, without an end goal of sounding like a human or passing a Turing test. This includes projects like Shimon the Rapper (Savery et al., 2020b), a robotic hip-hop system that develops new musical outcomes for human and robot performance. By combining both novel software and hardware development, our robotic musicians create innovative musical responses that push musical experiences and outcomes to uncharted domains.

Robotic platforms and embodied agents allow for social interactions not possible with computer interactive music systems, leading to the third design principle – be social. Interaction with gestures can significantly affect the musical experience, increasing the social engagement and leading to more fluent turn-taking (Hoffman & Weinberg, 2010). Each of our robotic platforms uses physical movements for visual choreography to add to the aesthetic impression for audience and performers. In particular, the percussion robot Haile was used to study the effect of ancillary gestures on co-player anticipation and audience engagement and to explore the subjects' perception of the robot and the music it generates (Weinberg et al., 2006). We have also used musical robotic platforms to develop new forms of interaction for nonmusician interactions, such as new methods for social robots to speak (Savery et al., 2020a).

Our robotic platforms also use artificial vision to watch and learn from human collaborators. In the music-making process, visual connection is key to taking advantage of social gestures and creating music as an ensemble. Musical gestures can act as cues to synchronise music and anticipate other musicians' future decisions. This work has been covered in many performances, from responding to guitar cue synchronisation to real-time detection of emotion and film analysis for live movie composition (Savery & Weinberg, 2018).

Our final robotic design principle – wear it – involves potential application as prosthetics to allow musicians with disabilities to enhance their performance ability, merging their biological body with technological enhancements. The current frontier of robotic musicianship research at Georgia Tech focuses on the development of wearable robotic limbs that allow not only amputees but also able-bodied people to play music like no human can, with virtuosity and technical abilities that are humanly impossible. This research is currently developing platforms for drum and piano performance.

Convnet

In the following section, we describe a new system designed to address many of our ethical goals, while exploring new areas of research in robotic musicianship. In previous work, we have not used real-time machine learning interaction, rather generated offline sequences before a performance. Many of the ethical goals align well with the development of a real-time system, such as portability

and low environmental impact. Real-time generation also increases the agency of the human performer, allowing much greater control over the performance as a whole.

A crucial ethical and design choice for the system was the use of data, with the goal of being trainable on a small dataset. A small dataset reduces training time and environmental impact while allowing more flexibility and future variation. Our system trains only on data given by the performer before a concert, requiring about an hour of recording. This allows us to ensure we can always have consent for the data used and allow human agency over the style that is created by the system. By allowing the user to supply their own data, we also hope to prevent bias from broader datasets that may act against the performer using the system. Custom datasets also increase the personalisation of the system to each user. Transparency is still a challenge from a technical perspective, although showing performers the training and exact data used helps develop an understanding of how ideas are being created.

The system was developed for interaction with expert performers, aiming to build off their musical vocabulary. We choose to allow for call and response and dialogue-like interaction where an improviser plays a phrase to which Shimon responds, creating a constant musical communication. These interactions can take place in strict four-bar trades or open, free-form exchanges. While implementation in Shimon was the primary goal, the system allows for software interaction using just a virtual instrument. The generative system combines a convolutional neural network (ConvNet) built-in Tensorflow 2.0 with Python and a MaxMSP patch which communicate using OSC. Figure 6.1 shows a system overview where the MaxMSP patch receives a monophonic instrument, converting it into symbolic data, and sends it to a U-Net-inspired ConvNet, which generates and returns new melodies.

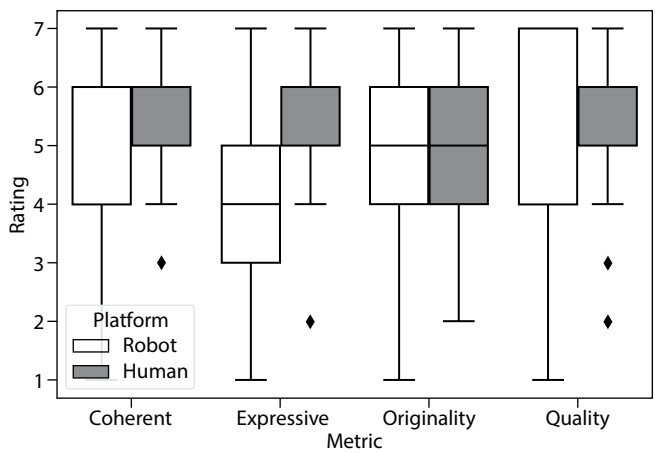

Figure 6.1 System diagram.

Considering the environmental implications, we believe it is useful for these systems to note their environmental impact. For our system, each training run uses 0.16 kg CO2 eq. (Lacoste et al., 2019), calculated based on training in an Nvidia 1080 GPU. Through iteration and testing, we trained the system eight times in total.

Interaction

There are two main forms of interaction available, either trading fours or free-form response. Both forms of interaction use the same model for processing but feature slightly different input and output methods. The system allows a MIDI keyboard or audio from a monophonic instrument for interaction; for an audio file, the notes are translated into MIDI values. For trading fours, Shimon listens to and processes the input for four bars and then generates an output for the following four bars. Before trading fours, a tempo is set, between 60 BPM and 180 BPM, with Shimon analysing the audio input at twenty-four samples per beat, allowing the system to learn triplet subdivisions. During the input cycle, a list of 384 notes is recorded, which is sent to the generation model. The model then returns a 384 list back, which is played as the melodic response.

The method for free interaction is more complicated and allows for much more variety from the performer. In this style of interaction, Shimon and the human improviser can choose to respond at any time. Shimon constantly listens to and stores the input from the human improviser, even while Shimon is playing. By constantly listening and processing the input, Shimon is able to respond back to the improviser at any time. To allow the recorded input to fit the 384 grid required by the system, Shimon has the ability to stretch or reduce the input sequence. In the trading fours version, each 384 value sequence can represent a length of 4.8 seconds (four bars at 180 BPM) to 16 seconds (four bars at 60 BPM). For the free version, we aim to keep the length of phrase in this range, but inputs often slightly vary. Variable-length input-and-output capability is achieved via a stretch-and-shift process, wherein an input melody less than four bars long is stretched temporally to fit across four bars, and its corresponding output is compressed and shifted back to the input's original time span and temporal location.

Shimon chooses when to respond based on three possibilities, either after two seconds of silence from the input stream, after twelve-second intervals, or with random interjection. Each second, there is a 10% chance of random interjection, where Shimon responds based on the most current input. In early experiments, we allowed longer silences; however, we have found that extended silences and gaps from computer performers can increase uncertainty from human collaborators.

Dataset

A key component of the system is the data representation and processing through the system. Both the trading fours and free interaction provide the model with a 384-length vector of MIDI note values. This vector is then converted to a 24×16

matrix, which is processed by the model. This novel data format arranges beats in a column such that time steps relative to the beginning of each beat are stacked on top of one another. This reshaping allows the model to learn relatively coherent musical structure and discover temporal dependencies within a phrase. We believe that the use of this data format is significant in the resulting musical coherency of this system.

We have trained the system on three datasets, for three different performances. Each performance used a dataset created by the performer specifically for the concert. We have currently worked with pianist and Hollywood composer Kris Bowers, pianist and Danish composer Signe Bisgaard, and vocalist Mary Carter. Each dataset consists of approximately 1,600 measures, which is fifty choruses of thirty-two bars recorded as MIDI data. In the future, we may allow for audio datasets that we then convert to MIDI, but currently all performers have played MIDI devices. For improvisers, recording the dataset requires about an hour of improvisation.

From tests using our own datasets, we developed multiple guidelines for the creation of the dataset. Firstly, the improvisations should be done with a click; however, the click can change to any range of tempos between 60 BPM and 180 BPM during the session. The recordings don't need to be quantised to the click, but the click should be a point of reference for the improvisations. The dataset works best if it is in a clearly defined style so should be based around the musical language that may be employed within a single improvisation and not cover a range of styles. Finally, we encouraged the improvisers to not worry about the improvisation being perfect, instead asking them to record continuously without deleting any material. Each version of our system is only trained on the improviser, who will be performing with the system.

After collecting the dataset, we then transpose each one up and down six chromatic steps to create a version of each improvisation in every key. This allows the improvisation to be independent of any key signature and is considered standard practice. To generate the call-and-response dataset, we then split the data into call (X) and response (Y), by taking four bars as the call and the following four as the response. These then overlap, so the first response then becomes a call that uses the next four as a response.

Model Architecture

Our data representation was built around our model choice of a convolutional network (ConvNet). ConvNets have had wide success in image- and video-related tasks (Khan et al., 2020). While less common in symbolic music generation tasks, ConvNets have been widely used for audio generation in WaveNet (Oord et al., 2016) and some music-generation tasks (Yang et al., 2017). From our experience, we found that for the short responses this system generates, a sequence-based model such as a recurrent neural network does not necessarily perform better and is prone to overfitting on small datasets.

The ConvNet model is based on U-Net (Ronneberger et al., 2015) and comprises a symmetrical encoder-decoder architecture in which the outputs of encoding layers are appended to the inputs of corresponding decoding layers. It uses the 384 vector as both input and output to the system. U-Net was originally developed for biomedical image segmentation and designed to work on limited training datasets, with very fast generation times. It is distinct from other ConvNet models as it returns an output of the same size as input and essentially performs a classification on every value from the input.

Embodiment

The interactive system described to this point can function as either a software system or embedded in a robot. Our end goal is always to have generations performed by robot; however, for this system we also maintained a software-only version to allow any potential performer to interact and allow for users to test the improvisation on their own computers. The software system follows Figure 6.1 with the output sent as MIDI to a software instrument. MIDI is created from the 384-vector list by connecting repeating numbers to create longer notes.

For the robotic performance with Shimon, we use our standard path-planning algorithm to turn the generation into something playable for Shimon. Shimon has four arms, with two mallets on each arm, that move linearly across a marimba but cannot cross over each other. Therefore, to reach different areas of the marimba requires careful path planning to avoid collisions. Our path-planning system uses a greedy algorithm to choose the most appropriate arm to play each note. For the robot-played version, we also use a stochastic, rule-based system to add extra embellishments in the form of tremolos. Tremolos are occasionally triggered on sustained generated notes, often at top speed and at times using syncopated, fast rhythms.

In addition to path planning, each interaction with Shimon requires head and body gestures to enhance the engagement for coplayers and audiences. For this system, we repurposed many of Shimon's standard gestures. These gestures include looking at the performer that is playing with Shimon, looking towards the marimba, and looking at the audience. These are interwoven with a robotic breathing gesture and moving to the pulse of the music.

Evaluation

Method

The usefulness and challenges of evaluating creative AI generation systems has been widely researched, although there is no single accepted practice (Sturm et al., 2019). For an interactive robotic system, a Turing test, where the computer system attempts to convince a viewer it is human, is not a relevant approach. There is also the argument that applying a Turing test to music generation is not

appropriate, as these systems are not designed explicitly to trick a human listener (Agres et al., 2016).

For this paper, we chose to base our evaluation on a repurposing of Boden's framework for computational creativity (Boden, 2009). This framework has been used previously for narrative rating (Riedl & Young, 2010) and in our own work on lyric generation (Savery et al., 2021b). Part of Boden's framework proposes that creativity is a combination of novelty and originality, expressiveness, and coherence. Boden further describes that the balance between coherent output and novelty is part of what defines the creativity of a work. We contend that these metrics further extend to the idea that in order to keep human collaborators curious and engaged, the system has to strike a balance between novelty and coherence. For our study we gathered metrics for originality, expressivity, coherence, and the overall quality rating for both the human and robot performer.

We developed one primary research question to evaluate the system: in an ensemble performance, can an improvising robot display similar levels of creativity as defined within Boden's framework? Our hypothesis was that Shimon and a human improviser would not have a significant difference in results. In addition to this evaluation using Boden's frameworks, we wished to gather some further qualitative data. To do this, we used questions developed by Sturm et al. (Sturm et al., 2019) for a live music performance. These questions asked for participants' favourite moments, surprising moments, if listening to a robot changed how they listened to the piece, and if they had any general comments.

Participants first read a virtual consent form, signed by entering their participant ID. This was followed by viewing two videos, each one ninety seconds long, of Shimon improvising in concert with a saxophone player in Denmark. The videos were played in a random order. At random times an attention check was given through an audio command in place of the stimuli that asked the participants to type a text phrase on the following page. Participants were not able to move back through the survey without restarting, which would prevent them from completing the survey. This caused any participant who missed the attention check to either stop the study or type an incorrect phrase, in which case their data was discarded. We also timed how long participants viewed the video. We had seven participants who were unable to complete the attention check or did not watch the complete video. After watching the videos, participants were shown a short text excerpt that described coherence and originality in creative work. They then answered questions related to Boden's metrics for the human and robot performer and filled out a short text response.

We recruited sixty participants on Mechanical Turk (Mturk), each classified as an Mturk master, which indicates a top-rated participant. After removing the seven who failed the attention check, we had a total of fifty-three participants. Each participant was paid $2 for the ten-minute survey. Participants were based predominantly in the USA (n=40), with the remaining participants from India (n=13). Seventeen participants identified as female, with the remaining thirty-six male. The average age was forty, with a standard deviation of 10.25 and a range of eighteen to sixty-four.

Results

Figure 6.2 shows a box plot of the results. The human and robot were rated very similarly across each metric. The means for Shimon were, coherence, 5.11; expressivity, 4.3; originality, 5.03; and overall quality, 5.28. The means for the human performer were, coherence, 5.51; expressivity, 5.80; originality, 5.05; and overall quality, 5.6. We conducted a pairwise t-test on each category and found that only expressivity had a significant value ($p<0.001$), proving our hypothesis to be partly correct.

From categorizing the text responses, we developed three main concepts. As is common with work in robotics and music, many participants commented positively on Shimon's head movements, despite the questions explicitly asking about the musical content. Overall, the participants found the gestures effective and a significant part of what they noticed in the performance. Participants described their favourite part as 'seeing the robot face move. It looked lifelike and gave it personality,' or when the robot 'moves its eye.'

The second common thread in the comments was a positive sentiment to the robot as a listener who was able to interact in meaningful ways. Participants wrote, 'I like how the robot tried to not interrupt the human and played directly with them,' and 'I was just impressed at how well the robot was able to "listen" and reply to the musical patterns before it.' Multiple participants also felt the robot was able to achieve good musical balance in the composition, such as, 'I was pleasantly surprised with how well the robot was able to play along with the human orchestra. The balance between human and robot was impressive.'

The final concept that arose throughout the comments was the impact of the use of a robot. There was no consensus on how having a robot impacted people's perception of the music, with some participants believing a robot made no difference to their perception, while others found the robot distracting and that it detracted from the human's music. Additionally, some participants thought, 'It was surprising to me that the robot did as well as it did.' However, one participant wrote, 'I was much more critical of the piece knowing that a robot played a role. The robot would have had more extensive knowledge and better motor control than the humans, so I judged it more harshly than I would a human.'

Discussion

Our results showed that Shimon performed without significant difference to the human performer for coherence, originality, and quality. This is a promising result, since an unbalanced response among these parameters could indicate problems. For example, an unbalanced high level of originality could relate to high levels of randomness and is not necessarily a positive for the system. We found that Shimon did perform worse in expressivity, perhaps due to a lack of dynamic and expressive ability in the system. It is also possible that from an audience perspective, an improvising saxophone has more expressive capabilities than a marimba sound, or that they found the human's body movements more expressive.

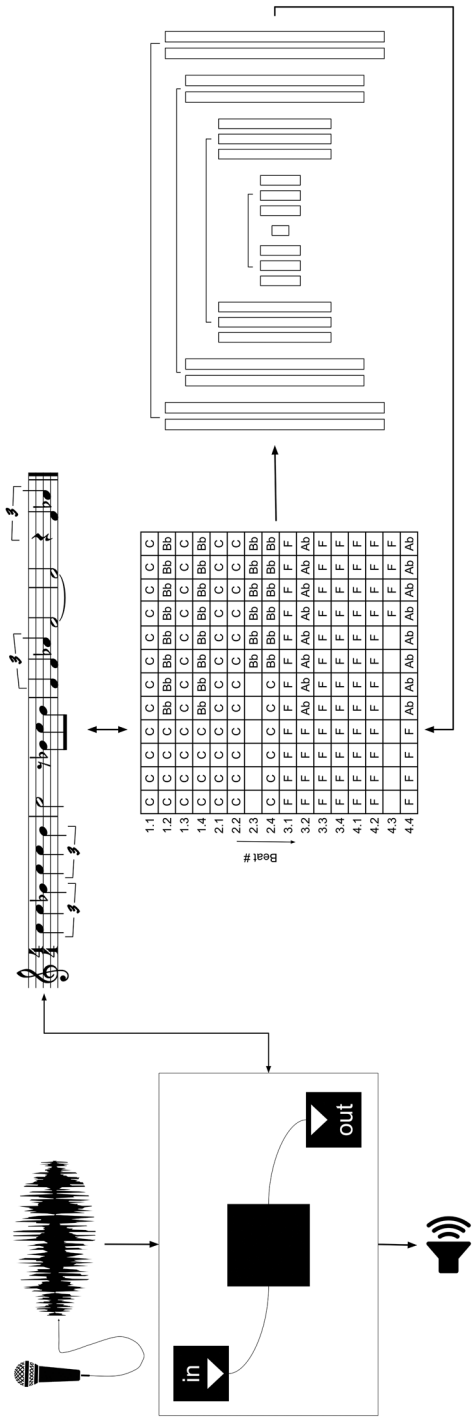

Figure 6.2 Box plot of results on Boden's framework.

As the comments indicated, gestures and robotic body movement were highly impactful to an audience. From our experience in concerts, gestures and movements are often the first thing noticed and, by many nonexpert musicians, one of the main memories that are taken from a performance. The ability for Shimon to look at the musician they are improvising with also encourages an audience to listen and notice musical sonic interactions between performers.

Our evaluation framework did not include the ethical goals we set out above. We believe that applying our ethical standards continuously through the design process framework and our postcreation reflections allows for better future development than attempting to incorporate them into a musical viewing experience.

Conclusion

Our system described here has so far been used in three concerts and recordings. The videos used for the evaluation were recorded in Denmark with the Aarhus Jazz Orchestra, in the concert *We, Robots* at Musikhuset Aarhus. This performance was awarded the Jazz Denmark Prize for 'the most innovative and creative concert experience of the year.' The system has also been featured in improvisations with film composer Kris Bowers for the BBC show *In the Studio* and has been used for a concert at the New Museum in New York. Audio and video samples are available online.[1]

We believe robotic musicianship offers a paradigm for innovative, new developments in AI and music. In this paper, we have framed our next stages of development around broader ethical goals and contend that these considerations are crucial for future musical AI design. From our prototype system, we have shown that ethical frameworks can lead to effective musical systems and encourage new directions for AI and music.

Note

1 https://richardsavery.com/project/shimonplays.

Bibliography

Agres, K., Forth, J. & Wiggins, G.A. (2016) Evaluation of musical creativity and musical metacreation systems. *Computers in Entertainment (CIE)*. 14 (3), 1–33.

Appiah, K.A. (2008) Cosmopolitanism: Ethics in a world of strangers. *Management Revue*. 19 (4), 340–341.

Barocas, S. & Selbst, A.D. (2016) Big data's disparate impact. *California Law Review*. 104, 671.

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

Briot, J.-P., Hadjeres, G. & Pachet, F. (2017) Deep learning techniques for music generation-a survey. *ArXiv preprint*. arXiv:1709.01620.

Clarke, E., DeNora, T. & Vuoskoski, J. (2015) Music, empathy and cultural understanding. *Physics of Life Reviews*. 15, 61–88.

Gomez, E., Castillo, C., Charisi, V., Dahl, V., Deco, G. et al. (2018) Assessing the impact of machine intelligence on human behaviour: An interdisciplinary endeavour. *arXiv preprint*. arXiv:1806.03192.

Hoffman, G. & Weinberg, G. (2010) Synchronization in human-robot musicianship. *19th International Symposium in Robot and Human Interactive Communication, IEEE*. 718–724.

Khan, A., Sohail, A., Zahoora, U. & Qureshi, A.S. (2020) A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*. 1–62.

Knotts, S. & Collins, N. (2020) A survey on the uptake of music AI software. *Proceedings of the International Conference on New Interfaces for Musical Expression*. 499–504.

Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. (2019) Quantifying the carbon emissions of machine learning. *arXiv preprint*. arXiv:1910.09700.

Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O. et al. (2016) Wavenet: A generative model for raw audio. *arXiv preprint*. arXiv:1609.03499.

O'Keefe, C., Lansky, D., Clark, J. & Payne, C. (2019) *Comment regarding request for comments on intellectual property protection for artificial intelligence innovation*. https://perma.cc/ZS7G-2QWF.

Riedl, M.O. & Young, R.M. (2010) Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*. 39, 217–268.

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*. Cham: Springer, pp. 234–241.

Savery, R., Rose, R. & Weinberg, G. (2019a) Establishing humanrobot trust through music-driven robotic emotion prosody and gesture. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE*. 1–7.

Savery, R., Rose, R. & Weinberg, G. (2019b) Finding shimi's voice: Fostering human-robot communication with music and a nvidia jetson tx2. *Proceedings of the 17th Linux Audio Conference*. 5.

Savery, R. & Weinberg, G. (2018) Shimon the robot film composer and deepscore. *Proceedings of Computer Simulation of Musical Creativity*. 5.

Savery, R., Zahray, L. & Weinberg, G. (2020a) *Emotional musical prosody for the enhancement of trust in robotic arm communication*. Trust, Acceptance and Social Cues in Human-Robot Interaction, Ro-MAN 2020. https://par.nsf.gov/servlets/purl/10286326.

Savery, R., Zahray, L. & Weinberg, G. (2020b) *Shimon the rapper: A real-time system for human-robot interactive rap battles*. International Conference on Computational Creativity, ICCC'20. https://researchers.mq.edu.au/en/publications/shimon-the-rapper-a-real-time-system-for-human-robot-interactive.

Savery, R., Zahray, L. & Weinberg, G. (2021a) Before, between, and after: Enriching robot communication surrounding collaborative creative activities. *Frontiers in Robotics and AI*. 8, 116.

Savery, R., Zahray, L. & Weinberg, G. (2021b) Shimon sings-robotic musicianship finds its voice. In: *Handbook of artificial intelligence for music*. Cham: Springer, pp. 823–847.

Schwartz, R., Dodge, J., Smith, N.A. & Etzioni, O. (2019) Green ai. *arXiv preprint*. arXiv:1907.10597.

Shelemay, K.K. (2013) The ethics of ethnomusicology in a cosmopolitan age. In: P.V. Bohlman (ed.), *The Cambridge history of world music*. Cambridge: Cambridge University Press, pp. 786–806.

Strubell, E., Ganesh, A. & McCallum, A. (2019) Energy and policy considerations for deep learning in nlp. *arXiv preprint*. arXiv:1906.02243.

Sturm, B.L., Ben-Tal, O., Monaghan, U´., Collins, N., Herremans, D. et al. (2019) Machine learning research that matters for music creation: A case study. *Journal of New Music Research*. 48 (1), 36–55.

Sturm, B.L., Iglesias, M., Ben-Tal, O., Miron, M. & Gomez, E. (2019) Artificial intelligence and music: Open questions of copyright law and engineering praxis. In: *Arts*. Basel: Multidisciplinary Digital Publishing Institute, Vol. 8, p. 115.

Tinker, A. & Coomber, V. (2004) University research ethics committees: Their role, remit and conduct. *Bulletin of Medical Ethics*. 203, 7.

Vochozka, M., Kliestik, T., Kliestikova, J. & Sion, G. (2018) Participating in a highly automated society: How artificial intelligence disrupts the job market. *Economics, Management, and Financial Markets*. 13 (4), 57–62.

Weinberg, G., Driscoll, S. & Thatcher, T. (2006) *Jam'aa-a middle eastern percussion ensemble for human and robotic players*. Geneva: ICMC, pp. 464–467.

Yang, L.-C., Chou, S.-Y. & Yang, Y.-H. (2017) Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint*. arXiv:1703.10847.

Yang, N., Savery, R., Sankaranarayanan, R., Zahray, L. & Weinberg, G. (2020) *Mechatronics-driven musical expressivity for robotic percussionists*. New Interfaces for Musical Expression – NIME 2020. https://arxiv.org/abs/2007.14850.