Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation

Ryan Louie*
Northwestern University
Evanston, IL
ryanlouie@u.northwestern.edu

Jesse Engel
Google Brain
Mountain View, CA, USA
jesseengel@google.com

Anna Huang
Google Brain
Montreal, Quebec, Canada
annahuang@google.com

ABSTRACT

There is an increasing interest from ML and HCI communities in empowering creators with better generative models and more intuitive interfaces with which to control them. In music, ML researchers have focused on training models capable of generating pieces with increasing long-range structure and musical coherence, while HCI researchers have separately focused on designing steering interfaces that support user control and ownership. In this study, we investigate how developments in both models and user interfaces are important for empowering co-creation where the goal is to create music that communicates particular imagery or ideas (e.g., as is common for other purposeful tasks in music creation like establishing mood or creating accompanying music for another media). Our study is distinguished in that it measures communication through both composer's self-reported experiences, and how listeners evaluate this communication through the music. In an evaluation study with 26 composers creating 100+ pieces of music and listeners providing 1000+ head-to-head comparisons, we find that more expressive models and more steerable interfaces are important and complementary ways to make a difference in composers communicating through music and supporting their creative empowerment.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; • **Computing methodologies** → *Neural networks*; • **Human-centered computing** → **User studies**.

KEYWORDS

quantitative methods; generative models; steering interfaces; human-ai co-creation;

ACM Reference Format:
Ryan Louie, Jesse Engel, and Anna Huang. 2022. Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3490099.3511159

*This work was performed during the first author's summer internship at Google Research.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03.
https://doi.org/10.1145/3490099.3511159

1 INTRODUCTION

There is an increasing interest from machine learning (ML) and human computer interaction (HCI) communities in empowering creators with better generative models and more intuitive interfaces with which to control them. In the domain of music, ML researchers have focused on training models capable of generating pieces with increasing long-range structure and musical coherence [29, 44], while HCI researchers have separately focused on designing better steering interfaces that support user control and ownership through overcoming AI-induced information overload and non-deterministic model outputs [37]. As these innovations are being transferred from the research lab into the hands of professional musicians [39] and into commercial tools [1, 49], it is increasingly clear that generative tools will have an impact on the music production and music consumption industries. These industries exist to provide technologies (Digital Audio Workstations or DAWs [38], instruments) in support of music creators and to understand the perceptions and desires of audiences as well. Given this potential for impact, it's important understand how the progress we are making across generative models and steering interfaces are actually affecting these downstream tasks as measured from the perspective of creators and audiences.

While the ML and HCI communities have similar aspirations for generative modeling, interdisciplinary collaborations have been limited by mismatches in the way each field evaluates progress. One the one hand, many ML researchers desire their models to be useful for creators, but models are not directly measured on metrics *downstream* from training, such as empowering individuals to achieve their creative goals. Instead, they are measured using proxy metrics that are easier to automate, such as the ability to generate realistic samples that imitate the training data. On the other hand, HCI researchers will conduct user studies with creators to evaluate whether an interactive generative tool is more controllable and promotes feelings of collaboration and ownership. While these studies focus on the experience of the creator as measured through self-report [37, 42, 62], an equally important metric is how the products of creation can have an effect on an outside audience. Within the domain music, it remains untested whether better steering interfaces for generative models can empower users of these tools to create compositions that better evoke an emotion or sound more musical, as judged by listeners.

As a step towards uniting these fields, this paper presents a study examining how recent developments in ML (more expressive models) and HCI (more steerable interfaces) can affect novice composers' experience when creating, as well whether the music they produces achieve specific creative goals as judged by outside-listeners. Our study follows the music creation scenario illustrated in Figure 1,

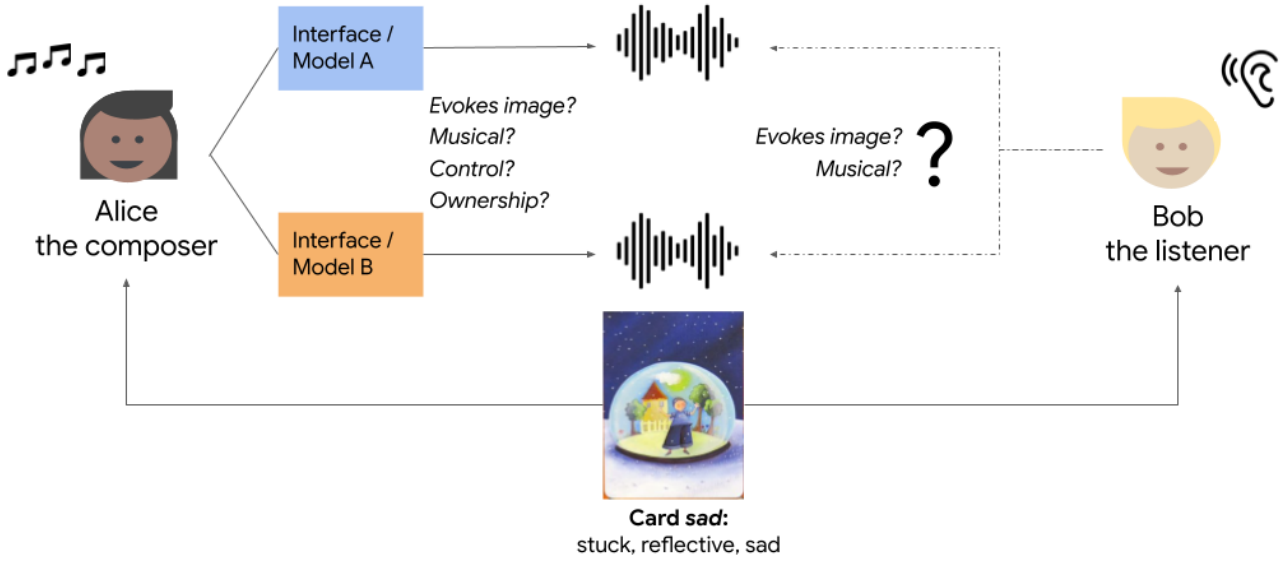


Figure 1: In our Expressive Communication study method, a novice composer (named Alice) is tasked with creating a piece of music using generative tools that expresses the imagery and words of a card. She creates two pieces of music, one for each of the different systems she is comparing, which can differ on their generative model, or their interface for control. After completing her compositions, she provides self-report answers about her experience and the compositions made with both systems. Afterwards, a listener (named Bob) is provided the same card and listens to both music compositions made by Alice. Bob answers questions about the comparison between Alice's two compositions, in terms of which better evokes the card's image and sounds more musical.

where a novice composer uses generative tools to create pieces of music that express a particular feeling and image, while an outside listener judges which music actually best evokes that feeling and image. We compare between two generative models capable of different degrees of long-range structure and musical coherence, and also between two different interfaces capable of different degrees of steering and iterative composition. Importantly, the study situates the developments of generative tools in the context of several *downstream* creative goals, and evaluates the effectiveness of varying both the model and interface of a tool in accomplishing these *downstream* metrics, i.e. composer's *subjective* self-reports of the co-creation process and their final generated music, and listener's *objective* judgements about evoking the imagery and the musicality.

The composers in our study created 100+ musical phrases [1], expressing the imagery and words in the set of illustrations in Figure 4, which were then evaluated by listeners in 1000+ head-to-head comparisons. Our results show that both the ML and HCI approaches (developing better pretrained models and better steering interfaces, respectively) are important and complementary ways to support composers in both communicating through music and feeling empowered in the process of co-creating with generative models. Our results also shed light on how biases in a pretrained model capabilities such as stronger coherence can make feelings such as fear more difficult to express with curation of random samples alone, and how the addition of steering interfaces can help to overcome model biases by creating samples that are less likely from the model, but more aligned with the user's expression and musical goals.

In summary, our paper makes the following contributions:

- We present a study that evaluates generative models and steering interfaces in the context of downstream creative goals of expressing emotion in music as measured by through novice composer's subjective self-reports and listener's objective judgements.
- Our results show how developments in more expressive generative models for music and more steerable interfaces for control are complementary approaches for impacting the products and processes of co-creation.

In the rest of the paper, we survey related work on how ML and HCI communities evaluate progress. Next, we describe the different experimental systems we built for our studies for comparing generative models with different levels of expressiveness, and interfaces with different levels of steering. Then, we introduce our evaluation study setup, measures, and analysis. We present our quantitative results and qualitative takeaways. Finally, we end with a discussion of how unified evaluations of ML and HCI progress in generative models can drive new questions and efforts towards the goal positively impacting creators and audiences.

[1] Listen to the music participants composed using generative tools for different imagery and words here: https://storage.googleapis.com/expressive-communication/index.html

2 RELATED WORK

While the ML and HCI communities have similar aspirations in building new tools for creative use, in practice their approaches and objectives differ, making it difficult to evaluate and compare their *downstream* impact on creators. We briefly survey how the two communities evaluate progress, their respective limitations, and show how our evaluation study addresses some of the challenges that arise, especially in the creative context.

2.1 ML Evaluation of Generative Models

In ML, researchers often evaluate generative models using proxy metrics that are easy to automate. The most common objective is to maximize the likelihood of the data according to the model, which is used to measure how well a model is able to fit a desired distribution. In different contexts, this can also be interpreted as the ability of the model to maximally compress the data with the smallest reconstruction error [2]. However, not only is likelihood an unreliable metric for comparing different model types, it is also not a measure indicative of sample quality [59]. Yet, most sequence models are trained to optimize for this metric. As a proxy to evaluating sample quality, researchers in image generation often use metrics based on classifiers trained on ImageNet [14], such as the Inception score [48] and the Fréchet Inception Distance (FID) [25], to measure how well generated objects bear the features learned from classifying the objects. In music generation, it is common to compute simple musically informed features to measure and compare if generative systems are able to produce samples with desirable features [61]. However, these metrics are only a proxy to human evaluations, and furthermore do not capture how useful the models will be *downstream* to creators.

Perhaps most related to this work is the framing of Reinforcement Learning (RL) tasks [58]. In RL, generative models are often used by agents to understand the environment and inform their actions [22]. While most of this work is focused on downstream tasks such as teaching artificial agents to play video games [40], an emerging thread of research is exploring ways of evaluating and optimizing for Human-AI collaboration [10]. Studies have show human feedback and interaction can be used to optimize generative models for collaboration on tasks such as classification or assistive game playing [33, 45]. This work examines a new approach towards evaluating and optimizing for Human-AI collaboration in the domain of creative expression.

2.2 HCI Evaluation of Interactive Systems and Interfaces

In HCI, many interactive systems and interface techniques have been explored to support human-AI creation with a generative system across domains such as drawing [19, 42], creative writing [21], and design ideation [35]. To evaluate these systems and techniques, researchers will conduct user studies to understand the impact they have on users' self-report feelings, their behaviors in the process, and the quality of the resulting artifacts.

To support the evaluation of creativity tools, significant research has focused on measuring the process and experience. For example, the Creativity Support Index (CSI) measures the impact on six dimensions such as exploration, expressiveness, immersion, enjoyment, results worth effort, and collaboration [11]. While such research efforts have sought to provide a standard and reliable measure of comparison for a tool, psychometric indexes like CSI rely on users' subjective self-reports. Ultimately, instead of reflecting a creator's *subjective* perspective on the impact of the tool, we argue that evaluation studies that seek to understand if tools support creators effectiveness need equal emphasis for evaluating the final creative outputs.

In this vein, several methods have been employed to measure the products of creation in other domains. Within design and ideation, measures like ratings from expert judges or the quantity of products within a creative ideation have been employed [52]. Within interpretibility of generative models, novel tasks have been proposed for the predefined goal of interactively reconstructing a given target artifact [47], allowing researchers to compare which models enabled users to complete the task better. Nonetheless, in the context of expressive creative practices like music, such objectives have been harder to define. As such, many user studies within the music domain are situated in more open-ended creative tasks [27, 37]. With our work, we seek to provide a comprehensive evaluation using *downstream* metrics such as composer's *subjective* self-reports and listeners *objective* listener judgements of an expressive communication task. In doing so, our work provides the necessary flexibility inherent in expressive creativity while also providing more objective measures upon which to evaluate the generated musical outputs.

3 COMPARISON OF GENERATIVE MODELS AND STEERABLE INTERFACES FOR MUSIC CREATION

With the goal of evaluating how developments in generative models and steering interfaces impact downstream metrics, we built several experimental systems that would be compared against each other in the study. To evaluate the impact of more expressive generative models, we referenced two generative music models from the ML literature to inform how we implemented two systems that provide different levels of expressiveness. To evaluate the impact of more steerable interfaces, we drew upon two interfaces from the HCI literature to inform how we built two systems that provide different levels of steering.

3.1 Comparing Generative Models with Different Levels of Expressiveness

Within the field of generative ML, researchers have sought to develop models that are capable of more expressive timing and dynamics [43] to models that are capable of building upon and developing expressive themes and motifs through long-range coherence [29]. For our studies, we compared two trained autoregressive generative models capable of generating polyphonic piano music:

- **Performance RNN** [43] is the baseline and less expressive model. It is trained on a smaller and more dramatic piano performance dataset (MAESTRO) [23], and has less capacity in the model architecture to model the distribution of music.

- **Music Transformer** [29] is the more expressive model. It is trained on a larger and more melodic piano performance dataset (YouTube) [53], and has larger capacity to model the distribution of music.

3.2 Comparing Interfaces with Different Levels of Steering

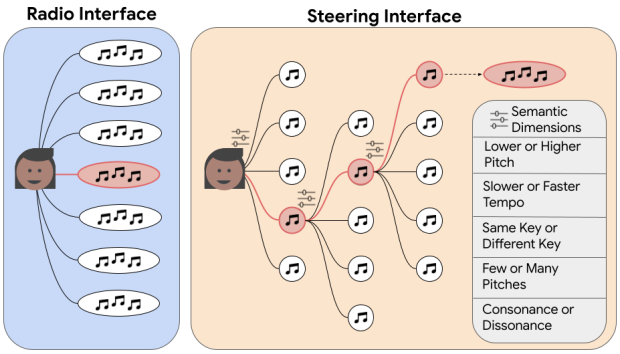


Figure 2: We compared two interfaces for creating music with an autoregressive generative model: the Radio Interface and the Steering Interface.

To design the two interfaces that provide different levels of steering, we drew upon existing interfaces for generative music tools from the HCI literature. For our experimental systems, we implemented a baseline and less steerable interface called the **Radio Interface**, and a **Steering Interface** which embodies several interaction principles for creating more steerable generative tools; see Figure 2. A composer using the Radio Interface requests randomly generated options of the full-length phrase of music, and selects the one that best matches her goals. A composer using the Steering Interface will explore and select three total generated chunks for the start, middle, and end of their musical phrase. For exploring any chunk, they can request a randomly generated set, or they can choose to constrain the generation along semantic dimensions (e.g., lower or higher pitch; consonance or dissonance). After selecting any chunk, subsequent chunks are generated as a continuation of the previous chunks, in an autoregressive fashion.

In what follows, we describe the theoretical motivations and design of the two interfaces and the implementation details of their front-ends and back-ends.

3.2.1 Theoretical Motivation and Design of Interfaces. In designing a baseline interface, we referenced studies of how composers currently interact with "end-to-end" deep generative models that are similar to Performance RNN and Music Transformer [28]. Based on these studies, a common approach when interacting with an "end-to-end" generative model is to generate a large quantity of musical samples, and curate through them to find desired results. As a baseline approach, musicians can manually curate by listening to many alternatives, before selecting generated options which are most interesting or "catchy" or which might fit well for their musical goals. In our experiments, we implemented this baseline interface and refer to it as the *Radio Interface* (see Figure 2, left), inspired by how a user chooses to listen to music from one station out of many stations on a radio [17]. A composer using the Radio Interface requests randomly generated options of a full-length phrase of music, and selects the one that best matches their goals.

When designing a steering interface for our experiments, we drew upon HCI research about building better interactive tools that provide users (e.g., novice composers) more control and agency when creating with deep generative models [37]. This research recommended two helpful interaction principles, which we used when building our experimental system: (1) generating music chunk-by-chunk and (2) having controls to constrain the generation along semantically-meaningful dimensions. However, since our Steering Interface would be used with an autoregressive generative model like Music Transformer—which is fundamentally different than the in-filling model [26] used in prior work [37]—we needed to adapt these interaction principles for autoregressive models.

We present a diagram of the user interaction flow for using the *Steering Interface* on the right-side of Figure 2. The steering tools support listening to generated music one chunk at a time before moving onward. To do this, the autoregressive output is split into three chunks (5 seconds each), and users can choose one of the generated options for the current chunk, before seeing the next chunk which is a continuation of all subsequent chunks. Additionally, the steering tools provide users with several semantic dimensions for controlling the generation. For example, a user can select from several dropdowns on whether they want "lower or higher pitch", "slower or faster tempo", etc. These dimensions were chosen to align with how novice composers think of meaningful dimensions of variation when expressing more abstract ideas and concepts through music.

3.2.2 Implementation Details of Front-End. The front-end for the Radio Interface and Steering Interfaces was implemented in a Google Colab notebook; see Figure 3. For the Radio Interface, a user can request 10 randomly generated options that are all approximately 15 seconds in length by running a notebook cell. The front-end for viewing the generated options is the same across the Radio and Steering Interface. Each generated option has an audio player, a visualized piano roll, and a text area where a user can jot notes to help them remember each option. A user can scroll horizontally to view all the generated options. For the Steering Interface, users can interact with the model by (1) choosing which 5 second chunk to generate (1st, 2nd, 3rd chunk) by navigating to the respective sections of the notebook; (2) specifying any semantically-meaningful-dimensions using dropdown widgets (2) generating options for this chunk by running the cell. The Steering Interface provides 10 generated options in the 1st chunk to mirror the diversity of parent nodes in the Radio Interface, whereas the tool provides 5 generated options in the second and third chunk.

3.2.3 Implementation Details of Back-End. To implement the back-end of the Steering Interface's chunk-by-chunk interaction, we used the autoregressive models to pre-generate a forest of trees with 100 parent seeds for the first chunk, 100 child continuations in the second chunk, and 100 child continuations in the 3rd chunk. This makes for 1 million possible 15 second full-length phrases that can be sampled from the model. The tool provides 10 generated options

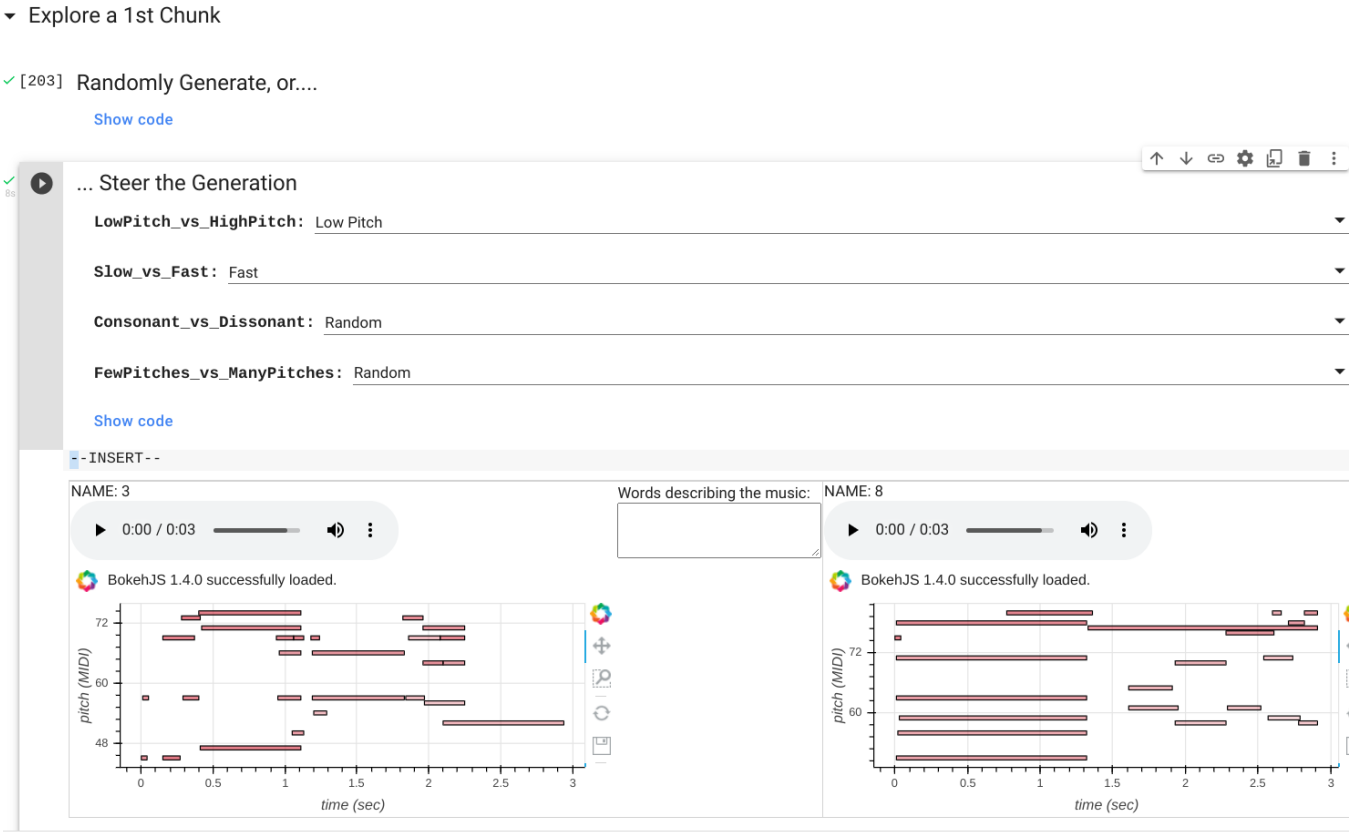


Figure 3: We built these systems using a Google Colab notebook, and participants used these systems from the notebook by adjusting the interactive widgets and running cells.

in the 1st chunk to mirror the diversity of parent nodes in the Radio, whereas the tool provides 5 generated options in the second and third chunk.

To implement the back-end of the Steering Interface's semantic dimensions interaction, we used a rejection sampling approach to filter generated output along different musically-meaningful dimensions. We defined several heuristic functions that can characterize some musical feature of a chunk through a numerical value. These functions fall under two categories: absolute musical features which use the absolute numerical value (i.e., slow vs fast); and relative musical features describing a current chunk's musical attribute relative to its previous chunk (i.e., make the second chunk "slower or faster" than the first chunk).

4 EVALUATION STUDY

We are broadly interested in evaluating how recent progress in ML and HCI for generative music tools is impacting downstream measures of creative effectiveness and empowerment. Thus we ask the following research questions: **RQ1** Are the final compositions better in the eyes of outside listeners (e.g., in evoking the desired feeling, sounding more musical) when they are created using (a) a more expressive generative model or (b) more steerable interface? **RQ2** (How) do composers feel more empowered (e.g., achieving creative goals, ownership, efficacy) when they use (a) a more expressive generative model or (b) a more steerable interface?

4.1 Experimental Conditions

In this within-subjects study, our participants created music compositions in two experiments: a *model comparison* between two pretrained generative models, and an *interface comparison* between two interfaces for exerting influence on the final generated output.

For the model comparison, participants created two more musical pieces with two systems that differed on their pretrained generative models: *PerformanceRNN* which is algorithmically simpler, less expressive and less able to elaborate a music theme; and *MusicTransformer* which is the more expressive model and has algorithmic capabilities that can model longer-range themes and structures in the music. To interact with these models, participants used the Radio Interface to curate and find best generated sample that matched the ideas of a card.

For the interface comparison, participants created two pieces of music with two systems that differed in their interfaces (as illustrated in Figure 2): a baseline *Radio Interface* from which they select a best match from a set of randomly generated full-length phrases; and a *Steering Interface* from which creators have the ability to iterate and perfect smaller sections of the phrase before moving

Figure 4: We selected 5 image cards that came from the board game Dixit [60]. They were selected to cover different parts of the valence and arousal dimensions of emotion. To ensure that composers and listeners had a common interpretation of the card, we attach three keywords to each card.

onto selecting new ones. These two interfaces were used to steer the generated outputs of the MusicTransformer model.

We chose to only use the Music Transformer for the interface comparison due to constraints on how many systems we could reasonably ask participants to use. In order to mitigate effects caused by fatigue or boredom, we opted for only performing these two comparisons, with a total of the four variants provided to users. While 3 out of the 4 system variants used the MusicTransformer model, we mitigated any effects due to familiarity by using a practice tutorial using both models to ensure they were equally familiar with the models before the experiment.

In our experimental setup, we counterbalanced both the ordering of the comparisons (e.g., whether they were to first use the two different interfaces, or the two different models), and the ordering within each comparison (i.e., for the interface comparison, whether they would first use the Steering or Radio Interface). For each system comparison (interface comparison, model comparison), we assigned each participant a randomly selected card from the set of five cards; see Figure 4. This ensured we controlled for participants having the same expressive communication goal when comparing within a comparison.

4.2 Composer Study Method

4.2.1 Study Setup. The 26 participants who completed the study included 12 females and 14 males, ages 24 - 60 (μ = 37). Users were recruited through mailing lists at our institution and came from a variety of professional backgrounds (e.g., designer, administrator, technical writer, engineer). Each received a $50 gift credit for their time. We chose to recruit N=26 participants by comparing our study plan to other within-subject studies who use similar measures (e.g., control, ownership) [36]. Since we anticipated a similar effect size for these measures, we recruited N>=21 to achieve similar statistical power.

The composer sessions were conducted remotely with participants over a video meeting where they shared their screen. Each user was given an overview of the goals of the study and the expressive communication game (10 minutes). Participants were assigned to a comparison (interface comparison or model comparison) according to the counterbalanced ordering described in Section 4.1, and completed a guided tutorial of the systems they would be using in their first comparison (15 minutes). For example, suppose a hypothetical participant, Alice, was assigned to the model comparison first; Alice would have been guided on a tutorial to create two sample musical phrases evoking the an emotion with both Performance RNN and Music Transformer. In the first comparison, participants were assigned a card and were asked to compose music that reflected the imagery and words of the card using each of the systems being compared for the comparison (10 minutes per system). Users were observed while composing using a think-aloud procedure. Finally, they answered a post-comparison questionnaire and completed a semi-structured interview comparing their experiences (10 minutes). This procedure was repeated for the second comparison, including being guided through a new tutorial, composing two compositions using the two systems to express a single card, and a post-comparison questionnaire and interview (40 minutes). At the end, they answered several questions about their overall experience composing in this expressive communication game (10 minutes).

4.2.2 Measures and Analysis. To answer our RQs about the impact of different generative tools on composers' experiences, we used a combination of quantitative survey ratings, and qualitative data from interviews and talk-alouds. For our quantitative measures, we asked participants to rate six survey measures for both system versions they had used. All measures below were rated on a 7-point Likert scale (1=Strongly Disagree, 7=Strongly Agree).

We were interested in measuring how participants felt about the final piece of music they created using each of the generative systems, which motivated the following set of metrics. **Expression**: Users rated *"I feel confident that my composition created with System X expresses the imagery/ideas of the card I chose."* **Communication**: Users rated *"I feel confident that others will be able to rank the Card I chose as high after listening to the composition created with System X."* **Musical Coherence**: Users rated *"I feel the composition I created with System X feels musically coherent."*

Motivated by the importance of supporting effective, human-centered partnerships with AI [4, 36, 42] and supporting intrinsically-rewarding creative activities [13], we asked three questions about participants' attitudes towards the generative tool and their composition experience using the tool. **Ownership**: Users rated *"Using System X, I felt the composition created was mine."* **Control**: Users rated *"I felt I had control creating the composition when using System X."* **Efficacy**: Users rated one item from the Generalized Self-Efficacy scale [51] rephrased for music composition, which asked *"Using System X, I could find several solutions to achieve my goals for the composition."*

To analyze these quantitative differences, we conducted a two-sample paired t-test, with the null hypothesis that the mean of their differences is zero. We used a two-sample paired t-test because our within-subjects design had a participant provide a rating for both system versions that they used. Since we use multiple statistical tests in our composer study, we decided to use a Bonferroni correction method to correct for the chance of a type 1 error (e.g., the chance we find a significant effect for a perceived composer metric, when there is truly no difference). When testing for significance, we used a Bonferroni-corrected threshold of 0.00416 to correct for the 12 tests performed (6 survey questions asked for the 2 comparisons of different models, and different interfaces).

For our qualitative analysis, we coded the quotes from interviews and talk-aloud sessions using a deductive, thematic analysis [8] according to the various dimensions we asked in quantitative results (e.g., expression; musical coherence; control), and grouped by the type of model or interface used.

4.3 Listener Study Method

4.3.1 Study Setup. We recruited 20 unique listeners for the listener study from an online crowd work platform. Each listener made head-to-head comparisons for all 51 pairs of musical samples created by the participants in our composer study (26 interface comparisons and 25 model comparisons). In total, our listeners provided 1020 head-to-head ratings comparing the pairs of music compositions.

We asked listeners to compare the two phrases of music which were created by the same composer, for the same card. We chose this to control for the variations in composers interpretations of a given card, and to control for variations in how easy it might be to express one card vs. the other. The ordering of the music options were randomized to prevent an association with ordering and experimental conditions.

4.3.2 Measures and Analysis. After being given two musical samples to compare, listeners rated two questions that asked *"Which one of these musical excerpts **most evokes the feelings** of the words and imagery on the card?"* and *"Which one **sounded more musical**"* and answered on a 5 point balanced scale (*"Strong preference for option 1"*, *"Weak preference for option 1"*, *"No preference"*, *"Weak preference for option2"*, and *"Strong preference for option2"*). In preparation for conducting our analysis, we converted this scale to a numerical scale from $[-2, 2]$. For for the model comparison, positive values correspond to a preference for Music Transformer; for the interface comparison, positive values correspond to a preference for the Steering Interface. To analyze these listener ratings, we used a one-sample t-test, with the null hypothesis that there would be no preference ($\mu = 0$) for either of the models or interfaces.

5 QUANTITATIVE RESULTS

5.1 Composer Study

5.1.1 Model Comparison. The composer self-report ratings for the model comparison is shown in Figure 5a. During the study, participants found the compositions made with Music Transformer to have more **musical coherence** ($\mu = 5.88$) as compared to ones made with Performance RNN ($\mu = 4.64$) where a paired t-test found the difference to be significant ($t = 3.3$, $p < 0.003$). We did not find a significant difference between the compositions made with two models on **expressing** the musical qualities of the card ($\mu_1 = 5.32$, $\mu_2 = 4.72$, $t = 1.78$, $p = 0.087$), or their confidence in the music **communicating** the ideas of the card for a listener to guess ($\mu_1 = 4.56$, $\mu_2 = 4.44$, $t = 0.35$, $p = 0.72$).

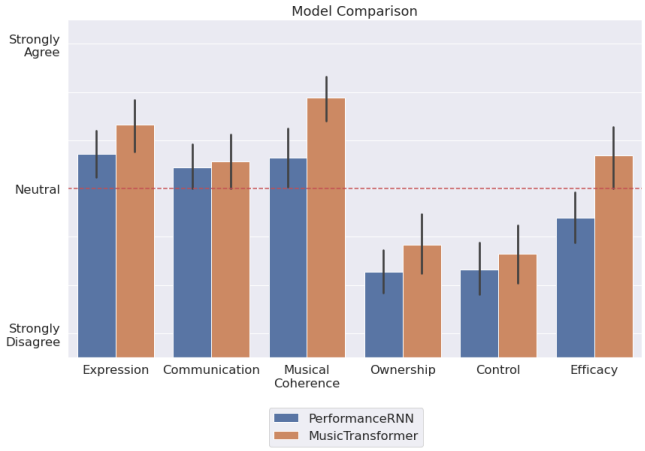
Participants felt greater **ownership** of the composition created using Music Transformer over the one made with Performance RNN ($\mu_1 = 2.84$, $\mu_2 = 2.28$, $t = 3.22$, $p < 0.004$), and also had greater **efficacy** in finding multiple solutions to achieve their goals using Music Transformer ($\mu = 4.68$) as compared to Performance RNN ($\mu = 3.4$), where the difference was statistically significant ($t = 5.02$, $p < .0001$). Accounting for our Bonferroni correction threshold, we did not find a significant difference in **control** between Music Transformer and Performance RNN ($\mu_1 = 2.64$ $\mu_2 = 2.32$, t=2.33, p=0.0307).

5.1.2 Interface Comparison. The composer self-report ratings for the interface comparison is shown in Figure 5b. No significant difference was found between the final music created with either interface for **expressing** the musical qualities of the card ($\mu_1 = 5.38$, $\mu_2 = 4.96$, $t = 0.98$, $p = 0.33$), their confidence in the music **communicating** the ideas of the card for a listener to guess ($\mu_1 = 5.15$, $\mu_2 = 4.92$, $t = 0.58$, $p = 0.56$), or **musical coherence** ($\mu_1 = 5.69$, $\mu_2 = 5.96$, $t = 0.96$, $p = 0.35$).

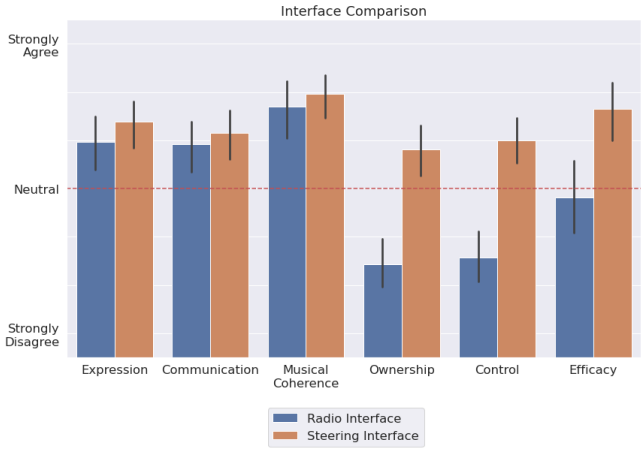
Participants did feel a greater sense of **ownership** of the composition created using the Steering Interface over the one made with the Radio Interface ($\mu_1 = 4.8$, $\mu_2 = 2.42$, $t = 7.7$, $p < 1e{-}7$), and felt they had more **control** using the Steering Interface as compared to Radio Interface ($\mu_1 = 5.0$, $\mu_2 = 2.58$, $t = 6.6$, $p < 1e{-}6$). They also had greater **efficacy** in finding multiple solutions to achieve their goals using the Steering Interface ($\mu = 4.68$) as compared to the Radio Interface ($\mu = 3.4$), where the difference was statistically significant ($t = 4.13$, $p < 0.001$).

5.2 Listener Study

5.2.1 Model Comparison. The listener results for the model comparison is shown in Figure 6a. Listeners felt compositions made with **Music Transformer** better evoked the feelings of the card over compositions made with **Performance RNN**, where the difference was statistically significant ($\mu = .24$, $t = 3.318$, $p < 0.001$). In addition, listeners rated compositions made with Music Transformer to sound more musical than those made with Performance RNN ($\mu = 0.378$, $t = 4.74$, $p < 0.00001$).

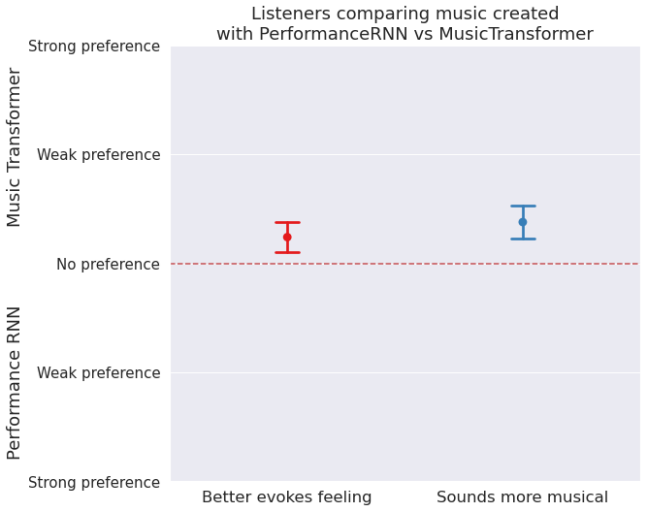


(a) Composers' ratings comparing between different models.

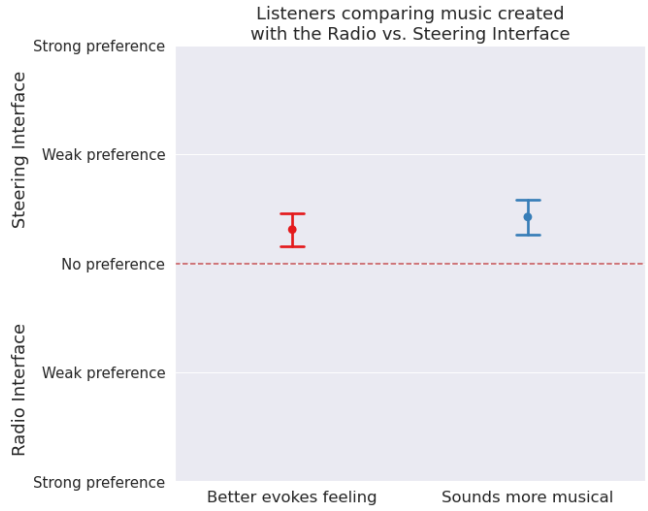


(b) Composers' ratings comparing between different interfaces.

Figure 5: Composers answered questions about on the final compositions made and about their experiences using the different interfaces, and different generative models.



(a) Listeners' ratings comparing between models.



(b) Listeners' ratings comparing between interfaces.

Figure 6: Listeners compared compositions created for different interfaces and different generative models, along the dimensions of evoking the feelings of the card, and sounding more musical. We visualize a 95% confidence interval for the point estimates of each question.

5.2.2 Interface Comparison. The listener results for the interface comparison is shown in Figure 6b. Listeners on average felt compositions made with the **Steering Interface** better evoked the feelings of the card over compositions made with the **Radio Interface**, where the difference was statistically significant ($\mu = 0.31$, $t = 4.09$, $p < 0.0001$). Additionally, listeners felt compositions created with the **Steering Interface** sounded more musical than those made with the **Radio Interface** ($\mu = -0.5846$, $t = 3.472$, $p < 0.001$).

5.3 Discussion of Composer and Listener Study Quantitative Results

In sum, providing composers with a better model (while keeping the baseline Radio Interface constant) resulted in a small but statistically significant difference in ownership and efficacy, which composers attributed to the model generating more viable options to choose from. Moreover, providing composers with a more steerable interfaces to the generation helped them feel more ownership, control, and efficacy in finding multiple solutions.

For creating a composition that better evokes the feeling of the imagery and words when a listener hears it, using a system with a more steerable interface, or a more expressive model, both make a difference. However, from composers' self report, no overall difference in expressiveness or communication was found between music created with one system versus the other, across interfaces and models.

For creating a composition that sounds more musical to listeners' ears, using a system with a more steerable interface, or a more expressive model, again both make a difference. From the perspective of composers, compositions made with the Music Transformer, a model capable of maintaining long-term structure in the music, also makes a difference for musical coherence; however, no difference was found in composer-perceived musical coherence between compositions made with the different interfaces.

6 CORE TAKEAWAYS

Having described our main quantitative results, we now enrich these findings with qualitative evidence from composer interviews, and secondary analyses of the listener results with respect to how the generative tools supported expressing some emotions better than others.

6.1 Better Models and Interfaces Support Creative Empowerment and Effectiveness

Our present study evaluated how using tools powered by more expressive autoregressive models, or better steering interfaces for these models can make a difference in downstream creative measures. Our results show that both better steering interfaces and more expressive models make a difference in composer's feelings of empowerment (i.e., control, ownership, efficacy) and their effectiveness in creating music that evokes the intended feelings and that sounds more musical to outside listeners.

6.1.1 Coherent and Expressive Models Help Evoke Imagery and Ideas. When reflecting on the difference between models, composers described Music Transformer as capable of evoking a coherent musical idea. For example, a participant said Music Transformer communicated a *"single thread"* which *"repeats and builds upon a small theme,"* ultimately helping to evoke a clear image and mood (P5). In contrast, composers commented that the musical phrases generated by Performance RNN were disjoint between sections, were less coherent, and *"don't know what they want to be"* (P12). Another participant elaborated that the music samples from Performance RNN would *"often change the tempo or tonality too fast, and there wasn't a consistent feeling in it"*, which made it difficult to find samples that *"expressed something that was clearly evocative"* (P8). As we found in our quantitative results, participants curating random samples from Performance RNN had lower efficacy in finding musical samples that matched the creative ideas they were trying to express. After asking composers to reflect on their experience using Performance RNN, they felt that since "there were fewer [options] that I was happy with, I felt like I did have less control and fewer solutions" (P21). Some participants elaborated that the lack of a consistent theme in the generated music made it difficult to find many candidates since *"in each of them there's always something, like an interjection, that does not quite match the card"* (P9).

Our results also highlight how models that generate music that are expressive and relatable also helped composers with the expressive communication task. Participants felt that Music Transformer tended to generate more modern music or pop-songs which were easier for them relate to. For example, composers sometimes would connect a certain music to a particular cultural reference, such as reminders to specific songs they have heard (e.g., *"it sounds very similar to the song 'Silent Night', so I already imagine snow and winter"*) and types of music they have heard used in specific contexts (e.g., *"this is music that plays at the beginning of a Japanese anime drama"*; *"it reminds me of the music that plays during a boss battle in a video game"*). They could use these connections to other contexts to guide their decision on whether this music expressed the matching imagery or words of the cards. In contrast, participants noticed that the Performance RNN model would generate music that associated with *"classical music"* or a *"20th century modern style* (P8). While the generated music was complex and sophisticated, it did not express something that was clearly evocative. These findings shed light on the importance of the cultural-relatedness and expressive moods of the pretrained models from which co-creation tools are built upon.

6.1.2 Better Steering Interfaces Empower Composers. Better Steering Interfaces empowered composers by making a positive impact in their feelings of control, ownership and efficacy; see the measurable quantitative difference in Figure 5b. Participants using the steering interface felt they *"could more piece it together"* through the ability to make choices for each chunk, and said they could *"compose a flow, a narrative"* and *"think about the direction it takes"* (P7). Through allowing the composer to be more involved in incrementally building up the piece, composers could see how "different elements show up in the different chunks that I chose" (P12). This extra involvement and work to match the music helped composers feel more ownership: "because I put that extra effort and put more thoughts into it, I liked it better" (P22).

In comparison, composing by curating random outputs provided very little control. Since composers had little ability to provide input beyond making a final selection, some participants described their role more as a *"listener and evaluator"*, rather than as composer or collaborator. This mode (or lack thereof) of interaction coincided with a loss of ownership when using the Radio interface. As one participant said, *"since I showed my preference only a little bit in choosing from the different examples, I don't feel it was mine"* (P4).

From our quantitative results, creators felt they had increased efficacy in finding multiple music phrases that matched the ideas and mood of the card when using the Steering Interface. For composers who did find multiple solutions, a common strategy was to focus on finding a first chunk that captured the general mood and tempo, through a combination of semantically constraining the generation and curating to find a match. From there, composers felt confident that they had set a good initial direction, and often noticed that many generated options in subsequent chunks continued the express the same feeling established by the first chunk.

Interestingly, since creators could find multiple options in the subsequent chunks that matched the ideas and mood, they had greater flexibility to control the musicality of the piece and use their musical preferences to guide the final output. For example,

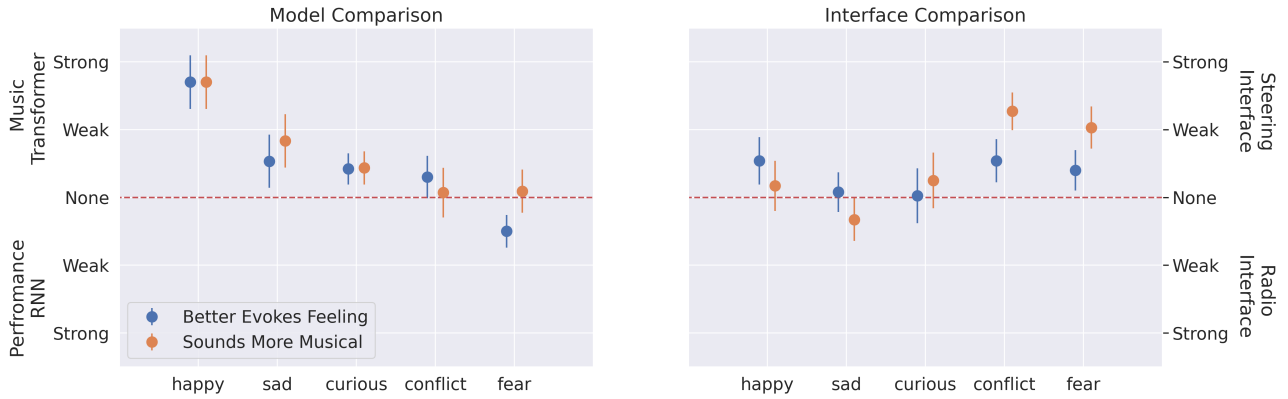


Figure 7: Listener ratings, comparing which model and interface best evoke the target feeling and/or sound more musical. For graphing, we use simplified descriptions to denote the feeling of each card. Comparing the model ratings by card (left) exposes bias in the pretrained models, where curated random samples from Music Transformer clearly better evoke the feelings of happy, sad, conflict, and curious, while Performance RNN samples better evoke fear. Interestingly, these communication preferences are strongly correlated with sounding more musical for happy, sad, and curious, but there is no musical preference for conflict and fear. This exposes that the pretrained Music Transformer has a model bias towards coherent musical output which is more aligned with straightforward feelings such as happy or sad. Further, one way of evoking conflict and fear is with musically incoherent pieces, as Performance RNN is biased to output. Comparing interface ratings across cards reveals that music created with the steering interfaces do no better than curated random samples on evoking feelings like sad and curious.

a participant mentioned that *"because I could find many phrases that match, I had room to think about how polished and consistent the phrases and chunks were"* (P3). As another way of focusing on musical coherence, many participants when composing the last chunk felt it was important to provide an ending that resolved the musical phrase, and so focused selecting a generated option that satisfied this musical goal. For example, one participant who was expressing the feeling for "satisfied" said *"I really like this third chunk, which ended on a proud note. It is very consonant and does resolve"* (P15).

6.2 Steering can Help Overcome Model Biases

6.2.1 Biases in Pretrained Models Help to Better Evoke Some Feelings Over Others. Comparing model preferences by card (Figure 7, left) exposes a bias in pretrained models, where curated random samples from Music Transformer evoke the feelings of happy, sad, curious, and conflict, while Performance RNN samples better evoke fear. Interestingly, effectively evoking the feeling is strongly correlated with sounding more musical, revealing that pretrained Music Transformer has a bias towards coherent musical output which is more aligned with straightforward feeling such as happy or sad.

From the perspective of composers, several felt that Music Transformer was especially useful for expressing imagery and ideas from some cards, such as dreamy, curiosity, sad, easy going, and satisfied. However, several participants felt that Music Transformer's generated outputs could be *"too steady"*, which they felt *"worked well for walking or peaceful but less so for something with a climax"* (P12). In contrast, participants using Performance RNN took advantage of the bias in its generated outputs towards feelings like fear using *"rhythms that are really not steady... with short-long-short-long patterns that interfere with each other"* which they felt made this sample *"the creepiest"* (P14).

6.2.2 Steering Interfaces are More Helpful When Expressing Feelings that are Misaligned with Model Biases. Comparing interface ratings across cards (Figure 7, right) shows that when the goal is to express feelings like sad and curious, music created with the steering interfaces do no better than curated random samples. This reveals that when random samples from an expressive model easily aligns with a goal, it can work just as well as steering to express those same feelings. This finding is corroborated with evidence from our composer study, where composers whose goal was to express straightforward feelings felt they could achieve their same goal just as well, but in a different way, when curating random samples generated from Music Transformer rather than steering. In these cases, they were open-minded to *"ceding control and seeing where it takes me"* by letting the model generate music that surprised them and was different than their initial expectations (P1). For example, another participant said *"I was surprised by the random composition, taking it into a musical direction that I would not have otherwise gone... this tool makes me go beyond my first idea, giving ideas that help one express an idea in a different way"* (P15).

On the contrary, when the goal is to express feelings like conflict and fear that are less likely to be represented in the random samples, steering interfaces made a difference in creating music that better evokes the feeling (see Figure 7, right). Observations during the talk-aloud composers sessions help to provide additional evidence

of how steering can help to express these feelings. As an example of a steering strategy to express the feelings of fear, a composer selected a first chunk that is *"slower and has less energy to signal something impending that hasn't happened yet"*, added more feelings of ominous by setting the semantic parameters to *"slower, lower, and different key"*, and steered the third chunk to be faster and much lower in the end to *"build tension at the end to signal that the character is being attacked by the killer plants"* (P3).

Interestingly, the compositions made through steering interfaces sound more musical too when the goal is to express feelings like fear and conflict. We observed in the composer studies that participants using curation of random samples would try to convey "fear" with music that would keep the listener uncomfortable or not knowing what would come next. For example composers picked up on *"short long note patterns"* and *"sudden pauses and tempo shifts"* to distinguish from the other samples; thus, they would sacrifice coherence in this context to evoke this emotion. In contrast, with the steering interface, they would find solutions that satisfied both the "fear feeling", through interesting dynamics and developments across the chunks, while also retaining a musically sophisticated sound. As one participant described, *"The [music created with] chunks could be more coherent when compared to the radio version... the radio version has the right elements, but because of that, they aren't glued together"* (P3).

7 DISCUSSION AND FUTURE WORK

7.0.1 Evaluations with Downstream Creative Tasks and Outside-Listeners can Help Discover Biases. Our evaluation situated the developments of generative tools in the context of several downstream creative goals, and evaluated the effectiveness of varying both the model and interface of a tool in accomplishing these downstream metrics, i.e. composer subjective self reports and listener's objective judgements. Our results showed that the Music Transformer model, trained on a larger corpus of simpler music, demonstrates a bias towards long-range structure and musical coherence, which can make emotions like fear more difficult to express with random curation alone. In contrast, the Performance RNN model, trained on a smaller corpus of virtuosic music, is biased towards musically incoherent outputs that more easily evoke conflict and fear.

The existing literature on bias in machine learning has investigated the source of bias in pretrained models, which span the training data input [5–7, 9, 16, 18, 34, 41] to the technical decisions made with regard to the learning, architecture, and implementation of the model itself [24, 50, 57]. Within the domain of music, existing work has researched how *popularity bias* of certain musical styles or genres (i.e., larger representation of content that appeals to a wider audience) can be mitigated when providing recommendations [20]. In our work with music generative models, these pretrained model biases could be attributed to choices in the music training data and the technical architectures of the model. Music Transformer was trained on piano music sourced from Youtube, whose platform may have a popularity bias towards musical styles like "pop piano" [31]. We can hypothesize that the representation of pop piano in the training data might have made it easier to generate music the evokes emotions like happiness and sadness, rather than ones like conflict or fear. At the same time, the technical decision to make Music Transformer's architecture powerful at modeling long-range dependencies also played a role in creating coherent, predictable repetitions of the phrase—which were less conducive to creating the sudden rhythmic or melodic shifts that could evoke feelings of fear or conflict. Future work could better characterize and document details about how various creative tasks are impacted by (1) the training datasets used for creative generative systems (expanding on the datasheet format used in [55]) and (2) the modeling decisions that were made for a particular pretrained generative model.

7.0.2 Communication as a Unified Task and Metric for Evaluating Generative Tools for Co-Creation. Our evaluation method is distinguished in that it uses communicating a target emotion through the co-created artifact as a common task and metric for creativity to evaluate developments in generative models and steering interfaces. Future work could employ this evaluation method to understand and compare other generative models and interfaces. For example, it can be useful to compare models within a family (e.g., autoregressive [30, 31], like we have done in this paper; infilling [26, 32]; latent-space models like VAEs [46]) that lend themselves to similar interfaces, since the experiment can control for interface. Researchers could explore how the same steering interface overcomes different types of pretrained model biases. Comparing different steering interfaces for a model family (e.g., semantic-dimension sliders [15, 36]; natural language prompts [3, 54]) could help us understand their impact on the products and process of co-creation.

7.0.3 Benchmarking Across Studies with Objective Listener-Centric Metrics. The Expressive Communication study method used in our evaluation (see Figure 1 can be commonly applied across more interactive tools and models as a useful measure for how well they do. We anticipate that our evaluation method can promote better cross comparisons across various music co-creation studies. For example, researchers and developers of new intelligent user interfaces for music co-creation can task participants with generating music pieces according to our Expressive Communication study method by using the same set of imagery and words of the cards used in our studies (as shown in Figure 4). Then, researchers can investigate how the music created using their new model or interface can better evoke the ideas as compared to the generated musical outputs from our studies as a baseline.

7.0.4 Objective for Training New Collaborative Human-AI Systems. A new frontier for ML research is explicitly training models to optimize downstream Human-AI collaborative tasks [10, 33, 45]. As human interaction is expensive and does not scale well to training, these works use limited human interactions and evaluations to learn approximations of human behavior and preferences that can be used for automated training [12, 56]. The Expressive Communication study method used in this paper could be adapted to these approaches to enable human evaluation to scale to providing direct training signal for adapting models to better communicate and create more musical samples. This line of work often considers the interface to be fixed and optimize a model for the collaboration objective. One future avenue to possibly bring together ML and HCI research would be to use HCI insights to parameterize a space of possible interfaces and use agents (with policies distilled

from imitating humans) as "proxy users", enabling automated optimization of a new user interface that best allows a proxy user to satisfy the approximate human evaluation objective. Such a system would enable ML and HCI researchers to better utilize limited human interaction and evaluation to bootstrap new systems where the models and interfaces are co-optimized for better Human-AI collaboration. While this line of thinking is speculative, this work represents a first step towards that direction, demonstrating the value of evaluating generative tools with a unifying study method that features both human demonstration and evaluation towards a common objective.

REFERENCES

[1] [n. d.]. Ableton Live. https://www.ableton.com/en/live/. Accessed: 2020-05-04.

[2] Alex Alemi, Ben Poole, Ian Fischer, Josh Dillon, Rif A Saurus, and Kevin Murphy. 2018. An information-theoretic analysis of deep latent-variable models. *OpenReview: https://openreview.net/forum?id=H1rRWl-Cb* (2018).

[3] Safinah Ali and Devi Parikh. 2021. Telling Creative Stories Using Generative Visual Aids. *arXiv preprint arXiv:2110.14810* (2021).

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 3, 13 pages. https://doi.org/10.1145/3290605.3300233

[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[6] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation* 13, 4 (2021), 1008–1018.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[8] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[9] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[10] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-AI coordination. *Advances in Neural Information Processing Systems* 32 (2019), 5174–5185.

[11] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014).

[12] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4302–4310.

[13] Kate Compton and Michael Mateas. 2015. Casual Creators.. In *ICCC*.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[15] Ruta Desai, Fraser Anderson, Justin Matejka, Stelian Coros, James McCann, George Fitzmaurice, and Tovi Grossman. 2019. Geppetto: Enabling semantic design of expressive robot behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173986

[17] Monica Dinculescu. 2020. Listen to Transformer. https://magenta.tensorflow.org/listen-to-transformer.

[18] Patrick Esser, Robin Rombach, and Björn Ommer. 2020. A note on data biases in generative models. *arXiv preprint arXiv:2012.02516* (2020).

[19] Judith E Fan, Monica Dinculescu, and David Ha. 2019. collabdraw: An Environment for Collaborative Sketching with an Artificial Agent. In *Proceedings of the 2019 on Creativity and Cognition*. ACM, 556–561.

[20] Andres Ferraro. 2019. Music cold-start and long-tail recommendation: bias in deep representations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 586–590.

[21] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 296.

[22] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* (2018).

[23] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2018. Enabling factorized piano music modeling and generation with the MAESTRO dataset. *arXiv preprint arXiv:1810.12247* (2018).

[24] Thomas Hellström, Virginia Dignum, and Suna Bensch. 2020. Bias in Machine Learning–What is it Good for? *arXiv preprint arXiv:2004.00686* (2020).

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[26] Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Doug Eck. 2017. Counterpoint by Convolution. In *Proceedings of the International Conference on Music Information Retrieval*.

[27] Cheng-Zhi Anna Huang, David Duvenaud, and Krzysztof Z Gajos. 2016. Chordripple: Recommending chords to help novice composers go beyond the ordinary. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*.

[28] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J Cai. 2020. AI song contest: Human-AI co-creation in songwriting. *ISMIR* (2020).

[29] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. *ICLR* (2018).

[30] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2019. Music Transformer. In *International Conference on Learning Representations*.

[31] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Generating Music with Rhythm and Harmony. *arXiv preprint arXiv:2002.00212* (2020).

[32] Daphne Ippolito, Cheng-Zhi Anna Huang, Curtis Hawthorne, and Douglas Eck. 2018. Infilling Piano Performances. In *NeurIPS Workshop on Machine Learning for Creativity and Design*.

[33] Natasha Jaques, Jennifer McCleary, Jesse Engel, David Ha, Fred Bertsch, Douglas Eck, and Rosalind Picard. 2020. Learning via social awareness: Improving a deep generative sketching model with facial feedback. In *Workshop on Artificial Intelligence in Affective Computing*. PMLR, 1–9.

[34] Hannah Rose Kirk, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, Yuki Asano, et al. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. *Advances in Neural Information Processing Systems* 34 (2021).

[35] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design Ideation with Cooperative Contextual Bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 633, 12 pages. https://doi.org/10.1145/3290605.3300863

[36] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[37] Ryan Louie, Andy Coenen, Cheng-Zhi Anna Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3313831.3376739

[38] Mark Marrington et al. 2017. Composing with the digital audio workstation. *The Singer-Songwriter Handbook* (2017), 77–89.

[39] Nathan Mattise. 2019. How YACHT fed their old music to the machine and got a killer new album. https://arstechnica.com/gaming/2019/08/yachts-chain-tripping-is-a-new-landmark-for-ai-music-an-album-that-doesnt-suck/.

[40] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.

[41] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[42] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 649, 13 pages.

https://doi.org/10.1145/3173574.3174223

[43] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. 2020. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications* 32, 4 (2020), 955–967.

[44] Christine Payne. 2019. MuseNet. https://openai.com/blog/musenet. Accessed: 2020-05-04.

[45] Siddharth Reddy, Anca D Dragan, and Sergey Levine. 2021. Pragmatic Image Compression for Human-in-the-Loop Decision-Making. *arXiv preprint arXiv:2108.04219* (2021).

[46] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *International Conference on Machine Learning (ICML)*. http://proceedings.mlr.press/v80/roberts18a.html

[47] Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L Glassman, and Finale Doshi-Velez. 2021. Evaluating the Interpretability of Generative Models by Interactive Reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[48] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. *Advances in neural information processing systems* (2016).

[49] Aiva Technologies SARL. 2022. AIVA The Artificial Intelligence composing emotional soundtrack music. https://aiva.ai/.

[50] Sebastian Schelter and Julia Stoyanovich. 2020. Taming technical bias in machine learning pipelines. *Bulletin of the Technical Committee on Data Engineering* 43, 4 (2020).

[51] Ralf Schwarzer and Matthias Jerusalem. 1995. Generalized Self-efficacy Scale. *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs* 1, 1 (1995), 35–37.

[52] Pao Siangliulue, Joel Chan, Steven P Dow, and Krzysztof Z Gajos. 2016. IdeaHound: improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 609–624.

[53] Ian Simon, Cheng-Zhi Anna Huang, Jesse Engel, Curtis Hawthorne, and Monica Dinculescu. 2019. Generating Piano Music with Transformer. *Magenta Blog: https://magenta.tensorflow.org/piano-transformer* (2019).

[54] Charlie Snell. 2021. Alien dreams: An emerging art scene. https://ml.berkeley.edu/blog/posts/clip-art/.

[55] Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. 2021. Artsheets for Art Datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=K7ke_GZ_6N

[56] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325* (2020).

[57] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[58] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[59] L Theis, A van den Oord, and M Bethge. 2016. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*.

[60] Wikipedia contributors. 2019. Dixit (card game) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Dixit_(card_game)&oldid=908027531. [Online; accessed 19-September-2019].

[61] Li-Chia Yang and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Computing and Applications* 32, 9 (2020).

[62] Yijun Zhou, Yuki Koyama, Masataka Goto, and Takeo Igarashi. 2021. Interactive Exploration-Exploitation Balancing for Generative Melody Composition. In *26th International Conference on Intelligent User Interfaces*. 43–47.