Expressive Communication: A Common Framework for Evaluating Developments in Generative Models and Steering Interfaces

RYAN LOUIE, Northwestern University, IL

JESSE ENGEL, Google Research, USA

ANNA HUANG, Google Research, Canada

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 through a common framework 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.

Additional Key Words and Phrases: quantitative methods; generative models; human-ai co-creation;

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 [15, 22], 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 [18].

While the ML and HCI communities have similar aspirations for generative modeling, interdisciplinary collaborations have been limited by a lack of common frameworks to evaluate progress. Many ML researchers desire their models to be useful for creators, but most evaluations of these generative models stop short of directly measuring and optimizing metrics *downstream* from training, such as empowering individuals to achieve their creative goals. Instead, progress is often measured with proxy metrics that are easier to automate, such as the ability of a model to compress a dataset
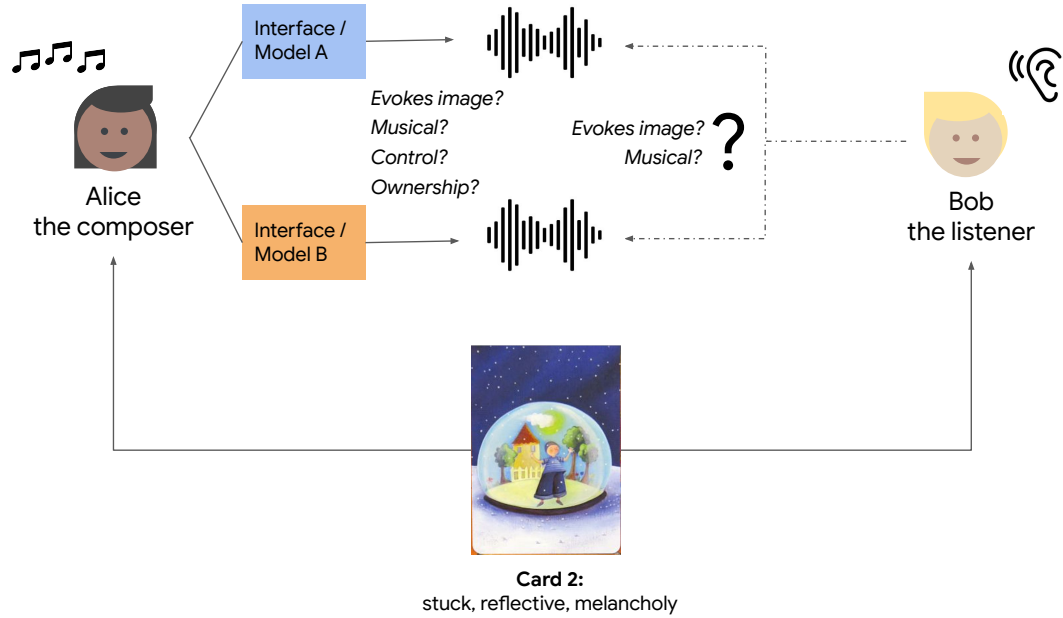
Fig. 1. In the Expressive Communication framework, Alice is tasked with composing a piece of music that expresses the imagery and words of a card. She creates two pieces of music, one for each of the different systems she is comparing, which can differ on their interfaces or their models. After completing her compositions, she provides self-report answers about her experience and the compositions made with both systems (e.g., how well does the music evoke the imagery and words? how musically coherent is your composition? how much control did you feel using this system?). Afterwards, Bob is provided the same card and listens to both music compositions made by Alice. Bob answers questions that have him compares these compositions (e.g., which music better evokes the imagery and words of the card? which sounds more musical?)

or generate realistic samples that imitate the training data. For example, in music, studying composers using deep generative models "in-the-wild" is a nascent field of study [14, 29], and not directly incorporated into the ML research process itself.

HCI researchers have separately focused on designing better interfaces for steering existing deep generative models. These studies often treat the model architecture (and just as importantly, the training data) as fixed, adapting off-the-shelf pretrained models. Further, many studies have focused on the experience of the creator, such as demonstrating that interfaces for steering generation and leading the collaboration can support users' experience during co-creation, including their control, ownership, engagement, and creativity [18, 20, 34]. In addition to self-report, separately measuring the products of creation and the effect they have on outside audience is an equally important metric. Within the domain of deep generative models for music, it remains untested if better steering can allow empower creators to better communicate feelings or create compositions that sound more musical, as judged by listeners.

2 THE EXPRESSIVE COMMUNICATION FRAMEWORK

As a step towards uniting these separate developments across ML and HCI, this paper proposes *Expressive Communication* as a common framework for evaluating progress in generative tools (including both the model and the interface). In

Expressive Communication, we not only consider the effect of a tool on the creator's subjective experience, but also how it objectively affects what they create. Consider the music creation scenario illustrated in Figure 1, where "Alice the composer" uses generative tools to create a pieces of music that express a particular feeling and image, while "Bob the listener" judges which music actually best evokes that feeling and image. Creators self-report their perceptions of the co-creation process and their final generated music, while outside listeners evaluate that music on the basis of how well they evoke the target creative ideas and their overall musical quality. This expression task was chosen to align with common real-world tasks in music content creation such as creating music that establishes a mood, or creating accompanying music for another type of media like a video [8]. Our framework enables us to situate user studies in the context of a range of *downstream* creative goals, and evaluate the effectiveness of varying both the model and interface of a tool in accomplishing these *downstream* metrics, i.e. composer's *subjective* self-reports and listener's *objective* judgements.

By using the Expressive Communication framework, we are able to compare both different generative models, and different interfaces to those models. As a proof of principle, 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. The composers in our study created 100+ musical phrases [1], expressing the imagery and words in the set of illustrations in Figure 3, 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.

3 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 framework addresses some of the challenges that arise, especially in the creative context.

3.1 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 [1]. However, not only is likelihood an unreliable metric for comparing different model types, it is also not a measure indicative of sample quality [31]. 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 [5], such as the Inception score [25] and the Fréchet Inception

---

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

Distance (FID) [12], to measure how well generated objects bear the features learned from classifying the objects. In music generation, it is common to use compute simple musically informed features to measure and compare if generative systems are able to produce samples with desirable features [33]. 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 [30]. In RL, generative models are often used by agents to understand the environment and inform their actions [10]. While most of this work is focused on downstream tasks such as teaching artificial agents to play video games [19], an emerging thread of research is exploring ways of evaluating and optimizing for Human-AI collaboration [2]. 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 [16, 23]. This work examines a new approach towards evaluating and optimizing for Human-AI collaboration in the domain of creative expression.

3.2   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 [7, 20], creative writing [9], and design ideation [17]. 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 [3]. 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 frameworks 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 [26]. Within interpretibility of generative models, novel tasks have been proposed for the predefined goal of interactively reconstructing a given target artifact [24], 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 [13, 18]. 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.

4   GENERATIVE MODELS AND STEERABLE INTERFACES

4.1   Steering Through Tree Search and Semantic Filtering

Today, a common approach when interacting with a generative model is to generate a large quantity of musical samples, and curate through them to find desired results [14]. 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

Radio Interface   **Steering Interface**
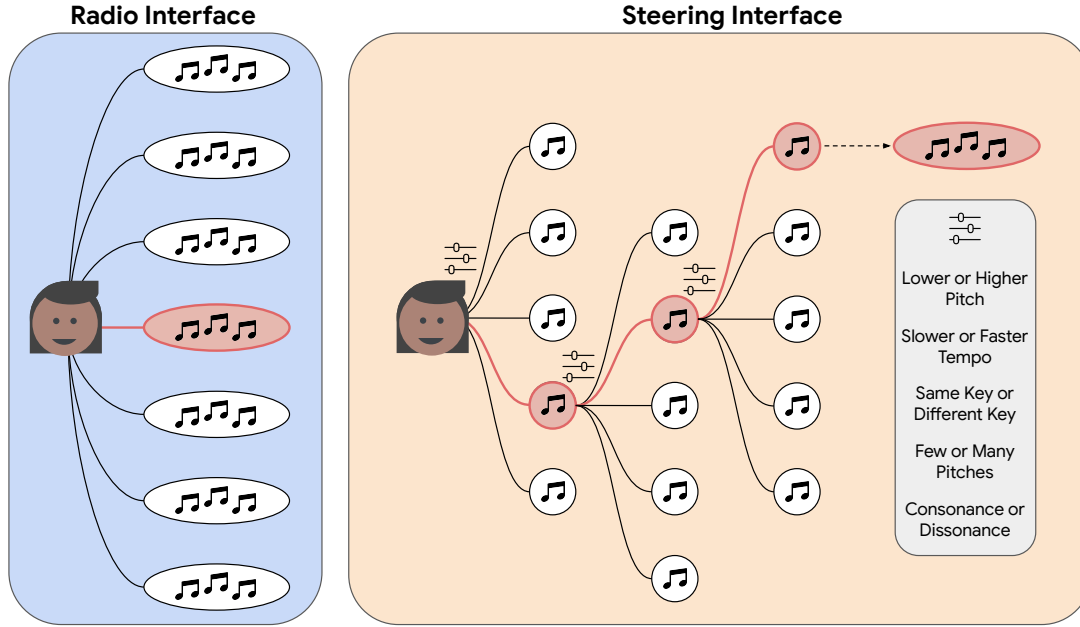

Fig. 2. In our experiments, we compared two interfaces for creating music with a generative model: the Radio Interface and the Steering Interface. A composer using the Radio Interface requests randomly generated options of the full-length phrase of music, and selects the one best matches her goals. A 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.

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 [6]. For the Music Communication task, creators request ten randomly generated phrases that are all approximately 15 seconds and can select the option that best matches the ideas and imagery of the card.

Recent developments in HCI research have sought to provide users (e.g., novice composers) more control and agency when creating with generative models. For example, researchers have developed steering tools for a deep generative AI models [18] and showed that two core capabilities can be helpful: (1) generating music chunk-by-chunk and (2) having controls to constrain the generation along semantically-meaningful dimensions. Thus, we used these design principles when designing the steering tools for the autoregressive Music Transformer model.

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 designed to align with how novice composers think of meaningful dimensions of variation when

expressing more abstract ideas and concepts through music. To control the generation, we use a rejection sampling approach to filter generated output along different semantic dimensions.

For our implementation of chunk-by-chunk tree search, 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 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.

For our implementation of semantic filtering, 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 chunks musical attribute relative to its parent (i.e., make the second chunk "slower or faster" than the first chunk).

4.2 More Capable and Expressive Generative Models

Within the field of generative ML, researchers have sought to develop models that are capable of more expressive timing and dynamics [21] to models that are capable of building upon and developing expressive themes and motifs through long-range coherence [15]. For our studies, we chose two trained autoregressive generative models capable of generating polyphonic piano music:

- PerformanceRNN [21] is the less expressive model. It is trained on a smaller and more dramatic piano performance dataset (MAESTRO) [11], and has less capacity in the model architecture to model the distribution of music.
- Music Transformer [15] is the more expressive model. It is trained on a larger and more melodic piano performance dataset (YouTube) [27], and has larger capacity to model the distribution of music.

5 EVALUATION STUDY

We are broadly interested in how the Expressive Communication evaluation framework can measure the progress in ML and HCI efforts to support creative empowerment with computer generated music.

Thus we ask the following research questions: **RQ1** How are the final compositions created using (a) a more expressive generative model and (b) more steerable interface perceived by outside listeners (e.g., in evoking the desired feeling, sounding more musical); and **RQ2** How do composers feel more empowered (e.g., achieving creative goals, ownership, efficacy) when using a tool that (a) is powered by a more expressive generative model, or (b) that has more steerable interfaces?

5.0.1 Experimental Conditions. In this within-subjects study, our participants created music compositions in two experiments: an *interface comparison* between two interfaces for exerting influence on the final generated output, and a *model comparison* between two pretrained generative models. 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 onto selecting new ones. These two interfaces were used to steer the generated outputs of the MusicTransformer model. For the model comparison, participants created two more musical pieces with two systems that differed on their pretrained generative models: *PerformanceRNN* which is algorithmically more simple, 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

Card *happy*: satisfied, easy-going, open   **Card *sad*:** stuck, reflective, sad   **Card *curious*:** curiosity, dreamy, peaceful   **Card *conflict*:** battle, climactic, intense   **Card *fear*:** preying, creeping, threatening

Fig. 3. We selected 5 image cards that came from the board game Dixit [32]. 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.

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.

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 Chunks 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 3. This ensured we controlled for participants having the same expressive communication goal when comparing within a comparison.

5.0.2 Composer Study Method. The 26 participants who completed the study included 12 females and 14 males, ages 24 - 60 ($\mu = 37$). 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.

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). They completed a guided tutorial of the systems they would be using in the first comparison (15 minutes). 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).

To test the difference between composers ratings, we conducted a two-sample paired t-test, with the null hypothesis that the mean of their differences is zero.

5.0.3 Listener Study Method. We recruited 20 unique listeners for the listener study from an online crowd work platform. They made head-to-head comparisons for each of the pairs of musical samples created by a composer expressing the imagery and words of a particular card. After finishing our composer studies, pairs of music compositions were made for

(a) Listeners' ratings comparing between models.   (b) Listeners' ratings comparing between interfaces.
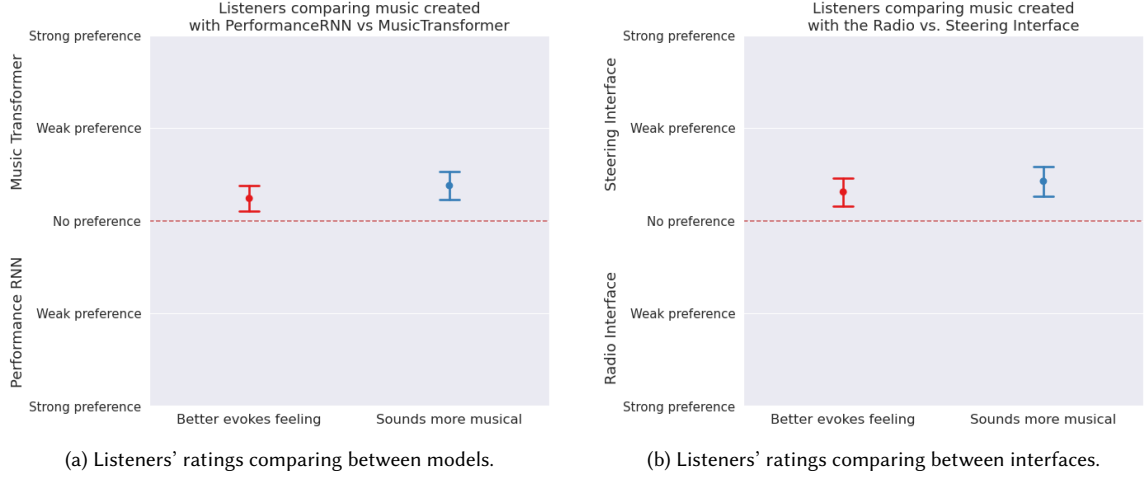
Fig. 4. Listeners compared compositions created for different interfaces and different generative models, along the dimensions of evoking the feelings of the card, and sounding more musical.

51 comparisons (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.

Listeners were 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 corresponded to a preference for MusicTransformer; for the interface comparison, positive values correspond to a preference for the Steering Interface.

6 QUANTITATIVE RESULTS

6.1 Listener Study

6.1.1 Interface Comparison. The listener results for the interface comparison is shown in Figure 4b. 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** ($\mu = -0.5846$, $t = 3.472$, $p < 0.001$).
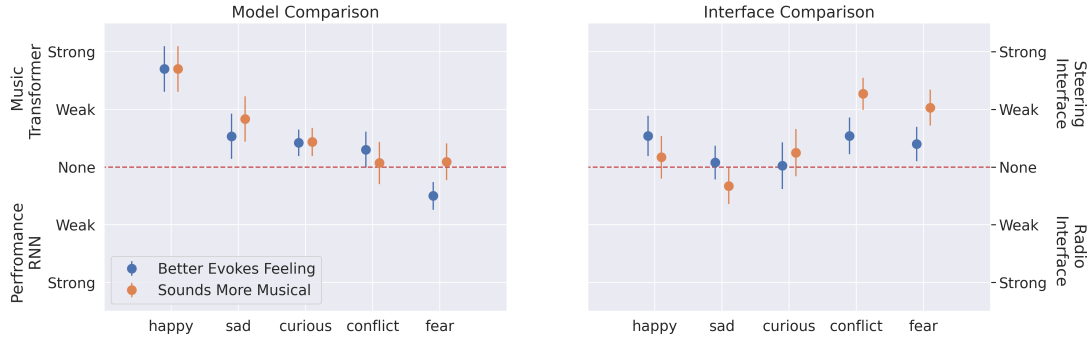
Fig. 5. Listener ratings, comparing which model and interface best evoke the target feeling and/or sound more musical. For graphing, we use simplified descriptions to denote the feeling of each card. Comparing the model ratings by card (left) exposes bias in the pretrained models, where curated random samples from MusicTransformer clearly better evoke the feelings of happy, sad, conflict, and curious, while PerformanceRNN 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 MusicTransformer 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.
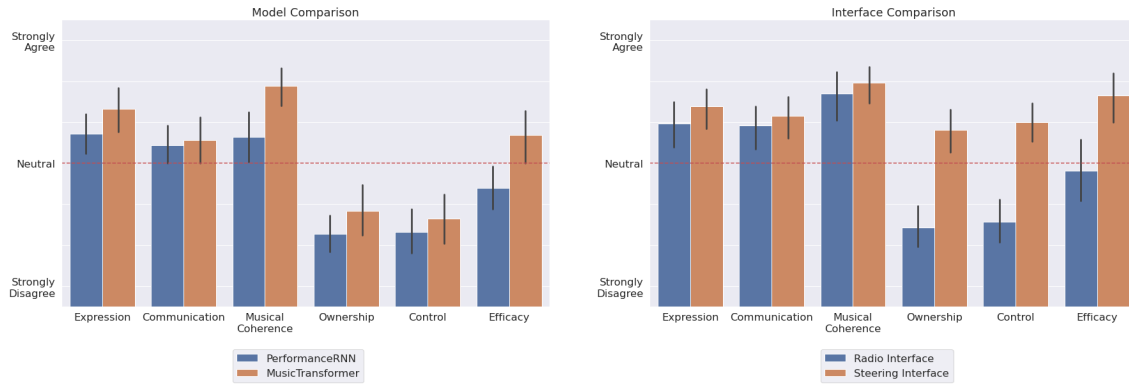
(a) Composers' ratings comparing between different models.   (b) Composers' ratings comparing between different interfaces.

Fig. 6. Composers answered questions about on the final compositions made and about their experiences using the different interfaces, and different generative models.

6.1.2 Model Comparison. The listener results for the model comparison is shown in Figure 4a. Listeners felt compositions made with **MusicTransformer** better evoked the feelings of the card over compositions made with **PerformanceRNN**, where the difference was statistically significant ($\mu = .24$, $t = 3.318$, $p < 0.001$). In addition, listeners rated compositions made with MusicTransformer to sound more musical than those made with PerformanceRNN ($\mu = 0.378$, $t = 4.74$, $p < 0.00001$).

6.2 Composer Study

6.2.1 Interface Comparison. The composer self-report ratings for the interface comparison is shown in Figure 6b. 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 ($t = 4.13$, $p < 0.001$).

6.2.2 Model Comparison. The composer self-report ratings for the model comparison is shown in Figure 6a. During the study, participants found the compositions made with MusicTransformer to have more **musical coherence** ($\mu = 5.88$) as compared to ones made with PerformanceRNN ($\mu = 4.64$) 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 MusicTransformer over the one made with PerformanceRNN ($\mu_1 = 2.84$, $\mu_2 = 2.28$, $t = 3.22$, $p < 0.004$), and felt they had more **control** using MusicTransformer ($\mu = 2.64$) as compared to PerformanceRNN ($\mu = 2.32$). They also had greater **efficacy** in finding multiple solutions to achieve their goals using MusicTransformer ($\mu = 4.68$) as compared to PerformanceRNN ($\mu = 3.4$), where the difference was statistically significant ($t = 5.02$, $p < .0001$).

6.3 Discussion of Composer and Listener Study Results

In sum, providing composers with a more steerable interfaces to the generation helped them feel more ownership, control, and efficacy in finding multiple solutions. Moreover, providing composers with a better model (while keeping the baseline Radio Interface constant) resulted in a small but statistically significant difference in ownership and control, which composers attributed to the model generating more viable options to choose from. Given that better interfaces and better models matter to these measures of empowering users, we emphasize that providing a more steerable interface led to a bigger increase in ownership, control, and efficacy.

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.

7 CORE TAKEAWAYS

7.1 Developments in Models and Interfaces are Complimentary for Creative Empowerment and Effectiveness

We used the Expressive Communication Framework to study how using tools powered by more expressive autoregressive models, or better steering interfaces for these models can make a difference. 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.

7.1.1 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 6b. 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, 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).

7.1.2 Coherent and Expressive Models Help Evoke Imagery and Ideas. When reflecting on the difference between models, composers described MusicTransformer as capable of evoking a coherent musical idea. For example, a participant said MusicTransformer 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

IUI '22, March 03–05, 2018, Finland

Louie, et al.

by PerformanceRNN 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 PerformanceRNN 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 PerformanceRNN 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 PerformanceRNN, 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 MusicTransformer 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 PerformanceRNN 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.

7.2 Steering can Help Overcome Model Biases

7.2.1 Biases in Pretrained Models Help to Better Evoke Some Feelings Over Others. Comparing model preferences by card (Figure 5, left) exposes a bias in pretrained models, where curated random samples from MusicTransformer evoke the feelings of happy, sad, curious, and conflict, while PerformanceRNN samples better evoke fear. Interestingly, effectively evoking the feeling is strongly correlated with sounding more musical, revealing that pretrained MusicTransformer 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 MusicTransformer 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 MusicTransformer'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 PerformanceRNN 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).

7.2.2 Steering Interfaces are More Helpful When Expressing Feelings that are Misaligned with Model Biases. Comparing interface ratings across cards (Figure 5, 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 MusicTransformer 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 5, 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).

8 DISCUSSION AND FUTURE WORK

8.0.1 Benchmarking Across Studies with Objective Listener-Centric Metrics. The Expressive Communication framework can be commonly applied across more interactive tools and models as other researchers develop them as a useful measure for how well they do. We anticipate that the Expressive Communication framework 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 framework using the imagery and words of the cards used in our studies 3. 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.

8.0.2 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 [2, 16, 23]. 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 [4, 28]. The Expressive Communication framework 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 framework that features both human demonstration and evaluation towards a common objective.

REFERENCES

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

[2] 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.

[3] 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).

[4] 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.

[5] 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.

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

[7] 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.

[8] Emma Frid, Celso Gomes, and Zeyu Jin. 2020. Music creation by example. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[9] 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.

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

[11] 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).

[12] 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).

[13] 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*.

[14] 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).

[15] 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).

[16] 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.

[17] 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

[18] 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

[19] 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.

[20] 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

[21] 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.

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

[23] 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).

[24] 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*.

[25] 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).

[26] 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.

[27] 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).

[28] 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).

[29] Bob L Sturm, Oded Ben-Tal, Úna Monaghan, Nick Collins, Dorien Herremans, Elaine Chew, Gaëtan Hadjeres, Emmanuel Deruty, and François Pachet. 2019. Machine learning research that matters for music creation: A case study. *Journal of New Music Research* 48, 1 (2019).

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

[31] 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)*.

[32] 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].

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

[34] 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.