AI SONG CONTEST: HUMAN-AI CO-CREATION IN SONGWRITING

Cheng-Zhi Anna Huang[1] **Hendrik Vincent Koops**[2] **Ed Newton-Rex**[3]
Monica Dinculescu[1] **Carrie J. Cai**[1]
[1] Google Brain [2] RTL Netherlands [3] ByteDance
annahuang,noms,cjcai@google.com, h.v.koops,ednewtonrex@gmail.com

ABSTRACT

Machine learning is challenging the way we make music. Although research in deep generative models has dramatically improved the capability and fluency of music models, recent work has shown that it can be challenging for humans to partner with this new class of algorithms. In this paper, we present findings on what 13 musician/developer teams, a total of 61 users, needed when co-creating a song with AI, the challenges they faced, and how they leveraged and repurposed existing characteristics of AI to overcome some of these challenges. Many teams adopted modular approaches, such as independently running multiple smaller models that align with the musical building blocks of a song, before re-combining their results. As ML models are not easily steerable, teams also generated massive numbers of samples and curated them post-hoc, or used a range of strategies to direct the generation, or algorithmically ranked the samples. Ultimately, teams not only had to manage the "flare and focus" aspects of the creative process, but also juggle them with a parallel process of exploring and curating multiple ML models and outputs. These findings reflect a need to design machine learning-powered music interfaces that are more decomposable, steerable, interpretable, and adaptive, which in return will enable artists to more effectively explore how AI can extend their personal expression.

1. INTRODUCTION

Songwriters are increasingly experimenting with machine learning as a way to extend their personal expression [12]. For example, in the symbolic domain, the dance punk band Yacht used MusicVAE [59], a variational autoencoder type of neural network, to find melodies hidden in between songs by interpolating between the musical lines in their back catalog, commenting that *"it took risks maybe we aren't willing to"* [46]. In the audio domain, Holly Herndon uses neural networks trained on her own voice to produce a polyphonic choir. [15, 45]. In large part, these human-AI experiences were enabled by major advances in machine learning and deep generative models [18, 65, 67], many of which can now generate coherent pieces of long-form music [17, 29, 38, 54].

Although substantial research has focused on improving the algorithmic performance of these models, much less is known about what musicians actually need when songwriting with these sophisticated models. Even when composing short, two-bar counterpoints, it can be challenging for novice musicians to partner with a deep generative model: users desire greater agency, control, and sense of authorship vis-a-vis the AI during co-creation [44].

Recently, the dramatic diversification and proliferation of these models have opened up the possibility of leveraging a much wider range of model options, for the potential creation of more complex, multi-domain, and longer-form musical pieces. Beyond using a single model trained on music within a single genre, how might humans co-create with an open-ended set of deep generative models, in a complex task setting such as songwriting?

In this paper, we conduct an in-depth study of what people need when partnering with AI to make a song. Aside from the broad appeal and universal nature of songs, songwriting is a particularly compelling lens through which to study human-AI co-creation, because it typically involves creating and interleaving music in multiple mediums, including text (lyrics), music (melody, harmony, etc), and audio. This unique conglomeration of moving parts introduces unique challenges surrounding human-AI co-creation that are worthy of deeper investigation.

As a probe to understand and identify human-AI co-creation needs, we conducted a survey during a large-scale human-AI songwriting contest, in which 13 teams (61 participants) with mixed (musician-developer) skill sets were invited to create a 3-minute song, using whatever AI algorithms they preferred. Through an in-depth analysis of survey results, we present findings on what users needed when co-creating a song using AI, what challenges they faced when songwriting with AI, and what strategies they leveraged to overcome some of these challenges.

We discovered that, rather than using large, end-to-end models, teams almost always resorted to breaking down their musical goals into smaller components, leveraging a wide combination of smaller generative models and recombining them in complex ways to achieve their creative goals. Although some teams engaged in active co-creation with the model, many leveraged a more extreme, multi-

stage approach of first generating a voluminous quantity of musical snippets from models, before painstakingly curating them manually post-hoc. Ultimately, use of AI substantially changed how users iterate during the creative process, imposing a myriad of additional model-centric iteration loops and side tasks that needed to be executed alongside the creative process. Finally, we contribute recommendations for future AI music techniques to better place them in the music co-creativity context.

In sum, this paper makes the following contributions:

- A description of common patterns these teams used when songwriting with a diverse, open-ended set of deep generative models.
- An analysis of the key challenges people faced when attempting to express their songwriting goals through AI, and the strategies they used in an attempt to circumvent these AI limitations.
- Implications and recommendations for how to better design human-AI systems to empower users when songwriting with AI.

2. RELATED WORK

Recent advances in AI, especially in deep generative models, have renewed interested in how AI can support mixed-initiative creative interfaces [16] to fuel human-AI co-creation [27]. *Mixed initiative* [32] means designing interfaces where a human and an AI system can each "take the initiative" in making decisions. *Co-creative* [43] in this context means humans and AI working in partnership to produce novel content. For example, an AI might add to a user's drawing [20, 49], alternate writing sentences of a story with a human [14], or auto-complete missing parts of a user's music composition [4, 33, 37, 44].

Within the music domain, there has been a long history of using AI techniques to model music composition [22, 30, 52, 53], by assisting in composing counterpoint [21, 31], harmonizing melodies [13, 42, 50], more general infilling [28, 34, 40, 60], exploring more adventurous chord progressions [26, 36, 48], semantic controls to music and sound generation [11, 23, 35, 62], building new instruments through custom mappings [24], and enabling experimentation in all facets of symbolic and acoustic manipulation of the musical score and sound [3, 7].

More recently, a proliferation of modern deep learning techniques [8, 9] has enabled models capable of generating full scores [39], or producing music that is coherent to both local and distant regions of music [38, 54]. The popular song form has also been an active area of research to tackle modeling challenges such as hierarchical and multi-track generation [51, 56, 63, 69].

Despite significant progress in deep generative models for music-making, there has been relatively little research examining how humans interact with this new class of algorithms during co-creation. A recent study on this topic [44] found that deep learning model output can feel non-deterministic to end-users, making it difficult for users to steer the AI and express their creative goals. Recent work has also found that users desire to retain a certain level of creative freedom when composing music with AI [25, 44, 64], and that semantically meaningful controls can significantly increase human sense of agency and creative authorship when co-creating with AI [44]. While much of prior work examines human needs in the context of co-creating with a single tool, we expand on this emerging body of literature by investigating how people assemble a broad and open-ended set of real-world models, data sets, and technology when songwriting with AI.

3. METHOD AND PARTICIPANTS

The research was conducted during the first three months of 2020, at the AI Song Contest organized by VPRO [66]. The contest was announced at ISMIR in November 2019, with an open call for participation. Teams were invited to create songs using any artificial intelligence technology of their choice. The songs were required to be under 3 minutes long, with the final output being an audio file of the song. At the end, participants reflected on their experience by completing a survey. Researchers obtained consent from teams to use the survey data in publications. The survey consisted of questions to probe how teams used AI in their creative process:

- How did teams decide which aspects of the song to use AI and which to be composed by musicians by hand? What were the trade-offs?
- How did teams develop their AI system? How did teams incorporate their AI system into their workflow and generated material into their song?

In total, 13 teams (61 people) participated in the study. The teams ranged in size from 2 to 15 (median=4). Nearly three fourths of the teams had 1 to 2 experienced musicians. A majority of teams had members with a dual background in music and technology: 5 teams had 3 to 6 members each with this background, and 3 teams had 1 to 2 members. We conducted an inductive thematic analysis [2, 5, 6] on the survey results to identify and better understand patterns and themes found in the teams' survey responses. One researcher reviewed the survey data, identifying important sections of text, and two researchers collaboratively examined relationships between these sections, iteratively converging on a set of themes.

4. HOW DID TEAMS CO-CREATE WITH AI?

The vast majority of teams broke down song composition into smaller modules, using multiple smaller models that align with the musical building blocks of a song, before combining their results: *"So my workflow was to build the song, section by section, instrument by instrument, to assemble the generated notes within each section to form a coherent whole"* (T12). A few teams first attempted to use end-to-end models to generate the whole song at once, such as through adversarial learning from a corpus of pop song audio files (T6) or through generating an audio track using SampleRNN [47] (T13). However, they

Music building blocks	Models & techniques
Lyrics	GPT2, LSTM, Transformer
Melody	CharRNN, SampleRNN, LSTM + CNN, WaveNet + LSTM, GAN, Markov model
Harmony	LSTM, RNN autoencoder, GAN, Markov model
Bassline	LSTM + CNN, WaveNet + LSTM, GAN
Drums	DrumRNN, Neural Drum Machine, SampleRNN, Markov model
Multi-part	MusicVAE trio (melody, bass, drums), MiniVAE trio, Coconet/Coucou (4-part counterpoint), MusicAutobot (melody, accompaniment), Transformer (full arrangement)
Structure	Markov model
Vocal synthesis	WaveNet, SampleRNN, Vocaloid, Sinsy, Mellotron, Emvoice, Vocaloid, custom vocal assistant
Instrument synthesis	SampleRNN, WaveGAN, DDSP

Table 1. Overview of musical building blocks used by teams.

quickly learned that they were unable to control the model or produce professional quality songs, and thus turned to the modular approach instead. In the following sections, we summarize how teams used modular building blocks, combined and curated them, and in some cases more actively co-created with AI to iterate on the outcomes.

4.1 Leveraging modular musical building blocks

Overall, teams leveraged a wide range of models for musical components such as lyrics, (vocal) melody, harmony, bassline, drums, arrangement, and vocal and instrument synthesis. Table 1 shows an overview of models used for each song component, and Figure 1 illustrates how the 13 teams co-created with AI along these different components. The number of unique model types used by teams ranged from 1 to 6 (median 4). Some teams used the same type of model for modeling different song components.

All teams used AI to generate lyrics and melodies, and more than half of the teams synthesized vocals for parts of the song. Some of these decisions were due to model availability and existing levels of model performance. For example, teams almost always generated lyrics using AI because high-performing models like GPT-2 [55] along with a fine-tuning script were readily available.

Teams with professional musicians often chose to only generate lyrics and melodic lines, in order to leave enough creative space to musicians to decide how the various lines can be put together and to arrange the song in their own style (T5, 8). One exception is T3 who generated a lead sheet with lyrics, melody and harmony.

Teams with more ML and less music expertise opt for minimal arrangements (T10, 6, 9), and often used multi-part models because they could generate melody, harmony, bass, drums together in a coherent way, providing teams with larger building blocks to work with. In one extreme case, the team was able to generate all sections of their song by traversing the latent space of MusicVAE (T9) (see "Bridging sections through latents" below for more detail).

4.2 Combining building blocks

Teams leveraged many strategies for combining outputs of smaller models, piecing together the musical building blocks to form a coherent song. These ranged from manually combining individual components, to using heuristics to automatically pair them up, to creating a pipeline between models, to adopting models that jointly model multiple components at once.

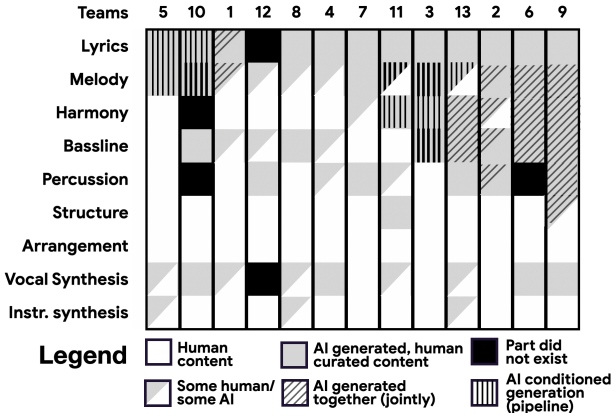

Figure 1. An overview of how 13 teams co-created with AI in songwriting. Each column shows whether each song's component was musician composed, AI generated then human curated, or both. Nearly all teams manually aligned AI generated *lyrics* and *melody*, except teams in the first three columns. T5 used a two-stage pipeline by first generating *lyrics* and *melody* separately, then algorithmically matching them up using their stress patterns. T10 first generated *lyrics*, and then conditioned on *lyrics* to generate *melody*. T1 jointly modeled *lyrics* and *melody*, generating syllables and notes in an interleaving fashion. T12, 8, 4, 7 all generated *melodic lines* first, and then manually **stitched** them together by layering them vertically as *melody* and *bassline* to yield *harmony*. In contrast, T11, 3 first generated *chords*, then conditioned on *chords* to generate *melody* (and *bassline* separately) in a **pipeline**. T13 iterated between conditioning on *melody* to jointly generate the other parts and vice versa. T2, 6 and 9 focused on models that could **jointly** generate multiple parts at once.

Stitching Many teams manually "stitched" together machine generated material, with the result informing the manual composition of other musical components. In

one team, a musician *"selected, tweaked, and weaved AI-generated melody, chords and drum parts into a ballad song form"*, while another musician wrote the bassline *"that seemed more or less implied by the AI materials"* (T7). This is echoed by another team, who composed the accompaniment *"based on chordal movements predicted by the melodic fragments"* (T5). Several teams layered melodic lines to yield harmony (T1, 12, 8, 4).

Pipelines Several teams leveraged model pipe-lining, feeding the output of one model into the input of another model. To generate melody that aligns well with lyrics, one team first used GPT-2 to generate lyrics, then a lyric-conditioned model to generate melody [68] (T10). One team decomposed melody generation into two steps, first using a LSTM to generate rhythm as note onset patterns, and then a CNN to assign a pitch to each onset (T8). While many teams "stitched" together melodic lines to create harmony, two teams first generated chords and then melody (and bassline) (T11 and T13). Pipeline approaches allowed teams to refine the output at each intermediate stage before passing content into the next model.

Joint modeling To generate multiple parts together, several teams adopted models such as MusicVAE trio [56], Coconet [34], MusicAutobot [61] or Transformers that are trained to jointly model multiple instrumental parts (T13, 2, 6, 9). One team experimented with jointly modeling notes and syllables from pairs of melodies and lyrics, but found it *"very hard to concurrently generate semantically meaningful lyrics and a melody that is not aimless"* (T1).

4.3 Generate then curate

A common approach was to generate a large quantity of musical samples, followed by automatically or manually curating them post-hoc. Teams took a range of approaches to curating the large quantity of results, ranging from brute-force manual curation, to a two-stage process of first filtering with AI, then curating manually.

Generation Often, teams used models to generate a large volume of samples to choose from. For instance, one team used their pipeline LSTM + CNN model to generate over 450 melodies and basslines (T8). Another team generated 10K lines of death metal lyrics (T13).

Manual curation While curating, teams were often looking for the key musical themes or motifs to use for their song. For example, one team used MusicVAE to generate several combinations of lead, bass, and drums, and *"handpicked the most appealing"* version to serve as their verse (T2). Another team was looking for a small, catchy snippet or *"earworms"* to flesh out the music (T11).

Two-stage curation A few teams first used automated methods to narrow down the choices, before manually selecting what would fit best in their song. For example, one team used a *"catchiness"* classifier trained on rankings of songs to filter melodies before handing them to an artist (T8). Another team curated their generated material by first algorithmically matching up the stress patterns of lyrics and melodies to make sure the material could be more immediately useful (T5).

Several teams found the process of generating and curating painstaking, or similar to a difficult puzzle (T1). However, one team described this massive generation process as exhilarating, or like *"raging with the machines"* (T5). They most appreciated the unexpected surprises, and actively engaged with this firehose of raw, AI-generated material: *"We couldn't resist including as much of the good and quirky machine output as possible...makes it much less repetitive than much of the music we might produce as humans. We really enjoyed having this firehose of generative capability...its constantly moving nature"* (T5).

4.4 Active co-creation

Some teams co-created music with AI in a more blended manner, where the model outputs influenced human composing and vice versa, similar to how human musicians might work together to write a song.

A few teams used AI-generated output as an underlying foundation for composing on top of, such as improvising a melody over AI-generated chords: *"we played the chords, and all of us around the table hummed along until we got to a simple and catchy melody"* (T11). Others took the AI output as raw material generated in a predefined structure, and manually composed an underlying beat (T8).

Others took AI output as an initial seed for inspiration and further elaboration. For example, one team trained SampleRNN on acappellas, which generated nonsensical output similar to babbling. A musician then tried to "transcribe" the words and the melody, and sang along with it. *"She found sections that sounded like lyrics. She wrote those lyrics down and sang them. She spent a day riffing on those lyrics, building a dialogue"*. These riffs and words fueled the formation of the larger story and song (T13).

For one participant, working with the AI was like jamming with another musician, an active back and forth process of priming the AI with chord progressions, hearing AI output, riffing on that output, then using those new ideas to seed the AI again (T12).

One team described making some deliberate decisions about how much agency to provide the AI vs. themselves as artists. To preserve the AI's content, an artist tried to only transpose and not "mute" any of the notes in two-bar AI-generated sequences, as he chose which ones to stitch together to align with lyrics. However, to bring the artist's own signature as a rapper, he decided to compose his own beat, and also improvise on top of the given melodies freely with a two-syllable word that was made up by ML (T8).

5. TEAMS OVERCOMING AI LIMITATIONS

In the previous section, we described the ways in which participants co-created with AI. Although teams made some headway by breaking down the composition process into smaller models and building blocks, we observed a wide range of deeper challenges when they attempted to control the co-creation process using this plethora of models. In this section, we describe participants' creative coping strategies given these challenges, and the strategies

they used to better direct the co-creation process.

5.1 ML is not directly steerable

Due to the stochastic nature of ML models, their output can be unpredictable. During song creation, teams leveraged a wide range of strategies in an attempt to influence model output, such as through data during fine-tuning, or through input or conditioning signals during generation. Below, we describe the most common patterns observed:

Fine-tuning After experimenting with a model, many teams tried to influence the mood or style of the generated content by fine-tuning models on a smaller dataset. For example, teams fine-tuned GPT-2 with lyrics from uplifting pop songs to German death metal (T13), in order to steer the generation towards phrase structures more common in lyrics and also sub-phrases that reflect the desired sentiment and style.

Priming While co-creating, teams often desired to create musical content in a particular key, chord, contour, or pitch range. To do so, many attempted to reverse engineer the process by priming the model's input sequence using music containing the desired property, in the hopes that it would produce a continuation with similar characteristics: *"I found that I could control the generation process by using different kinds of input. So if I wanted a sequence that is going upward, I would give that kind of input. If I wanted the model to play certain notes, I would add those notes in the input pattern"* (T12). This seemingly simple way of requesting for continuations led to a wide range of controls when used creatively.

To further direct content in lyrics, a team used specific words to start off each sentence (T5). Another team wondered *"can pop songs convey insightful messages with only two words?"*, and put together a verse with frequent bigrams from that dataset, such as *"my love"*, *"your heart"*. They entered this verse through the TalkToTransformer [41] web-app interface as context for GPT-2 to generate the next verses (T11).

Interpolating Models such as MusicVAE provide a continuous latent space that supports interpolating between existing sequences to generate new sequences that bear features of both [58]. Some teams leveraged this latent space as a way to explore. For example, one participant found by chance that *"interpolating between really simple sequences at high temperatures would end up giving me these really cool baseline sequences"* (T12).

5.2 ML is not structure-aware

Because large, end-to-end models were not easily decomposable into meaningful musical structures, most teams used multiple smaller, independent models. Yet, these smaller, independent ML models are not able to take into account the holistic context of the song when generating local snippets. To address this, users created their own workarounds to fill in this contextual gap, by arranging and weaving these independent pieces into a coherent whole.

Creating an overall song structure To create a backbone for the song, some teams used their musical knowledge to first curate chord progressions that can serve well for each song section (i.e. verse, chorus). One team then used a conditioned model (pretrained to be conditioned on chord progressions) to generate melody and basslines that would go well with those chords (T3).

Creating contrast between sections The verse and chorus sections of a song often carry contrast. However, participants did not have a direct way to express this desire for structured contrast while preserving overall coherence between verse and chorus. To address this, one team used their verse as a starting point to generate various continuations to the melody and bass lines, and finally chose a variation for the chorus section (T2). Another team used similar priming approaches but used different temperatures to generate the verse and chorus in order to *"add some randomness into the generation"* (T12). These approaches gave users a way to manually create structured variety.

Rewriting to add variation Rewriting allows one to generate new material while borrowing some structure from the original material. For example, one team was able to generate a *"darker version"* of the chorus of another song by rewriting it multiple times, alternating between regenerating the melody conditioned on the accompaniment, and then re-generating the accompaniment conditioned on the melody. To create a coherent song structure, the team initially attempted to *"repeat the same rave section twice"*, but later *"realized that was boring"*. The team then decided to vary how the second section ended by reharmonizing the melody with a new flavor: *"Coconet is great at reharmonizing melodies with a baroque flavor. We entered in the notes from our rave chorus. After a few tries, it made this awesome extended cadence"* (T13).

Bridging sections through latents One team devised an unusual strategy for connecting different sections of a song in a meaningful way (T9). They first trained multiple miniVAEs [19], one for each section of the song (e.g. intro, verse, chorus or genre such as rock and pop). They then composed the song by computing the *"shortest path"* through these latent spaces, allowing the sections to share elements in common with each other, despite each being distinctive. The genre miniVAEs also made style transfer possible by interpolating an existing trio towards their latent areas, allowing them to tweak the style of each section.

5.3 ML setup can interfere with musical goals

A logistical hurdle faced by teams was the significant setup and customization issues encountered to even start composing with a model. Aside from musical considerations, many teams chose or abandoned models based on what was available out-of-the-box, such as pre-trained checkpoints, fine-tuning scripts, scripts for retraining on a different dataset, data pre-processing scripts. In addition, different models expected different types of music representation as input (e.g. MIDI, piano roll, custom), adding additional pre-processing overhead to begin experimenting with them. To decrease time spent model wrangling, large teams sometimes divide-and-conquered, exploring several models in parallel to see which worked best. Ultimately,

teams' co-creation process involved navigating not only their musical goals, but also the logistical costs and benefits of using one model or another.

6. DISCUSSION

6.1 Decomposeable and context-aware modeling

Writing a song involves composing multiple sections, and multiple vocal and instrumental parts that all need to work well together as a whole. Because end-to-end models lacked meaningful hierarchical structures, and because smaller models lacked global awareness, participants often needed to reverse engineer a multitude of models, applying heuristics and domain knowledge to approximate high-level structure or to achieve desired musical effects. To ameliorate this process, one approach could be to infuse smaller models with more context-awareness, and exposing the common ways that they can be customized through user-facing controls. For example, a melody model could allow users to express whether they are creating a verse as opposed to a chorus, or whether they would like it to contrast with the next section. Another possibility is to design end-to-end models to have intermediate representations that align with the musical objects that musicians already know how to work with. The sweet spot is probably a hybrid that combines the flexibility of smaller models with the benefits of global context in end-to-end modeling.

6.2 AI-defined vs. User-defined building blocks

The design of ML models for music involves a series of upstream decisions that can have a large impact on how musicians think about music when they co-create with these models downstream. Whereas regular songwriting often starts with an initial spark of a human idea, in this work we found that the practical availability and limitations of AI tools were instead key drivers of the creative process, defining the scope of what's musically legitimate or possible. For example, most teams broke down songwriting into lyrics, melody, chords, etc., in part because these models were readily available. Yet, there are also other music building blocks that do not have corresponding, ready-made generative modsels (e.g. motif, verse, chorus, coda, etc.) or that currently are not treated as separate, first-class building blocks in deep models (e.g. rhythm, pitch). Likewise, a musician's creative thought process can be unintentionally influenced by the order of steps taken to fine-tune, condition, prime, and apply a model. In the future, the design of ML models should be coupled with a more careful consideration of what workflows and building blocks end-users already use in their existing practice, and perhaps start with those as first-class principles in guiding the design of AI systems.

6.3 Support for parallel music and ML exploration

A central aspect of the creative process involves a "flare and focus" [10] cycle of ideating, exploring those ideas, selecting ideas, then rapidly iterating. We found that a key challenge of human-AI co-creation was the need to juggle not only this *creative* process, but also the *technological* processes imposed by the idiosyncrasies and lack of steerability of learning algorithms. For instance, while ideating motifs for a song, participants needed to carry out a large additional burden of sub-tasks, such as selecting which combination of models to use, re-training or conditioning them as necessary, chaining them together, and "gluing" their outputs together. In essence, the typical "flare and focus" cycles of creativity were compounded with a parallel cycle of having to first explore and curate a wide range of models and model outputs (Figure 2). While some of these model-wrangling tasks led to new inspiration, many interfered with the rapid-iteration cycle of creativity.

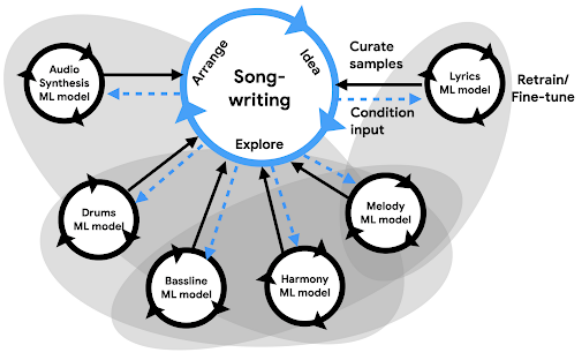

Figure 2. Parallel music and ML feedback loops in human-AI co-creative songwriting.

These issues raise important questions around how best to support users in juggling the dual processes of creative and technological iteration cycles. One approach is to have ML models readily available to musicians in their natural workflows. For example, Magenta Studio [57] makes available a suite of model plug-ins to music production software such as Ableton Live [1], and Cococo [44] allows users to semantically steer a model directly in its user interface. Beyond this, human-AI interfaces could scaffold the *strategic* part of the model exploration and selection process by surfacing effective model combinations (e.g. using general infilling models for rewriting or to reharmonize another generated melody) or fruitful workflows (e.g. matching lyric and melody stress patterns), so that new users can benefit from past users' experiences. Reducing this overhead of model-based decisions could empower users to more easily prototype their creative ideas, accelerating the feedback and ideation cycle.

7. CONCLUSION

We conducted an in-depth examination of how people leverage modern-day deep generative models to co-create songs with AI. We found that participants leveraged a wide range of workarounds and strategies to steer and assemble a conglomeration of models towards their creative goals. These findings have important implications for how human-AI systems can be better designed to support complex co-creation tasks like songwriting, paving the way towards more fruitful human-AI partnerships.

8. ACKNOWLEDGEMENTS

We thank Karen van Dijk and VPRO for organizing the AI Song Contest, and also NPO Innovatie, 3FM and EBU. We want to thank all participants for their insights and contributions. We also thank Tim Cooijmans for creating early versions of the figures and Michael Terry for feedback on this manuscript.

9. REFERENCES

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

[2] Anne Adams, Peter Lunt, and Paul Cairns. A qualititative approach to hci research. 2008.

[3] Carlos Agon, Gérard Assayag, and Jean Bresson. *The OM Composer's Book 1*. Editions Delatour France / Ircam-Centre Pompidou, 2006.

[4] Théis Bazin and Gaëtan Hadjeres. Nonoto: A model-agnostic web interface for interactive music composition by inpainting. *ICCC*, 2019.

[5] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[6] Virginia Braun and Victoria Clarke. What can "thematic analysis" offer health and wellbeing researchers? *International journal of qualitative studies on health and well-being*, 9, 2014.

[7] Jean Bresson, Carlos Agon, and Gérard Assayag. *The OM Composer's Book 2*. Editions Delatour France / Ircam-Centre Pompidou, 2008.

[8] Jean-Pierre Briot. From artificial neural networks to deep learning for music generation – history, concepts and trends, 2020.

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

[10] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.

[11] Mark Brozier Cartwright and Bryan Pardo. Social-eq: Crowdsourcing an equalization descriptor map. In *ISMIR*, pages 395–400, 2013.

[12] Barbican Centre. 12 songs created by ai how musicians are already embracing new technologies. https://artsandculture.google.com/theme/12-songs-created-by-ai/FwJibAD7QslgLA. Accessed: 2020-05-03.

[13] Ching-Hua Chuan and Elaine Chew. A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*.

[14] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 2018.

[15] Jack Denton. Future 25: Holly herndon, artist and inventor of spawn. https://www.rollingstone.com/music/music-features/future-25-holly-herndon-889072/. Accessed: 2020-05-03.

[16] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. Mixed-initiative creative interfaces. In *CHI 2017 Extended Abstracts*.

[17] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[18] Sander Dieleman, Aaron van den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. In *Advances in Neural Information Processing Systems*.

[19] Monica Dinculescu, Jesse Engel, and Adam Roberts. Midime: Personalizing a MusicVAE model with user data. 2019.

[20] Judith E Fan, Monica Dinculescu, and David Ha. collabdraw: An environment for collaborative sketching with an artificial agent. In *Proceedings of the 2019 on Creativity and Cognition*.

[21] Mary Farbood and Bernd Schöner. Analysis and synthesis of Palestrina-style counterpoint using markov chains. In *Proceedings of the International Computer Music Conference*, 2001.

[22] Jose D Fernández and Francisco Vico. AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, 48.

[23] Lucas Ferreira and Jim Whitehead. Learning to generate music with sentiment. In *ISMIR*, 2019.

[24] Rebecca Anne Fiebrink. Real-time human interaction with supervised learning algorithms for music composition and performance. *PhD dissertation, Princeton University*, 2011.

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

[26] Satoru Fukayama, Kazuyoshi Yoshii, and Masataka Goto. Chord-sequence-factory: A chord arrangement system modifying factorized chord sequence probabilities. 2013.

[27] Werner Geyer, Lydia B. Chilton, Ranjitha Kumar, and Adam Tauman Kalai. Workshop on human-AI co-creation with generative models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 2020.

[28] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. DeepBach: a steerable model for Bach chorales generation. In *International Conference on Machine Learning*, pages 1362–1371, 2017.

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

[30] Dorien Herremans, Ching-Hua Chuan, and Elaine Chew. A functional taxonomy of music generation systems. *ACM Computing Surveys (CSUR)*, 50(5), 2017.

[31] Dorien Herremans and Kenneth Sörensen. Composing fifth species counterpoint music with a variable neighborhood search algorithm. *Expert systems with applications*, 40(16):6427–6437, 2013.

[32] Eric Horvitz. Principles of mixed-initiative user interfaces. In *SIGCHI Conference on Human Factors in Computing Systems*.

[33] Cheng-Zhi Anna Huang, Sherol Chen, Mark Nelson, and Doug Eck. Mixed-initiative generation of multi-channel sequential structures. In *International Conference on Learning Representations Workshop Track*, 2018.

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

[35] Cheng-Zhi Anna Huang, David Duvenaud, Kenneth C. Arnold, Brenton Partridge, Josiah W. Oberholtzer, and Krzysztof Z. Gajos. Active learning of intuitive control knobs for synthesizers using gaussian processes. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI, 2014.

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

[37] Cheng-Zhi Anna Huang, Curtis Hawthorne, Adam Roberts, Monica Dinculescu, James Wexler, Leon Hong, and Jacob Howcroft. The Bach Doodle: Approachable music composition with machine learning at scale. *ISMIR*, 2019.

[38] 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. Music transformer. In *International Conference on Learning Representations*, 2019.

[39] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Generating music with rhythm and harmony. *arXiv preprint arXiv:2002.00212*, 2020.

[40] Daphne Ippolito, Anna Huang, Curtis Hawthorne, and Douglas Eck. Infilling piano performances. In *NIPS Workshop on Machine Learning for Creativity and Design*, 2018.

[41] Adam King. Talk to Transformer. `https://talktotransformer.com/`. Accessed: 2020-05-03.

[42] Hendrik Vincent Koops, José Pedro Magalhães, and W. Bas de Haas. A functional approach to automatic melody harmonisation. In *Proceedings of the First ACM SIGPLAN Workshop on Functional Art, Music, Modeling Design*. Association for Computing Machinery, 2013.

[43] Antonios Liapis, Georgios N. Yannakakis, Constantine Alexopoulos, and Phil Lopes. Can computers foster human users' creativity? Theory and praxis of mixed-initiative co-creativity. *Digital Culture & Education*, 8(2):136–153, 2016.

[44] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 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*, 2020.

[45] Jonathan Maas. Holly herndon: We are ai. `https://www.vprobroadcast.com/titles/ai-songcontest/articles/we-are-ai.html`. Accessed: 2020-05-03.

[46] Nathan Mattise. How yacht used machine learning to create their new album. `https://www.wired.com/story/how-yacht-used-machine-learning-to-create-their-new-album/`. Accessed: 2020-05-03.

[47] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[48] Eric Nichols, Dan Morris, and Sumit Basu. Data-driven exploration of musical chord sequences. In *Proceedings of the international conference on Intelligent user interfaces*, 2009.

[49] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 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*, CHI '18, 2018.

[50] François Pachet and Pierre Roy. Musical harmonization with constraints: A survey. *Constraints*, 6(1):7–19, 2001.

[51] Alexandre Papadopoulos, Pierre Roy, and François Pachet. Assisted lead sheet composition using flowcomposer. In *International Conference on Principles and Practice of Constraint Programming*. Springer, 2016.

[52] George Papadopoulos and Geraint Wiggins. AI methods for algorithmic composition: A survey, a critical view and future prospects. In *AISB Symposium on Musical Creativity*, volume 124, 1999.

[53] Philippe Pasquier, Arne Eigenfeldt, Oliver Bown, and Shlomo Dubnov. An introduction to musical metacreation. *Computers in Entertainment (CIE)*, 14(2):2, 2016.

[54] Christine Payne. Musenet. `https://openai.com/blog/musenet`, 2019. Accessed: 2020-05-04.

[55] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[56] Adam Roberts and Jesse Engel. Hierarchical variational autoencoders for music. In *Proceedings NIPS machine learning for creativity and design workshop*, 2017.

[57] Adam Roberts, Jesse Engel, Yotam Mann, Jon Gillick, Claire Kayacik, Signe Nørly, Monica Dinculescu, Carey Radebaugh, Curtis Hawthorne, and Douglas Eck. Magenta studio: Augmenting creativity with deep learning in Ableton live. 2019.

[58] Adam Roberts, Jesse Engel, Sageev Oore, and Douglas Eck. Learning latent representations of music to generate interactive musical palettes. In *IUI Workshops*, 2018.

[59] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. *ICML*, 2018.

[60] Andrew Shaw. A multitask music model with bert, transformer-xl and seq2seq. `https://towardsdatascience.com/a-multitask-music-model-with-bert-transformer-xl-and-seq2seq-3d80bd2ea08e`. Accessed: 2020-05-03.

[61] Andrew Shaw. Musicautobot. `https://musicautobot.com/`. Accessed: 2020-05-03.

[62] Ian Simon, Dan Morris, and Sumit Basu. Mysong: automatic accompaniment generation for vocal melodies. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 725–734. ACM, 2008.

[63] Ian Simon, Adam Roberts, Colin Raffel, Jesse Engel, Curtis Hawthorne, and Douglas Eck. Learning a latent space of multitrack measures. *arXiv preprint arXiv:1806.00195*, 2018.

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

[65] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[66] Karen van Dijk. AI Song Contest. `https://www.vprobroadcast.com/titles/ai-songcontest.html`. Accessed: 2020-05-03.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[68] Yi Yu and Simon Canales. Conditional LSTM-GAN for melody generation from lyrics. *arXiv preprint arXiv:1908.05551*.

[69] Yichao Zhou, Wei Chu, Sam Young, and Xin Chen. Bandnet: A neural network-based, multi-instrument Beatles-style midi music composition machine. *arXiv preprint arXiv:1812.07126*, 2018.