Co-Creativity between Music Producers and ‘Smart’ versus ‘Naive’ Generative Systems in a Melody Composition Task

Marinus van den Oever[1][2], Anna Jordanous[3], Rob Saunders[1]

[1] Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands
[2] Zooma, Leiden, The Netherlands
[3] School of Computing, University of Kent, Canterbury, Kent, UK

Abstract

Research in computational co-creativity frequently focuses on technical performance of computational systems and subjective quality of end-products. How computational co-creative systems impact the creative process of users has received less attention. This paper reports on a two-way double-blind crossover study to investigate how the creative processes of thirteen electronic music producers were impacted while interacting with two computational co-creative systems providing melody suggestions. The two systems shared a common user interface, however, the ‘smart’ co-creative system suggested melodies that expanded on producers’ input melodies, while the ‘naive’ co-creative system produced melodies unrelated to the producers’ inputs. To capture participants’ subjective experience, aspects of creativity were rated on Likert scales and further explored with semi-structured interviews. Each system’s output and producer’s intermediate melodies were compared for change (in compositions), dissimilarity (new AI-generated elements), and adoption (into melodies). Producers considered the ‘smart’ co-creative system to produce the most novel and valuable contributions to their process. Outputs from the ‘naive’ co-creative system were judged to be more dissimilar than smart expansions. Nonetheless, the changes that producers incorporated into their intermediate melodies were similar between the systems. This study suggests that co-creative interactions can stimulate the creative process by offering both related and unrelated musical suggestions.

Introduction

Computational co-creative systems have been broadly categorised into three types: creativity support tools, generative systems, and computer colleagues (Davis et al. 2015a). Computational co-creative systems have been developed to support human-computer collaboration on a range of creative tasks including drawing (Davis et al. 2015b), design (Karimi et al. 2020), games design (Yannakakis, Liapis, and Alexopoulos 2014), songwriting (Huang et al. 2020), and music improvisation (Hoffman and Weinberg 2010). But few have been developed for electronic music production (hip hop, dance, etc.). In addition, most electronic music producers use digital audio workstations (DAWs): software tools used for recording, editing, and playing back digital audio, that are yet to integrate AI technology (Davis 2022). According to Nash and Blackwell (2014), music software focuses primarily on transcribing and editing existing ideas, and not necessarily on generating inspiration. In practice, however, producers start their composition with a DAW and use it to create melodies. But current DAWs have not been designed to initiate, or help resolve creative blocks during, the composition process.

Nash and Blackwell (2014) emphasize that creativity is closely related to the transfer of ideas from the unconscious to the conscious mind, and suggest that this process can be stimulated through computational tools. In the field of design, collaborative ideation has been shown to stimulate the production of more creative ideas by exposing individuals to ideas beyond their own (Chan et al. 2017). Knotts and Collins (2020) performed a survey among music technologists, who indicated that they used tools like Magenta Studio[1] (Roberts et al. 2019) to generate ideas as a starting point for composition. Research in the use of AI in music creation has mostly focused on technical complexity (Sturm et al. 2019) and music information retrieval (Downie 2003), rather than the creative process. Some subjective user evaluations have been published (Karimi et al. 2018), and a recent study reports the (altered) experience of interactions of musicians with a keyboard player pretending to be an AI-system (Thelle and Fiebrink 2022). No studies, however, have looked at actual human-AI interactions during composition, or compared these to unaided conditions.

This paper presents a study that compared two computational co-creative music systems, a ‘smart’ system that processed user input and a ‘naive’ generator that did not. The naive generator was used as a comparator to test the assumption that any musical proposition might be helpful when a producer is in need of suggestions, regardless of how “smart” the generator is. The main hypothesis was that the smart generator will provide more valuable suggestions that are more readily incorporated into the composition; whereas the naive generator’s proposals may be considered novel and surprising, but less useful.

Method

A randomized crossover study was performed in which participants were assigned to two conditions across two consec-

[1] `https://magenta.tensorflow.org/studio/`

utive sessions in double-blinded random order: co-creating with a ‘smart’ system and with a ‘naive’ system. This robust experimental setup corrects for inter-individual differences in experimental variables (e.g., comprehension of instructions, use preferences, compositional approaches, and interpretation of questionnaires and terminologies). Both systems were presented as tools to expand on a MIDI file provided by the participant. The ‘smart’ system considered the input when generating its expansion, whereas the ‘naive’ system did not. For each session of this study, participants were asked to create two melodies using their preferred DAW, while actively collaborating with one of the systems. The participants were unaware that only one of the conditions actively interacted with their submitted MIDI files. Thirteen participants were recruited through social media and personal contacts. Participants were required to have experience in producing music with software, but it was unnecessary to have a degree in music.

Generative Systems

For this experiment, pre-trained applications from Magenta Studio (Roberts et al. 2019) were modified. The smart condition interacts with Continue, which uses a recurrent neural network to expand note sequences. The naive condition works with Generate, which uses a variational autoencoder to produce melodies based on music it has been trained on. To avoid unblinding subject and researcher to the condition, both systems had identical interfaces and file sizes.

Procedure

Participating producers were requested to make two compositions with the help of the smart system on one day, and the naive system on another day, in random order. Both sessions were conducted online at the participant’s home. The lead researcher and participant communicated via Zoom, which recorded the participant’s voice, screen, and computer sound. The sessions were conducted either in English or Dutch. Participants were asked to think aloud during the experiment.

The session began with an explanation of the system – how the various settings work, and how to produce, include and export MIDI files. Participants received a file to install the software. They were allowed to use their preferred DAW to work on generated outputs to minimize disruption to their usual workflow. The producer was encouraged to solicit help from the generator whenever they desired by exporting and uploading their intermediate MIDI files to the generator.

When the participant was ready, the researcher shut off the video connection and no longer interfered with the experiment, but stayed online for questions. The producer then started with the assignment to create two 8-bar melodies at a tempo of 120 beats per minute (BPM) in 40 minutes while actively collaborating with the system. Additional sounds could be added to the composition if this helped the producer to get into their flow.

After the compositional assignment, participants completed a questionnaire about their demographic information, musical expertise, and experience with the system. Using 7-point Likert scales, they indicated the extent to which they found the software’s output to be novel, valuable, and surprising. The interpretation of these terms was left to the participants, but their understanding was subsequently evaluated in a semi-structured interview, where participants further explained their answers. Additionally, they evaluated whether the software made idea generation easier, if it disrupted their creative process, and their intention to use it in their daily practice. These aspects were also explored further during the interview. Finally, participants were asked to submit their DAW project and generated MIDI. The first session lasted approximately 75 minutes and the second one hour.

Analysis

Video recordings and English transcripts were downloaded from the Zoom web portal. Dutch interviews were manually transcribed. For each participant and topic, a short summary of responses was made, including representative quotes, which were tabulated for further analysis and integration. DAW project files and MIDI files produced by the user and the generators were collected for the smart and naive systems. Project files were used to export the final melodies to MIDI. To allow comparison between the monophonic suggestions (single notes played at a time) made by the systems, and the sometimes polyphonic melodies (multiple notes played together) created by producers, these melodies were reduced to monophonic through manual extracting the top melody and truncation of overlapping notes.

MIDI files were analyzed using MIDI Toolbox (Eerola and Toiviainen 2004). The *meldistance* function was used to measure similarity between two MIDI files, on a scale from 0 to 1, based on the distribution of pitch classes (*pcdist1*). Using this measure three scores, *dissimilarity*, *adoption* and *change*, were obtained for each iteration i of the producer’s process, where $i \in 1 \ldots n-1$ and n is the number of (intermediate) melodies created by a producer. Dissimilarity, δ^i, was calculated as the average similarity between the participant’s input melody, p^i, and the system’s outputs based on this input, $G^i = \{g_1^i, \ldots, g_m^i\}$, where m is the number of outputs generated from the same input, such that $\delta^i = \frac{1}{m}\sum_{j=0}^{m} meldistance(p^i, g_j^i)$. Adoption, α^i, is the highest similarity measure between the generator’s outputs, G^i, and the producer’s next intermediate melody, p^{i+1}, such that $\alpha^i = \max_{j=0}^{m} meldistance(p^{i+1}, g_j^i)$. Change, γ^i, is the similarity between p^i and p^{i+1} such that $\gamma^i = meldistance(p^i, p^{i+1})$. For further details see van den Oever (2022).

Differences in these measures between the two systems were statistically analyzed with paired two-sided Student’s t-test and Fisher’s exact test, with a significance level of 0.05.

Results

Thirteen male electronic music producers (mean (M) age 23; range 19-30) participated in the study. On average, they had been actively composing music for 5.5 years (standard deviation (SD) 3.3; range 2-16). Five had a formal musical education. Six considered themselves amateurs, the others

	Smart	Naive	Difference	p-value
Dissimilarity	0.459	0.737	-37.7%±9.5	**0.000002**
Adoption	0.655	0.543	20.8%±15.4	**0.0219**
Change	0.318	0.384	-17.0%±16.8	0.1857

Table 1: Average scores for dissimilarity, adoption, and change for all producers are presented as proportions of altered elements (see methods section).

were professionals or semi-professionals. The mean time spent on making music was 14.2 (SD 8.7) hours per week.

Almost all experiments went smoothly without technical difficulties. Not all subjects adhered to the instructions, however this was considered an element of artistic liberty. One producer inadvertently used the naive system in both sessions. After discovery, a third session was conducted with the smart system. Results of the two naive sessions were averaged.

Analysis of Intermediate MIDI Files

The numbers of interactions with the system ranged between 2 and 10. Although participants varied their interactions considerably between sessions (from 0 to 5), the average numbers were similar for the smart generator (M±SD 5.5±2.9) and naive system (5.3±2.2). During each interaction, producers requested between 2 and 8 melody suggestions from their generator. These requests also did not differ significantly between systems (5.5±2.2 vs 4.8±1.8, difference 14.3±2.0%, p=0.230).

Dissimilarity For all participants, the average dissimilarity scores of melodies produced by the naive system was higher than for ‘smart’ melodies. The difference was highly significant (p=0.000002, Table 1). This was in line with the hypothesis that the smart generator modulates on the input and will therefore return suggestions that resemble or relate to the producer’s melody. In contrast, the naive system generates output autonomously, irrespective of the input.

Adoption It was expected that the ‘smarter’ generator would provide more useful suggestions, leading the producer to incorporate more elements of the system’s suggestions in their composition. This was the case for most producers, and the difference between the two systems was statistically significant (p=0.0219, Table 1).

Change For each interaction, the melody that was fed into the system, p^i, was compared to the (intermediate) composition made by the producer, p^{i+1}. During this complex process, producers could freely incorporate musical elements generated by the system or reject the suggestions altogether. They did this to variable degrees, to follow their own flow and inspiration, or to start with an entirely new composition. The resulting changes between p^i and p^{i+1} did not differ significantly among the two generators (Table 1).

Questionnaire and Interview

Different aspects of the interaction with the smart and the naive generators were evaluated using Likert scales and in a semi-structured interview. During the interviews, many participants made comparable comments on whether they agreed or disagreed with a certain qualification of the generator. These agreements or disagreements were scored for numerical comparisons between the two conditions, using Fisher’s exact test. The results of these numerical evaluations are presented in Table 2.

Value For both systems, participants generally agreed that the software outputs were valuable. The value of the smart generator was considered somewhat higher than for the naive system. The difference in Likert scores showed a trend in favor of the smart system (p=0.06766, Table 2), which evoked appreciative comments about value from most subjects. This contrasted significantly with the naive system, on which all participants gave at least one statement of disagreement (p=0.0272).

Smart Generator: Subjects largely agreed that the output of the system or the system itself was valuable. Participants frequently mentioned that the suggestions were easy to integrate. P13: “It was much better than I expected, I only had to change the timing of a single note, and it was perfect.” The processing of user input allowed participants to create variations of the same melody. P8: “The idea that came out of it was quite different from what I was initially going for, but it really provided like a nice bridge from, I guess, the general vibe I was trying to create.” Few disapproving comments addressed the inefficiency, as not every suggestion was equally good, requiring participants to evaluate multiple outputs. P6 describes how unfitting results can be valuable: “Even the wrong notes let you think about the possibilities.”

Naive Generator: Eight subjects stated that they found the naive system of some value. Some mentioned that the gener-

	Likert Scores (M±SD)			Participants’ Comments (n)				
	Smart	Naive		Smart		Naive		
			p-value	Agree	Disagree	Agree	Disagree	p-value
Value	5.62±1.19	4.77±1.48	0.06766	10	2	8	13	**0.0272**
Novelty	5.69±0.48	4.54±1.13	**0.00929**	11	2	7	8	0.0546
Surprise	5.54±1.33	5.04±1.56	0.40780	8	4	8	4	1
Idea Generation	5.31±1.44	4.58±1.26	0.16604	8	1	6	2	0.5765
Disruption	3.00±1.87	3.96±2.05	0.23732	6	5	9	6	0.4517
Daily practice	4.23±1.74	3.62±1.56	0.27461	10	2	9	7	0.2232

Table 2: Questionnaire Likert scores, and number of agreeing/disagreeing comments during interviews.

ator was most useful at the beginning of the process, to offer ideas to build upon. P5: "It did output things that I thought were useful and that I could use to make a new melody or composition." There were complaints that the system did not stick to the participant's key and rhythm. Nonetheless, after generating many results, or changing quite a bit, participants were still able to find something of value. P9: "It is productive if you're open to anything, or willing to push you in different directions, then it's definitely super valuable."

Novelty Participants largely agreed that the smart system was novel, and the naive system only slightly. The difference in Likert scores was highly significant (p=0.00929, Table 2). Comments also tended to be more supportive of novelty among the smart system compared to the naive system.

Smart Generator: Positive comments often mentioned how the system provided new insights, directions, and inspiration. Some participants appreciated the modesty of the changes suggested by the system. P6: "Although it was so simple and so minimal, it immediately gave me a new inspiration. Something I could have played myself, but didn't have in my mind at that time." Few negative comments addressed the fact that the system partially repeated their input.

Naive Generator: Participants agreeing with the novelty of the naive generator mainly talked about the dissimilarity of the output. P7: "It came with completely different things than what I imagined." Several negative comments also used the term 'randomness' to express dissatisfaction. P10: "It is a bit too random to get a melody out that works." Sometimes the naive system generated unrelated samples that were helpful. P7: "Something completely different came out, which I thought was very cool, and because of that I discarded my own piece."

Other categories Surprise, idea generation, disruption, and daily practice did not show significant results (Table 2), however occasionally helpful comments were made. The two passively collaborating systems were repeatedly stated to be perceived as 'fellow musicians'. P5 and P13 felt that the smart system provided the same effect as collaborating with human peers. Also working with the naive system approached similar stimulation. P11: "It's almost like having an additional musician who plays something in."

Discussion

We hypothesized that the smart system would be perceived as more valuable, but less novel and surprising compared to the naive generator; and that this would be reflected in higher adoption of 'smart' suggestions, and lower dissimilarity indices. The results show that participants considered the smart system more valuable and novel than the naive system. Other categories (surprise; ideation; disruption; daily practice) showed no noticeable differences. The participants' preference for the smart system is also evident in higher adoption, meaning that more elements were incorporated in the intermediate compositions. The smart output was less dissimilar compared to the naive output. There are two possibilities for this apparent discrepancy between higher value and lower dissimilarity. First, participants could have favored expansions that shared characteristics with their own input. Secondly, adoption could be higher because the smart generator repeated elements already present. Both possibilities may have contributed to the higher adoption index, but co-creative interplay also played an important role. Despite having the freedom to explore alternative melodic directions, participants tended to stick to their initial inputs even after interacting with the smart system.

The smart and naive systems both seemed to have similar and limited effects on the compositions: in both conditions, producers changed roughly 35% of their input melody (p^i) to make their next melody (p^{i+1}), see Table 1. The majority of the melodies were unchanged, suggesting that regardless of the system used, participants were disinclined to deviate too much from their ongoing composition. This could be one reason why the smart generator, which modulates on the producer's input melody, is considered significantly more valuable than the naive system (Table 2). Participants commented that the smart generator was most useful for progressing an existing composition. However, producers also mentioned that the smart generator offered value when their input (p^i) was only a few notes. Unexpected, in view of these relatively conservative preferences, the smart generator was judged to be more 'novel' than the naive system. This makes sense considering the large number of comments on how the smart system provided options that the participants did not think of.

Limitations of the study include the experimental setup, the generative systems, and the metrics used. Despite the careful design of the experiment and use of follow-up interviews, interpretations of novelty, value, and surprise between participants may still have varied, due to the inherent ambiguity of the terms (Grace et al. 2015), and thus affected the reliability of the analysis of outcomes between participants. The experimental setup was limited to comparing the naive and smart systems and did not include an unaided condition to establish how much producers alter intermediate compositions between iterations. In addition, the number of interactions with the system varied significantly between participants (2-10). The think-aloud protocol was found to be uninformative because participants were too preoccupied with the task to verbally reflect on the activity. Moreover, the generative systems used in the study were limited to the generation of monophonic melodies, and the similarity indices used do not capture all differences between melodies, e.g., rhythm and melodic contour. Furthermore, lack of diversity and gender inclusion restricts the generalizability of the results. Addressing these shortcomings opens up avenues for further work.

Conclusions

In this study, we investigated the creative processes of music producers who composed melodies, assisted by two computational co-creative systems that provided melody suggestions: a 'smart' system that processed user input, and a 'naive' generator that did not. We observed that two operationally identical systems had distinct but valuable interactions with the creative process. Participants particularly liked 'smart' system's expansions related to their own melodies, but also appreciated unexpected 'suggestions' from the naive system. Despite the systems' passive interaction, some participants felt a sense of collaboration, similar to working with other musicians. These findings highlight the importance of studying the human creative process *in situ* when developing systems for human-AI collaboration. Evaluating the system in isolation against static metrics of quality is insufficient to understand its place in a collaboration. Instead, we must consider how it integrates into the creative process. Moving forward, we encourage future studies to adopt similar methodologies that combine quantitative and qualitative metrics in a controlled blinded evaluation, to accelerate research progress in evaluating co-creative systems. To extend this research, we suggest including an unaided condition and exploring systems that can generate polyphonic melodies. Moreover, there is a need for psychometric instruments to quantify different aspects of creative processes.

Author Contributions

MvdO conducted the research, carried out the experiments, and drafted the paper. AJ and RS provided supervision and edited the document.

Acknowledgements

This paper is based on Marinus van den Oever's master thesis for the Media Technology MSc programme at Leiden University, The Netherlands. We would like to thank Oded Ben-Tal and Oliver Bown for their role as critics, providing valuable feedback during the defence.

References

Chan, J.; Siangliulue, P.; Qori McDonald, D.; Liu, R.; Moradinezhad, R.; Aman, S.; Solovey, E. T.; Gajos, K. Z.; and Dow, S. P. 2017. Semantically far inspirations considered harmful? accounting for cognitive states in collaborative ideation. In *Proc. of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 93–105. New York, NY, USA: Association for Computing Machinery.

Davis, N.; Hsiao, C.-P.; Popova, Y.; and Magerko, B. 2015a. An enactive model of creativity for computational collaboration and co-creation. In Zagalo, N., and Branco, P., eds., *Creativity in the Digital Age*. London, UK: Springer. 109–133.

Davis, N.; Hsiao, C.-P.; Singh, K. Y.; Li, L.; Moningi, S.; and Magerko, B. 2015b. Drawing apprentice: An enactive co-creative agent for artistic collaboration. In *Proc. of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 185–186. New York, NY, USA: Association for Computing Machinery.

Davis, D. 2022. Compute and resonate: An ongoing experiment in creating acid music using accessible artificial intelligence and computer-based generative tools. In Filimowicz, M., ed., *Designing Interactions for Music and Sound*. London, UK: Focal Press. 65–82.

Downie, J. S. 2003. Music information retrieval. *Annual Review of Information Science and Technology* 37(1):295–340.

Eerola, T., and Toiviainen, P. 2004. *MIDI Toolbox: MATLAB Tools for Music Research*. Jyväskylä, Finland: University of Jyväskylä.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3–4):125–147.

Hoffman, G., and Weinberg, G. 2010. Gesture-based human-robot jazz improvisation. In *2010 IEEE International Conference on Robotics and Automation*, 582–587.

Huang, C.-Z. A.; Koops, H. V.; Newton-Rex, E.; Dinculescu, M.; and Cai, C. J. 2020. Ai song contest: Human-ai co-creation in songwriting. In *Proc. of the 21st International Society for Music Information Retrieval Conference*, 708–716.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. In *Proc. of the 9th International Conference on Computational Creativity*, 104–111. Association for Computational Creativity.

Karimi, P.; Rezwana, J.; Siddiqui, S.; Maher, M. L.; and Dehbozorgi, N. 2020. Creative sketching partner: An analysis of human-ai co-creativity. In *Proc. of the 25th International Conference on Intelligent User Interfaces*, 221–230. New York, NY, USA: Association for Computing Machinery.

Knotts, S., and Collins, N. 2020. A survey on the uptake of music ai software. In *Proc. of the International Conference on New Interfaces for Musical Expression*, 499–504.

Nash, C., and Blackwell, A. F. 2014. Flow of creative interaction with digital music notations. In Collins, K.; Kapralos, B.; and Tessler, H., eds., *The Oxford Handbook of Interactive Audio*. New York, NY, USA: Oxford University Press. 387–404.

Roberts, A.; Engel, J.; Mann, Y.; Gillick, J.; Kayacik, C.; Nørly, S.; Dinculescu, M.; Radebaugh, C.; Hawthorne, C.; and Eck, D. 2019. Magenta studio: Augmenting creativity with deep learning in ableton live. In *Proc. of the International Workshop on Musical Metacreation*.

Sturm, B. L.; Ben-Tal, O.; Monaghan, U.; Collins, N.; Herremans, D.; Chew, E.; Hadjeres, G.; Deruty, E.; and Pachet, F. 2019. Machine learning research that matters for music creation: A case study. *Journal of New Music Research* 48(1):36–55.

Thelle, N. J. W., and Fiebrink, R. 2022. How do musicians experience jamming with a co-creative "ai"? In *NeurIPS 2022 Workshop: Machine Learning for Creativity and Design*.

van den Oever, M. 2022. Co-creativity between music producers and 'smart' versus 'naive' generative systems in a melody composition task. MSc. thesis, Leiden University. Available from: https://theses.liacs.nl/pdf/2021-2022-OevervandenMarinus.pdf.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *International Conference on Foundations of Digital Games*.