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

Marinus van den Oever
Graduation Thesis
Media Technology MSc program, Leiden University
July 2022
Thesis advisors: Rob Saunders, Anna Jordanous

Abstract

Human-AI collaboration is suggested to foster creativity in art and design. Research in computational creativity focuses mainly on technical performance and subjective quality of the end-products. It is unclear how AI-systems function when they collaborate with music producers. The current study analyzed creative processes in thirteen producers, who composed melodies with input from either a 'smart' or 'naive' AI-system. Aspects like novelty and value were rated on 7-point Likert-scales, which guided 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 system most novel and valuable. They particularly liked 'smart' expansions related to their own melodies, but occasionally also appreciated unexpected 'suggestions' from the naive system. Naive output was more dissimilar than smart expansions. Nonetheless, a significantly higher proportion of smart elements were adopted into melodies than for naive suggestions. Despite these differences, changes in intermediate melodies were similar between systems. Comparisons of interviews and generated melodies suggested that producers tended to adopt fewer AI-suggestions, if AI-output was either too similar ('same') or dissimilar ('weird') compared with their own melodies. This study indicates that music producers can benefit from AI-support, particularly if AI-suggestions fit optimally with their own melodies. Unrelated suggestions can be useful with creative blocks. These results can be used for the development of co-creative AI-plugins for DAWs, which are currently unavailable.

Introduction

Computational creativity is concerned with the development and evaluation of creative artificial systems. According to Davis et al. (2015a), this field works on three types of systems: generative systems, creativity support tools, and computer colleagues (co-creativity). In co-creativity, humans collaborate with computers on creative tasks. This has been used in a wide range of creative domains including drawing (Davis et al. 2015b), design (Karimi et al. 2020), dance (Jacob et al. 2013), songwriting (Huang et al. 2020), music improvisation (Hoffman and Weinberg 2010), and music composition (Louie et al. 2020; Suh et al. 2021). To our knowledge, co-creative systems are rarely used in contemporary music production, because most producers use digital audio workstations (DAWs) which do not yet incorporate 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, many producers start their composition with a DAW and interact with the software to generate melodies. However, these systems are not designed to initiate music compositions, and also not to resolve creative blocks that artists can experience.

Nash and Blackwell (2014) emphasize that creativity is closely related to the transfer of ideas from the unconscious to the conscious mind. They suggest that this process can be stimulated through computational tools. New music AI systems have been developed that aim to enhance human creativity, including Google Magenta, OpenAI's MuseNet and Jukedeck, Sony CSL's Flow Machines, AIVA, and Amper Music. Knotts and Collins (2020) performed a survey among music technologists, who indicated that they used tools like Google Magenta to generate ideas as a starting point for composition. It is unclear how such tools truly affect composing, as AI research tends to focus on the technical performance of the systems, rather than the creative process (Sturm et al. 2019). Some user studies have been conducted, but these are mostly based on subjective evaluations (Karimi et al. 2018). None of these studies look at the human-AI interactions during the compositional process or compared the contribution of AI to 'normal' unaided conditions.

To explore how generative systems can assist music producers in the process of composing melodies, the current study compared a 'smart' AI system that processed user input, with a 'naive' generator that provided musical content, unrelated to the user input. The naive generator is used as a 'dumb' 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 of this study was that the AI generator will provide more valuable suggestions that are more readily incorporated into the composition; whereas the naive generator's proposals may be more novel and surprising, but less useful for the producer.

Method

Study Design

A randomized crossover study was performed in which participants were assigned to two conditions across two consecutive sessions in double-blinded random order: co-creating with a 'smart' system and with a 'naive' system. Both systems expanded an input (MIDI) sequence provided by the participant. The smart generator considered this input sequence when generating its expansion, whereas the naive system did not. On each session of this study, music producers were asked to produce two melodies using their personal digital audio workstations, while actively collaborating with one of two different 'artificial intelligence systems'. The participants were unaware that only one of the conditions actively interacted with their music. The study was approved by the LIACS - Media Technology MSc Ethics Board.

Participants

Thirteen participants were recruited through social media channels and personal contacts of the researcher. Participants were required to have experience in producing music with software, but it was unnecessary to have a degree in music.

Apparatus and Procedure

Experimental Sessions 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 researcher and the participant were in contact via Zoom. Informed consent was obtained to record the participant's voice, screen, and computer sound through Zoom. Participants were asked to express their thoughts and deliberations aloud throughout the experiment.

The session began with a brief explanation of the generative system – how the various settings work, and how to produce, include and export MIDI files. Participants received a file to install the software. When the participant was ready, the researcher shut off the video connection and no longer interfered with the compositional process, but stayed online for questions or technical problems. The producer then started with the assignment to create two 8-bar melodies within 40 minutes while actively collaborating with the system. Additional sounds could be added to the composition if this helped the producer to get into his flow.

At the end of each session, the participant completed a questionnaire, followed by a brief semi-structured interview performed by the researcher. In addition, the participant was asked to submit his DAW project, generated MIDI and log files. The first session lasted approximately 75 minutes and the second one hour.

Compositional Software Producers were allowed to use their DAW of choice. Although there are differences between DAWs, they all offer the same basic functionality of recording, editing, and playing back digital audio. The producer was encouraged to solicit help from the generator whenever they felt like it, by exporting and uploading their intermediate MIDI files to the generator.

Generative Systems For this experiment, the Continue and Generate applications from Magenta Studio (Roberts et al. 2019) were modified. The smart condition interacts with Continue, which uses a recurrent neural network (RNN) to expand note sequences. The naive condition works with Generate, which uses a Variational Autoencoder (VAE) to produce melodies based on the music it has been trained on. A fake input field was added to the naive system to give the impression that it processes the producer's MIDI. To avoid unblinding of both subject and researcher to the condition, both systems were modified to have equal user interfaces and installation file sizes.

Questionnaire and Semi-structured Interview After the compositional assignment, participants completed in a questionnaire about their demographic information and musical expertise. They rated their experience of working with the system using seven-item verbal Likert scales. The propositions were that the software's output was valuable; novel; surprising; idea generating; disruptive; adoptable in daily practice. The answers to these questions were subsequently used to guide a semi-structured interview, in which participants were asked to provide further context to their answers, and to comment on technical issues, experimental setup and suggestions for features or improvements.

Data Extraction

Questionnaires and Interviews Video recordings and English transcripts were downloaded from the Zoom web portal. The recorded interviews were transcribed with the help of oTranscribe (`http://otranscribe.com/`). For each participant and topic, a short summary of responses was made, including representative quotes, which were tabulated for further analysis and integration. Verbal Likert scales were changed to numerical values: 1 'strongly disagree'; 2 'disagree'; 3 'somewhat disagree'; 4 'neither agree nor disagree'; 5 'somewhat agree'; 6 'agree'; 7 'strongly agree'.

Generative Systems Raw data were collected from the smart and naive systems, including DAW project files; MIDI files produced by the user and the generators; and the generator logfiles. The DAW project files were not routinely analyzed, but could be used as a backup to follow the compositional process if necessary.

To allow comparisons between the monophonic suggestions made by the generator systems, and the sometimes polyphonic melodies created by the producer, these melodies were reduced to monophonic melodies through manual extraction of the top melody and truncation of overlapping notes.

The MATLAB package, MIDI Toolbox by Eerola and Toiviainen (2004) was used for analyzing the MIDI data. For each generated melody, *meldistance* function was used to obtain similarity scores, scaled from 0 to 1. The melodies were compared on their distribution of pitch classes (*pcdist1*) using the taxicab distance metric. The smart and naive systems were compared for dissimilarity and adoption, using the analysis procedure illustrated in figure 1.

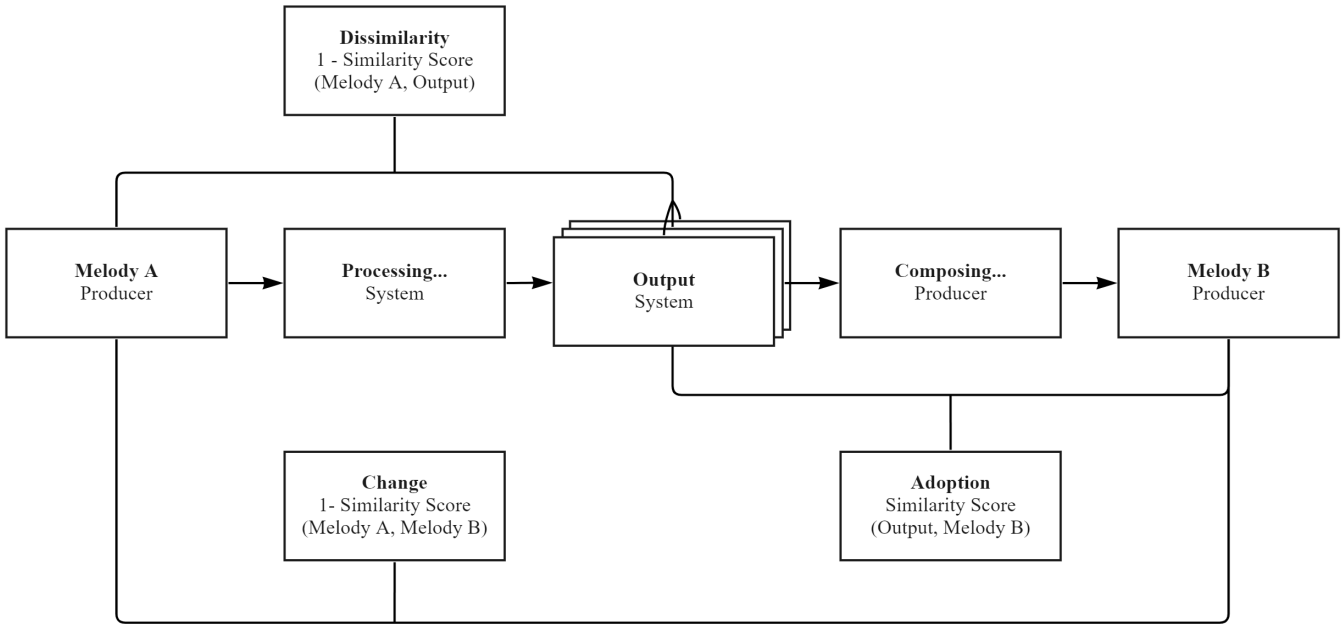


Figure 1: Analysis scheme.

Statistical Analysis After data extraction and cleanup, differences between the two systems for numerical outcomes of the experiment (Likert scores and numbers of agreement/disagreements for the questionnaires; dissimilarity-, adoption- and change-indices for the generators) were statistically analyzed with paired two-sided Student's t-test and Fisher's exact test, with a significance level of 0.05. Since this was an exploratory study, no corrections for multiple comparisons were made. For statistically significant results, patterns and relationships were plotted graphically for interpretation and secondary exploration.

Results

Demographic Characteristics

Thirteen music producers participated in the study. All subjects were males, with a mean (M) age of 23 years (range 19-30 years). On average, they had been actively composing music for 5.5 years (standard deviation (SD) 3.3 years; range 2-16 years). Eleven of the subjects played different musical instruments, on average three and mostly including piano. Five had a formal musical education. Six considered themselves amateurs, the others were (semi)professionals. The mean time spent on making music was 14.2 (SD 8.7) hours per week. Three of the participants had prior experience with AI programs in music.

Analysis of Intermediate MIDI Files

Almost all experiments went smoothly, without technical difficulties. One of the producers inadvertedly installed the same (naive) software package twice, for each of his two planned sessions. After this was discovered, this participant made a third composition with the other (smart) system. The results of his two naive sessions were averaged.

Figure 2: Wordcloud of producers' preferred musical genres.

System Suggestions and Interactions During the session, the numbers of interactions between the producer and 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 the systems (5.5±2.2 vs 4.8±1.8, difference 14.3±2.0%, p=0.230). Beforehand, it was considered possible that the numbers of interactions and suggestions could influence the composition and would therefore have to be taken into account as covariates in the statistical analysis. Since the interactions with the two systems were very similar, more complex statistics were put aside.

Dissimilarity In all participants, the average dissimilarity scores of melodies produced by the naive system was higher than for 'smart' melodies (Figure 3). The difference was highly significant (p=0.000002, Table 1). This was in line with the hypothesis that the smart generator modulates on

	Smart (M)	Naive (M)	Difference %±SD	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: Scores for dissimilarity, adoption and change are presented as proportions of altered elements (see methods section). Multiplication by 100% will generate percentages, which are sometimes used in the article.

the input and will therefore return suggestions that resemble or relate to the producer's melody (Figure 1). In contrast, the naive system generates output autonomously, irrespective of the input.

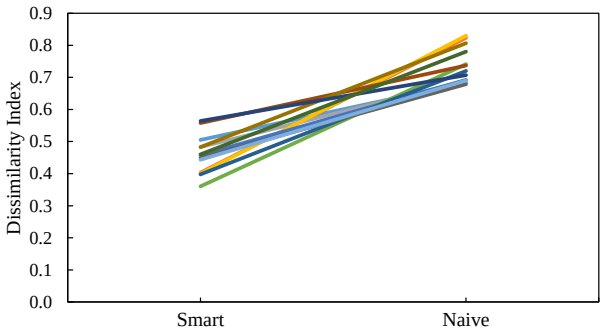


Figure 3: Dissimilarity

Adoption It was expected that the 'smarter' AI generator would provide more useful suggestions, leading the producer to incorporate more elements of the system's suggestions in his composition. As shown in Figure 4, this was the case for most producers, and the difference between the two systems was statistically significant (p=0.0219, Table 1).

Participants acted according to instructions to create an 8-bar composition. During his second session, one producer started working on a much longer composition of roughly 32 bars, which he fed entirely into the smart generator. Consequently, the suggestions that this producer adopted from the system, formed a much lower proportion of his composition (33.8%) than for his other (naive) session in which he made a much smaller melody (53.9%), or for most of the other participants (Figure 4). Although this became apparent during data processing and before unblinding, we considered this an (extreme) part of the producer's compositional freedom, and it was therefore decided to incorporate these data into the analysis unchanged.

Change For each interaction, the melody that was fed into the system (Melody A in the analysis scheme of Figure 1) was compared to the composition made by the producer (Melody B). 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 Melodies A and B did not differ significantly among the two generators (Figure 5, Table 1).

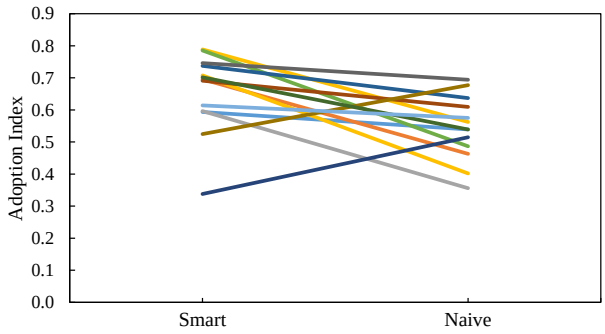


Figure 4: Adoption

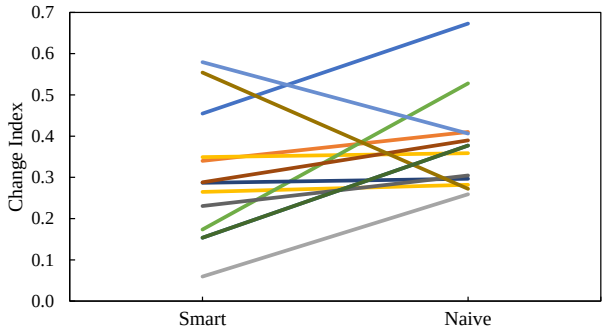


Figure 5: Change

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 questionnaire. Almost all spontaneous utterances in the think aloud recordings were repeated during the interviews. 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,

	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.

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 generator 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."

Surprise Both systems were rated almost equally surprising (Table 2). Twice as many comments expressed agreement rather than disagreement that the systems were surprising. The numbers were the same for the two generators however.

Smart Generator: Positive comments often mentioned how the generated melodies were different, although relating well to the producer's piece. P7: "I played in three notes, and what came out was a rather complex melody that sounded good and stayed in my chosen key." Negative participants described that not much changed in comparison to their input.

Naive Generator: Participants who commented that the naive system was surprising, often used terms expressing unexpectedness. The extensions were often quite different in terms of pitches, lengths and rhythms. Some found this "wackiness" interesting to work with (P11), whereas for others it was too "weird" (P6). P12: "I didn't know what to do with it."

Idea Generation Ratings for idea generation did not differ significantly between the two systems (Table 2). The number of participants who during the interview agreed were similar, because both systems were considered helpful albeit on different aspects.

Smart Generator: Participants who agreed that the system gave them new ideas, talked about how the system is most helpful at moments when they do not know how to proceed, P9: "It really helped me to get the ball rolling, rather than to be stuck." A single disagreeing participant (P12) was not open to the system's suggestions because he wanted to proceed with his own piece.

Naive Generator: There was agreement as well about how the naive system could help when a producer was stuck. The emphasis however was on using the suggestions as a starting point. P10: "Mostly in the beginning, when you can head in different directions." One participant disagreed that the naive generator added much to his own ability to come up with ideas, since he was trained as a professional musician (P5).

Disruption of Compositional Process The producers' creative processes seemed to have been slightly more disrupted by the naive than by the smart system. However, the difference in Likert scores were not statistically significant (Table 2). Opinions also differed during the interviews.

Smart Generator: Observations regarding disruption were mainly related to the fact that the software was a standalone plugin. Participants had to perform a series of manual operations, like creating folders and transferring files between the DAW and the generator. Some participants felt they were forced to actively use the tool, whereas in daily practice they would only use it a couple of times.

Naive Generator: The fact that the software is not an integrated plugin was also a problem here, but in addition the output was not always desirable. Participants mentioned how they had to generate several melodies, evaluate them one-by-one, and then still modify them, which slowed down their workflow. Subjects who disagreed that experienced much hindrance, talked about how the software could still spark ideas even when they did not incorporate the system's exact suggestions.

Suitability in Daily Practice On average, somewhat more agreement was found for the use of the smart system in daily practices compared to the naive system (Table 2). Nearly equal confirmative remarks were observed, however a large difference in disconfirming remarks present among the naive condition.

Smart Generator: During the interviews, only two producers commented that they would not use the software in their daily practice. They would rather create music on their own. Participants who would adopt the software, would primarily use it when they encounter problems. P4: "I could see myself using it when I'm stuck. However, I wouldn't know whether this would be the main instrument to my process." Participants also mentioned that it can be powerful for extending an existing piece. P5: "Adding the finishing touches, like embellishments and the melodies on top, to add some contrast in other parts."

Naive Generator: Seven participants disagreed that they would use the naive system in practice. Several emphasized that the tool can only be used at the onset of the process. For other subjects, this same argument was used to disagree with the statement. P8: "It would be hard to integrate this software when you already have an idea or genre in mind already. It's a lot more valuable if you're using it as the foundation or as the starting point, so that you can build around it, rather than using it to add to an already sort of half complete process."

Other Feedback *Smart Generator:* P5 and P13 felt that the smart system provided the same effect as collaborating with human peers. P5 describes how hearing other people's melodies can spark creativity: "Hearing someone else perform gives me ideas about what I could add to it; having a software that does that for you therefore gives you more ideas."

Naive Generator: Several subjects also described the interaction with the naive system as a collaboration with real musicians. P11: "It's almost like having an additional musician who plays something in." In this condition, however, it was more common for producers to discard their original melody and continue with the system's suggestion. P7: "I removed my part and turned the output into something new. The foundation of that melody was clearly originating from the generator." The system generated a wide range of pitches because it did not consider the producer's input, which inspired many to create a bass line from the melody, P7: "I went for a piano melody, and what I got back from the system was a bass line, which caused the top melody to turn into a bass line, which I thought was really cool."

Feature Suggestions Producers suggested that both systems could be improved by making them available as integrated plugins. A selection for beats per minute (BPM) and the scale (key) of the song was also missing. Few subjects also suggested a drum/percussion generator, since they find that aspect of production generally more difficult.

Relations between Questionnaire Evaluations and Generated MIDI Files

To better understand how producers interacted with the generators, the Likert ratings that differed significantly between the two systems ('value' and 'novelty') were related to the MIDI-indices ('adoption', 'dissimilarity' and 'change'). The emphasis was not on statistical analysis of the correlations, but on graphical illustration to facilitate interpretation and contextualization for the discussion.

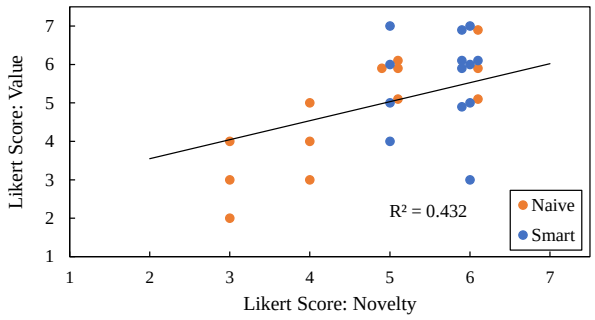
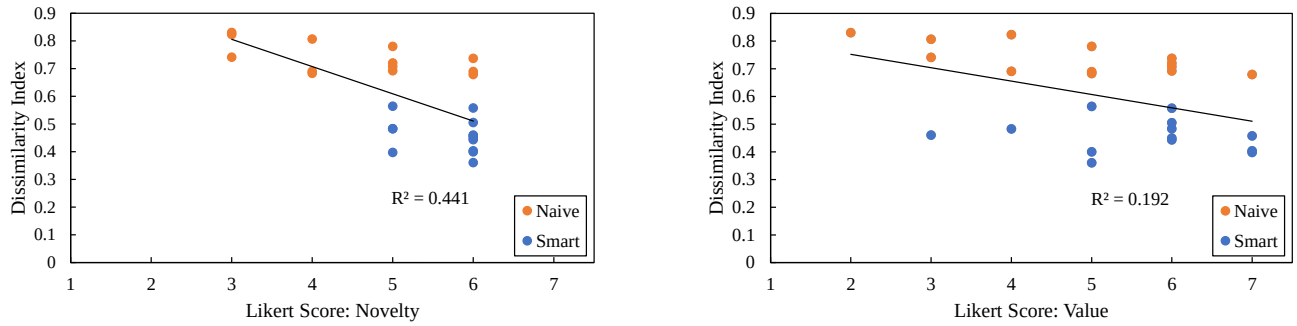


Figure 6: Novelty vs Value.

In our original study protocol, the dissimilarity index was named 'novelty', because by definition this parameter represents new elements added by the AI-system to the producer's input melody (Melody A – see analysis scheme in Figure 1). The name was changed, mainly to avoid confusion with the term 'novelty' used in the questionnaire and interview. However, the narratives revealed that the participants also understood the concept 'novelty' differently: not so much as novel musical elements, but also in terms of added value. This is illustrated in Figure 6, which shows that the numerical Likert scores for 'novelty' explains 65.7% of the variance of the 'value' scores (p= 0.000265).

Unexpectedly, the participants' Likert scores for 'novelty' showed inverse relationships with the 'dissimilarity' index produced by the AI-systems (Figure 7a). Value



(a) Novelty.

(b) Value.

Figure 7: Novelty (a) and Value (b) vs Dissimilarity.

scores showed similar negative relationships with dissociation (Figure 7b). Correlations were statistically significant for the two conditions combined, but the relationship was particularly clear for the naive system, where higher novelty and value scores were significantly associated with lower dissimilarity indices ($p = 0.033$ and $p < 0.001$, resp.; naive trendlines not shown separately).

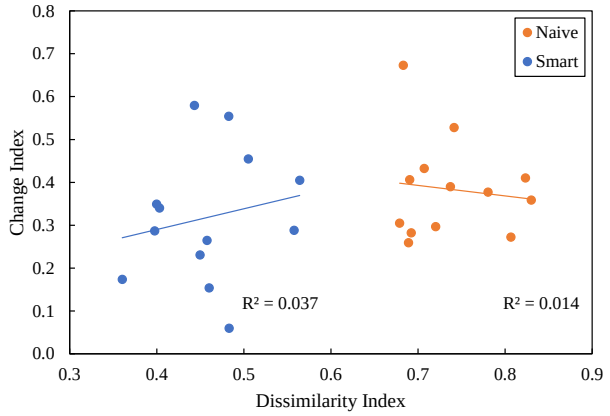


Figure 8: Dissimilarity vs Change.

Dissimilarity was not associated with any of the other Likert scores that showed no differences between the two systems (surprise; generation of new ideas; disruption of creative process; useability in daily practice).

The unexpected associations of high novelty with higher value scores and lower dissimilarity indices, and the participants' comments on these factors in the interviews, suggested a complex nonlinear association between the AI-output and the incorporations into the musical productions. This was further explored by plotting dissimilarity scores against the actual changes made in the melodies (Figure 8). This clearly shows that the unrelated output from the naive generator was much more dissimilar from the producer's input, than the modulations generated by the smart system. Moreover, the associations between dissimilarity and change seem to differ between the two systems.

Discussion

We conducted a study to investigate the creative process of music producers while they are composing melodies assisted by either a co-creative smart system or a naive generator. It was 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' AI 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) did not show 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 that were already present. It is likely both possibilities contributed to the high adoption index, but a co-creative interplay played an important part. Participants were free to move into different directions, but after smart-exposure, they still decided to continue in the same vein.

The smart and naive systems both seemed to have similar and limited effects on the compositions: in both conditions, producers changed roughly 35% of Melody A to make Melody B (Figure 1, 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. Producers also mentioned that the smart generator offered value when their input (Melody A) 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.

Producers stated that the naive system often provided unrelated yet interesting output. This was often mentioned to be most valuable at the beginning of the composition, or to start over when a producer got stuck. During the process, producers generally adopted a lower proportion of new 'naive' than 'smart' suggestions (54.3% vs 65.5%, Table 1). At the same time, the naive system produced much more dissimilar content (73.7%) than the smart generator (45.9%, Table 1). Simple multiplication of dissimilar and adopted elements would imply that about 33% more new musical suggestions were taken over from the naive output (73.7 × 54.3 = 40%) than for the 'smart' suggestions (45.9 × 65.5 = 30.1%). Unfortunately, absolute numbers of incorporated elements cannot be determined from the MIDI-file analysis. Still, our findings indicate that producers readily incorporated 'naive' music suggestions into the compositions, but they seemed to have worked differently with both systems. For the smart system, producers indicated that they highly valued the interactions with their own ideas, whereas the naive system was moderately appreciated for its unexpectedness. This was experienced by some participants as working with the systems as a collaboration. Producers commented that the interactions with the systems during the experiment approached the stimulation offered by working in a studio with other musicians. This is noteworthy because co-creativity requires that the computational actor takes an active role in the process (Davis et al. 2015a). In the current experiment, however, the two passively collaborating systems were repeatedly also stated to be perceived as 'fellow musicians'.

Both systems seem to have partially fulfilled the requirement of co-creative tools. However, there were some restrictions to both the systems and the experimental setup. The study was restricted to the production of monophonic melodies. Participants overall did not experience this as a limitation. Similarity indices (dissimilarity, adoption, change) were extracted from MIDI files, by comparing the distributions of pitch classes; the study did not take other melodic aspects into account, such as rhythm and melodic contour. Therefore, the three indices might not include all differences between the systems, in generated output or changes in compositions.

The questionnaires and semi-structured interviews provided important context for the interpretation of the system's performance. However, the questions were limited to predefined concepts that were considered relevant for the understanding of how producers interacted with different generative systems. This process was complex, however, and the controlled approach only offered limited insights. One of the objectives of the study was to record spontaneous reflections of the participants, while performing the experiment. This had previously been successful in capturing the creative process of composing (Collins 2007). This 'think-aloud' approach did not meet these expectations, mostly because participants were unused to talking to themselves, and were too absorbed with the process to verbally reflect simultaneously. Most of the recorded utterings were covered during the interviews, but important psychological aspects were undoubtedly missed.

Pre-trained models were used, which may have limited their usefulness, as the training data did not satisfy each producer's preferences. One producer never found a match for his favorite afro-beat. An option would be to pre-train the models on MIDI files provided by the participant before conducting the experiment. Real-time machine learning models may be even more appropriate. A limitation of the systems in this study is that they expand on input note sequences. It would also be useful to examine models which can make insertions or other modifications.

Several participants stated that they would use the tool less actively in their regular production sessions than was requested in the experiment. Even within the limits of instructions, the interactions with the system (2-10 times) or requested outputs (2-8 suggestions) varied widely among producers. When used freely in daily practice, they likely to differ substantially between users, and this will probably also develop over time. A longitudinal study could provide a more realistic picture on how these tools assist the process outside of a lab setting, and also elucidate which co-creative strategies will be developed. This study compared a 'smart' with a 'naive' generator, with the idea that the dumb system represented a 'sham' condition. The systems turned out to be used differently, but lead to just as much change of about 1/3 of the compositions. It is unknown how much producers normally change their intermediate compositions, and to understand how this is influenced by co-creative interactions, a comparison with an unaided 'natural' session would be useful.

Conclusions

Our study provides insights into the requirements for a co-creative tool that would suit different needs of producers during their compositional process. On the one hand, producers need a tool that modulates their compositions, stimulating their ideas and providing new directions. This feature requires AI-output that is not too similar to the producer's input melody (meaning a minimal dissimilarity index). On the other hand, producers can sometimes break through creative obstructions when the system offers an entirely new suggestion. In these situations, the dissimilarity index should be high enough to boost creativity, but not so high that it becomes too 'weird' to work with. Changes may therefore require an optimal level of dissimilarity (cf Figure 8), which should be adaptable to the producer's specific needs during the compositional process. An important aspect of a good co-creative system is an appropriate balance between 'going with the flow' and 'coming up with something new' at the right moment.

References

Collins, D. 2007. Real-time tracking of the creative music composition process. *Digital Creativity* 18:239–256.

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*, Springer Series on Cultural Computing, 109–133. Springer, London.

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 *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, CC '15, 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*, 65–82. Focal Press.

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

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. *International Society for Music Information Retrieval (ISMIR)* 708–716.

Jacob, M.; Coisne, G.; Gupta, A.; Sysoev, I.; Verma, G. G.; and Magerko, B. 2013. Viewpoints ai. In *Proceedings of the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE '13, 16–22. AAAI Press.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. In *Proceedings of the 9th International Conference on Computational Creativity*, ICCC '13, 104–111. Association for Computational Creativity (ACC).

Karimi, P.; Rezwana, J.; Siddiqui, S.; Maher, M. L.; and Dehbozorgi, N. 2020. Creative sketching partner: An analysis of human-ai co-creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, 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 *Proceedings of the International Conference on New Interfaces for Musical Expression*, NIME '20, 499–504.

Louie, R.; Coenen, A.; Huang, C. Z.; Terry, M.; and Cai, C. J. 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*, CHI '20, 1–13. New York, NY, USA: Association for Computing Machinery.

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*, 387–404. Oxford University Press.

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 *Proceedings of the International Workshop on Musical Metacreation (MUME)*.

Sturm, B. L.; Ben-Tal, O.; Úna Monaghan; 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* 48:36–55.

Suh, M. M.; Youngblom, E.; Terry, M.; and Cai, C. J. 2021. Ai as social glue: Uncovering the roles of deep generative ai during social music composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–11. New York, NY, USA: Association for Computing Machinery.