Jahrbuch Musikpsychologie
jbdgm.psychopen.eu | 2569-5665

Research Reports

The Creative Performance of the AI Agents ChatGPT and Google Magenta Compared to Human-Based Solutions in a Standardized Melody Continuation Task

Die Leistungen der Künstlichen Intelligenzen ChatGPT und Google Magenta im Vergleich mit Musikstudierenden bei einer standardisierten Melodie-Fortsetzungsaufgabe

Anton Schreiber[1], Kilian Sander[1], Reinhard Kopiez*[1], Raphael Thöne[1]

[1] Hanover University of Music, Drama and Media, Hannover, Germany.

Abstract

Many generative artificial intelligence (AI) systems have been developed over the last decade. Some systems are more of a generic character, and some are specialized in music composition. However, whether these AI systems are serious competitors for human composers remains unclear. Despite increased public interest, there is currently little empirical foundation for a conceivably equivalent performance for creative AI when compared to human experts in a controlled task. Thus, we conducted an online experiment to evaluate the subjectively perceived quality of AI compositions with human-made products (by music students) in a standardized task. Based on a melody continuation paradigm, creative products using AI were generated by the AI agents *ChatGPT* (Version 3.5) and *Google Magenta Studio* (Version 2.0). The human creative performances were realized by 57 melodic continuations, composed by music students. In the online evaluation study, listeners (N = 71, mainly musicians) rated the aesthetic quality of the outcomes of the various systems. Additionally, the raters' musical experience level was controlled as well as the length of the given melody completion task (two probe positions). As a main result, the overall quality of the AI compositions was rated significantly lower on all four target items compared to the human-made products (large effect sizes). Musical experience and the length of the melody did not influence the ratings. We conclude that the current capabilities of AI in the domain of musical creativity determined by a standardized composition task are far below human capabilities. However, we assume rapid progress will be made in the domain of generative music-specific AI systems.

Keywords: artificial intelligence, AI, composition, generative AI, empirical aesthetics, creativity

Zusammenfassung

Aktuell wird eine zunehmende Anzahl an generativen Systemen Künstlicher Intelligenz (KI) entwickelt. Einige Systeme sind eher von generischer Natur, andere wurden speziell für die Komposition von Musik entwickelt. Wie in anderen kreativen Bereichen ist noch unklar, welche Auswirkung diese KIs auf Musikschaffende haben wird. Trotz des kontroversen Themas, existiert bisher wenig Evidenz für die subjektiv bewertete Qualität von KI-Kompositionen im Vergleich zu menschlichen Kompositionen. Daher untersuchten wir in einem online Rating-Experiment die subjektiv bewertete Qualität von KI-Kompositionen im Vergleich zu Kompositionen von Musikstudierenden in einer standardisierten Aufgabe. Basierend auf einem Melodiefortsetzungsparadigma wurden Kompositionen mit den KIs *ChatGPT* (Version 3.5) und *Google Magenta Studio* (Version 2.0) erstellt. Musikstudierende generierten insgesamt 57 Fortsetzungsvarianten der gleichen Fortsetzungsaufgabe. In einem online Rating-Experiment bewerteten Teilnehmende (N = 71) die ästhetischen Qualitäten der Melodien. Zusätzlich wurde die musikalische Erfahrung der Teilnehmenden, sowie die Länge der vervollständigten Anfangsmelodie (zwei Probe Positionen) kontrolliert. Als Hauptergebnis wurden die Kompositionen der KIs für alle vier Bewertungs-Items schlechter als die menschlichen Lösungen bewertet (große Effekte). Musikalische Erfahrung, sowie die Länge der Anfangsmelodie hatten keinen Einfluss auf die Bewertung. Wir schlussfolgern, dass die kompositorischen Fähigkeiten musikalischer KIs noch deutlich hinter menschlichen Fähigkeiten liegen. Allerdings sind zukünftig rasante Entwicklungen im Bereich der generativen musikalischen KI-Systeme zu erwarten.

Schlüsselwörter: Künstliche Intelligenz, KI, Komposition, generative KI, empirische Ästhetik, Kreativität

Jahrbuch Musikpsychologie, 2024, Vol. 32, Article e195, https://doi.org/10.5964/jbdgm.195

Received: 2024-04-16. Accepted: 2024-08-21. Published (VoR): 2024-09-05.

Handling Editor: Kai Lothwesen, Staatliche Hochschule für Musik Trossingen, Trossingen, Germany

Reviewed by: Wolf-Georg Zaddach; Theresa Demmer.

*Corresponding author at: Hanover Music Lab, Hanover University of Music, Drama and Media, Neues Haus 1, 30175 Hannover, Germany. E-mail: Reinhard.kopiez@hmtm-hannover.de

The idea of a creative machine that can generate music at the touch of a button has fascinated composers since the invention of musical dice games in the 18th century (for an overview see Steinbeck, 2016). The idea of an automated creative process has continued into the 21st century, accelerated by the availability of powerful computers since the late 1960s (Strawn & Shockley, 2014). This development accelerated once again in the 1990s when personal computers with sufficient computing power became accessible to individuals. This can be considered the starting point of the first systematic exploration of the capabilities of Artificial Intelligence (AI). Based on the computer language LISP (a precursor of programming languages of AI) the American composer David Cope pursued the idea that musical compositions contain sets of instructions (style features) that can be used after identification to create highly related replications of themselves (Cope, 2000, p. 20). As a consequence, the understanding of creativity as a process that perpetually produces previously unheard music is replaced by a "recombinancy" paradigm. The elegance of recombinations contributes to the quality of these musical style copies. Using style copies of composers such as Bach, Mozart, Beethoven, Chopin, and others, Cope provides compelling evidence for the recombinancy paradigm (Cope, 1991, 1996, 2000, 2001). The development of the software for algorithmic composition was accompanied by Cope's perceptual experiments. In his *Experiments in Musical Intelligence* (Cope, 1996), he reports the results of listening tests with large numbers of participants on the ability to discriminate between original works by Mozart and style copies. For example, based on data from about 2,000 participants (delegates from a conference listening to the musical demonstrations and blind to the composition process of the respective musical example), musical amateurs outperformed expert musicologists (Cope, 2001, p. 21). The hit rates were usually around chance level (between 40 and 60 percent). In another series of listening tests, Cope (2001) suggests a discrimination test called "The Game": four completely machine-composed examples of music in the styles of Bach, Chopin, Smetana, and Cui are compared to similar style copies (recorded as human performances and added to the book on CD). Listeners scoring at least eight out of 12 correct responses (66%) are labelled as "High Scorers".

Now that AI agents have become available to a broader public (we will use the terms "agent" and "agency" in line with Latour's definition of agency as a property of human and non-human actors with the ability to alter a "state of affairs"; see Gioti, 2021, p. 55), these systems can be used to generate musical output, which in turn can influence musical thinking. Although this might occur in the context of co-creativity between humans and machines, our study focuses on automized composition. For example, the text-based AI agent *ChatGPT* (OpenAI, 2023), first released to the public in November 2022, is capable of accepting prompts for musical tasks in symbolic language (e.g., in MIDI code) and therefore can generate symbolic music as output (e.g., in MIDI format or other representations that can be converted into sounding music). Consequently, AI could pose a threat to creative musicians by offering a cheaper way to create music—an assumption that is supported by a more recent survey of composers (GEMA, 2024). Although the

use of AI agents for music production is a hot topic in the current technological development of AI agents, only a few studies have conducted a blind comparison of subjective evaluation of AI-generated musical output compared to human-generated music (e.g. Frieler & Zaddach, 2022; Tigre Moura & Maw, 2021; for an overview see Oksanen et al., 2023 and Yin et al., 2023).

For example, in their overview of subjective and objective evaluation strategies for AI-generated music, Xiong et al. (2023) have shown that basic empirical subjective evaluations of AI and human-generated music are rather rare. In their review of perceptual studies on the aesthetic quality of computer-generated music, Oksanen et al. (2023) found only ten empirical investigations between 2003 and 2021. In an early study on the influence of different narratives regarding the composition source (AI vs. human), Tigre Moura and Maw (2021, Study 2) used the AI-based song "Daddy's Car" (Vincent, 2016; for more details on the Flow Machines project see Pachet et al., 2021) and an AI-based symphonic film music titled "Genesis Symphonic Fantasy" created by AIVA software (https://www.youtube.com/watch?v=Ebnd03x137A). Surprisingly, listeners (blind to the composition process of the music) did not give negative evaluations of the AI condition. However, a fundamental problem in the handling of "AI music" can be observed in this and other studies that could be a reason for missing or inconsistent negative/positive evaluations: the stimulus used in the study by Tigre Moura and Maw (2021) resulted from the AI agent AIVA and represents a highly polished demonstration audio track released by the company. It does not represent the typical audio output of the AI agent but is the result of massive post-editing by a human arranger and is based on high-quality orchestra samples. This kind of promotional audio clip should be regarded as the outcome of co-creativity between AI and humans but not as representative of the output quality of automized AI composition agents (Gioti, 2021; Gioti et al., 2022). When co-creativity is not known to the listener, framing effects can also influence the evaluation of the aesthetic qualities of (widely unknown) human-sounding music from classical composers. For example, Shank et al. (2023) labelled short musical excerpts as (a) composed by AI, (b) composed by a human composer, or (c) a composer identity was not given. Results showed a significant effect of composer identity on the ratings of musical quality. Excerpts labelled as AI-generated music were rated significantly lower than those labelled as human-generated or those that provided no information on the composer's identity. Shank et al. (2023) label this effect as the "AI composer bias". Bias effects in the evaluation of AI vs. human-made artistic products have also been confirmed in a study by Millet et al. (2023): The authors investigated the influence of anthropocentric beliefs on the aesthetic evaluation of artworks such as paintings or music. As a result, two pieces of AI-generated music (created by the same AI agent AIVA in the style of symphonic film music) were rated as less creative (medium effect size) and less awe-inducing (small to medium effect size) compared to when they were labelled as human-made. This bias against AI-made art is known as "anthropocentric bias" in the evaluation of artistic creativity (Millet et al., 2023). From the perspective of music production in popular music, Deruty et al. (2022) studied the use of AI tools in the recording studio. The authors conclude that future production routines will likely be based on AI tools in terms of co-creative support in sound mixing, arrangements, and the production of rhythm tracks (for an overview of current tendencies see also Moffat, 2021).

Most similar to our study, Frieler and Zaddach (2022) compared the ratings of jazz solos of a generative model with human improvisation. They found that solos by jazz masters were generally rated better than algorithmically composed solos. However, in the classification task, even jazz experts only achieved 64.6% correct identifications. Therefore, in our explorative study, we conducted an online rating experiment to get more insight into the qualitative differences between human- and AI-generated compositions (beyond a mere discrimination paradigm) and thus, tested the aesthetic rating of compositions resulting from ChatGPT (Version 3.5, OpenAI, 2023) and Google Magenta Studio (Version 2.0; Google AI, 2023) compared to compositions by music students.

Research Question and Study Aim

The main research question is as follows: In terms of aesthetic quality, how similar are evaluations of compositions generated by AI systems and based on a standardized melody continuation paradigm compared to human-based compositions generated by music students? The aim of this study is to develop an empirical basis for future research into the aesthetic evaluation of creative products generated by musical AI agents. However, we cannot answer the question of how AI produces music in other musical systems outside of Western culture. Thus, in this study, we will focus on melodies in the style of Western music, which will be evaluated by listeners familiar with Western musical grammar. Due to the potential future impact of AI on the music industry and musicians, we argue that empirical research with an objective approach is needed in this field in order to assess the power and potential dangers of musical AI.

Hypotheses and Study Aims

In this explorative study, we did not test for any specific hypotheses. No hypotheses could be inferred due to the lack of specific previous studies and a theoretical framework, and it was therefore also impossible to calculate an a priori power analysis. However, in a post-hoc power analysis (see Results section), we tested whether a sufficient number of participants took part in the experiment to find the observed effect. Our goal was to learn more about the effects to provide a basis for future research.

Method

Design

For the subjective quality ratings of AI- and human-based compositions, we measured four dependent variables with one item each. These items were in part derived from previous research on related topics. The first item *musically convincing* was derived from Frieler and Zaddach (2022), and the second item *musically logical and meaningful* was obtained from Webster's (1994) measurements of creative thinking in music. Loosely related to Charyton et al.'s (2008) scale of originality, the third item asked whether the melody was *interesting*, and the fourth item asked how much the participants *liked* the melody (Frieler & Zaddach, 2022).

As an independent variable, we varied the composer in three conditions (human-based vs. AI agent ChatGPT vs. AI agent Magenta). Because the focus of this research was to evaluate differences between AI and human-based compositions, the statistical analysis focuses on the dichotomous differentiation with only two conditions for the independent variable (human-based vs. AI). Additionally, we controlled for the length of the prescribed melody that had to be continued. According to Schmuckler (1989), this independent variable was called *probe position* with two conditions (long vs. short). Finally, we controlled for the prior musical experience of the participants based on Zhang and Schubert's (2019) single-item scale of musical sophistication. This resulted in a 2 × 2 × 3 repeated measures design with the control variable of musical sophistication as a between factor.

Musical Stimuli

In line with previous research on musical expectation (e.g., Schmuckler, 1989; Unyk & Carlsen, 1987) we decided to use a melody continuation paradigm. Thus, the prescribed material for the compositions was a melody generated by one of

the co-authors (RT), a professional composer in the style of film music and arranged for strings (see Figure A1 in the Appendix; the full piece can be heard at https://www.youtube.com/watch?v=eYKdZBeY2eE). Due to different harmonic implications depending on where the melody was truncated, we decided to use two different lengths of the melodic material for the continuation task (see Figures A2 and A3 in the Appendix). The first condition used four bars of the original melody ending on an E♭4, therefore implying a harmonic context in C minor or E♭ major (probe position 1; PP1; Figure A2). The second condition had a length of 7 bars and ended on a D4, implying a harmonic context in D or G minor/major or even B♭ major (probe position 2; PP2; Figure A3). The musical stimuli were generated as follows: (a) music students from Hanover University of Music, Drama and Media continued either of the melodies following a standardized instruction (see Appendix 2). This resulted in a total of N = 57 melodies (PP1 and PP2). (b) As part of a seminar, the students also composed 42 melodies in the PP2 condition by means of ChatGPT. Six of the 42 melodies in the GPT-PP2 condition were not implemented in the rating experiment as they were too long. For this purpose, we used a Python syntax as a prompt (see Appendix 3) to transfer the prescribed original melody sections to ChatGPT. The results were exported in MIDI format using the Python module SCAMP (Evanstein, 2023). (c) To assess the quality of another AI agent, we produced an additional 50 melodies (25 for each condition) with the AI agent Google Magenta (Google AI, 2023; option *continue,* temperature [i.e., number of notes and pitch changes] = 1; length = 4 or 6 bars to add, number of variations [e.g., how many outputs the program will produce] = 4). (d) Additionally, we produced 25 melodies for the PP1 condition with ChatGPT (V3.5). The results from both AI agents were combined in the analysis of the participants' ratings. The resulting MIDI files were exported as mp3 audio files (timbre oboe) using MuseScore software (Version 4.0; MuseScore Team, 2023), normalized for loudness with Audacity (Audacity Team, 2023) to −10 LUFS.

The following total number of N = 168 melodic continuations was tested:

- n = 29 human-made continuations based on PP1
- n = 28 human-made continuations based on PP2
- n = 25 AI-based continuations (ChatGPT) based on PP1
- n = 36 AI-based continuations (ChatGPT) based on PP2
- n = 25 AI-based continuations (Magenta) based on PP1
- n = 25 AI-based continuations (Magenta) based on PP2

All materials are available in the Supplementary Materials section (see Schreiber et al., 2024).

Procedure

This study was conducted as an online experiment: N = 71 participants completed a questionnaire on the SosciSurvey platform (https://www.soscisurvey.de). In a randomized blind trial, each participant rated 20 melodies (five for each condition: AI agents/human-made for the two melody lengths of PP1 and PP2). In the instructions, participants were informed that some of the melodies were composed by AI (for the original wording see Supplementary Material, Schreiber et al., 2024). Consequently, no participant rated all 168 melodies, which resulted in an incomplete study design. The melodies were rated on a 5-point rating scale (1 = *not all all* [*gar nicht*] to 5 = *very much* [*sehr*]) and no anchor melody or prime was given. In addition, we asked for age, gender, and musical experience. The experiment took about 20 minutes to complete. Participants gave informed consent before starting the experiment. No reimbursement was paid.

Sample

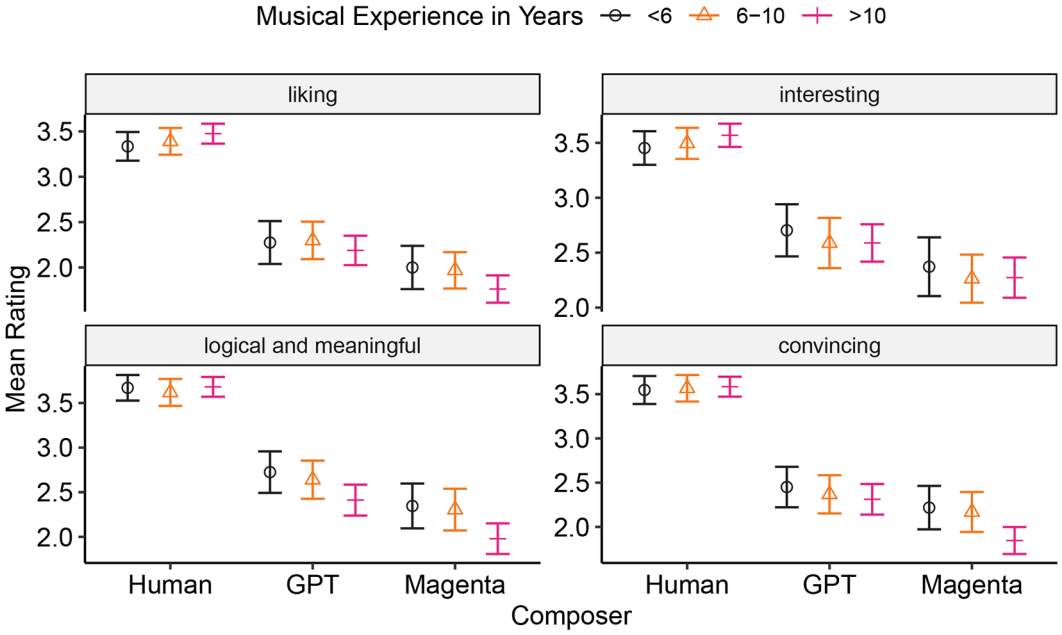
Out of the 71 participants, one person was excluded due to missing values and improbable responses. Participants were recruited from university seminars and personal networks of the authors. All had a Western cultural background. The sample showed a broad range of musical sophistication necessary to detect rating differences influenced by various levels of musical education. From the convenience sample, 31 (43.7%) were female and most of the participants studied music. The participants were aged between 20 and 79 years (*Mdn* = 27, *IQR* = 29). Their amount of musical training was as follows: 32 (45.1%) had more than 10 years of instrumental or vocal training, 21 (29.6%) had six to 10 years, and 17 (24.3%) had less than six years. This means that our sample for the rating task can be characterized as having a high degree of musical experience. In terms of Zhang and Schubert's (2019) categorization, this means that about 75% of the participants can be considered as raters with musical identity (6–10 years of musical training) or strong musical identity (> 10 years of musical training).

Results

As shown in Figure 1, as an overall effect, the compositions from music students were rated higher for all four dependent variables. For the AI agents, melodies from ChatGPT were rated slightly higher than those from Google Magenta. To test for the main effect of the independent variable "composer", a repeated measures MANOVA was calculated in R (R Core Team, 2023) using the stats software package. Since the comparison of results between the AI agents and humans was of special interest, ratings of ChatGPT and Google Magenta Studio were aggregated and regarded as representatives for the "AI generated" category.

Figure 1

Error Bar Diagram for the Mean Ratings of the Melodies for Each of the Four Items Grouped by Musical Experience and Sources of Melodic Creation



Note. The rating scale ranged from 1 (*not at all*) to 5 (*very much*). Error bars represent 95% confidence intervals.

Statistical analyses revealed a significant effect of the factor composer (source of melodic creation) on the rating of the four dependent variables, $F(1, 67) = 91.114$, $p < .001$, Pillai's trace = 0.857, $\eta^2 = 0.576$. The length of the completed melody had no influence on the rating, $F(1, 67) = 1.265$, $p = .293$, Pillai's trace = 0.073. This was also the case with the raters' degree of musical experience, $F(1, 67) = 0.554$, $p = .813$, Pillai's trace = 0.066, $\eta^2 = 0.019$. There were also no interactions between the factors (see Table 1). In terms of effect sizes, the observed rating differences between AI agents and human-made compositions were larger than 1.5 standard deviations for all dependent variables (see Table 2).

Table 1

MANOVA for all Factors With Simple Differentiation Between Human and AI Compositions

Effect	Factor	*df*	*V*	F_{approx}	df_F	*p*
Main effects	Musical experience	2, 67	0.066	0.554	8, 130	.813
	Composer	1, 67	0.857	96.114	4, 64	< .001
	Probe position	1, 67	0.073	1.265	4, 64	.293
Interactions	Composer: musical experience	2, 67	0.110	0.946	8, 130	.481
	Probe position: musical experience	2, 67	0.070	0.591	8, 130	.784
	Composer: probe position	1, 67	0.119	2.152	4, 64	.084
	Composer: probe position: musical experience	2, 67	0.061	0.509	8, 130	.848

Note. *V* = Pillai's trace. Colons in the Factor column represent interactions.

Table 2

Effect Sizes (Cohen's d_Z) for the Four Dependent Variables for the Comparison of AI- and Human-Made Compositions

	Effect size	
Dependent variable	**Cohens d_Z**	**95% Confidence Interval**
Interesting	−1.74	[−2.11, −1.37]
Logical and meaningful	−1.93	[−2.32, −1.53]
Liking	−2.23	[−2.66, −1.79]
Convincing	−2.11	[−2.53, −1.69]

Note. Negative d_Z values indicate lower ratings for AI compositions compared to human-made compositions.

Even though an a priori power analysis was not possible due to a lack of results from previous studies, a post hoc power analysis was calculated by means of G*Power (V3.1.9.6; Faul et al., 2007) to evaluate whether enough participants were included to unveil the observed effects. Based on $N = 70$ participants and a Pillai's trace value of $V = .857$ (the value for the main effect "composer") the calculated power was $1-\beta = 1.0$. Overall, the results show large effects in favor of the human compositions for all target variables. Neither the musical expertise of the participants, nor the length of the composed melody affected the striking difference in evaluation between human and AI compositions.

Discussion

The results show that the subjectively perceived and empirically confirmed quality of AI compositions is far below human-made compositions. This effect seems to be so large that the degree of musical experience had no influence on the rating. When listening to the stimuli, it became clear that, with just a few exceptions, the AI melodies sounded illogical and strange to our Western understanding of melodic construction (see Supplementary Materials for sound

examples, Schreiber et al., 2024). For example, the harmonic context of the prescribed melody was left intentionally ambiguous, but most of the AI-generated continuations did not even finish in the same key, resulting in a feeling of unresolved tonal tension and coherence at the end. Some melodies also contained breaks at unexpected metrical positions. These properties of the AI-generated melodies could explain why no effect of musical experience was observed and even musically naïve listeners gave lower ratings for the AI-generated versions. In terms of the perspective of Bigand and Poulin-Charronnat (2006) that all Western listeners are considered as "musically experienced listeners" with musical capacities acquired through exposure to music, we conclude that evaluation tasks like ours can be successfully performed without the help of explicit training or expertise. Interestingly, the ChatGPT compositions were rated slightly better for all dependent variables than those generated by Google Magenta, although this direct comparison between AI agents was not the focus of this study. Even though it is not a significant effect and may be due to variation in the data, it is surprising that results from the music-unspecific and text-trained AI agent ChatGPT were rated better than those from the music-specific AI agent Google Magenta Studio. However, although trained for the processing of musical material, we should bear in mind that Google Magenta Studio represents the predecessor generation of AI agents (first released in 2019) and ChatGPT the next generation (released by the end of 2022). We also conclude that our study provides first empirical evidence that contradicts the often anecdotal scenarios of the potential threat posed by the musically creative capabilities of AI systems. Under the condition of a standardized creative task, at least, we found no support for this assumption.

Our skeptical assessment of the current performance level of AI agents in the domain of music is in line with the critique by Rohrmeier (2022), who argues that computational creativity has to face four main challenges: (a) the cognitive challenge (creativity requires a cognitive model of music cognitions, e.g., for tonality); (b) the challenge of the external world (creativity includes the semantic, semiotic, and pragmatic world references, e.g., by references to an extra-musical program such as in Smetana's *Moldau*); (c) the embodiment challenge (creativity requires a model of the human body, e.g., for the possibilities of playing techniques); and (d) the challenge of creativity at the meta-level (creativity refers to meta-creative strategies such as form embeddings of a fugue within the classical sonata form, e.g., in Liszt's *B minor sonata*, or by the use of musical quotations). This general capacity of music and its creation requires the capacity for Artificial General Intelligence (AGI). As long as these prerequisites of musical creativity remain unresolved, these challenges will remain a fundamental AI-specific problem. However, according to the theoretical framework developed by Morris et al. (2024), AI agents such as ChatGPT have currently reached Level 2 (out of 6 levels) of AGI. In other words, musical AI agents have a long way to go before they at least reach the expert level (Level 4). However, we cannot exclude a significant increase in creative musical capacities for the next generation of AI agents, particularly when trained with music-specific materials of high quality and enriched with information about the external world. Finally, it remains to be seen whether extended training based on additional musical material will increase the outcome quality of the AI. For example, Pachet et al. (2021) report that "the most interesting music generation was not obtained with the most sophisticated techniques" (p. 512). Instead, the combination of various tools produced a more interesting musical output; second, the training of current AI systems depends on the availability of high-quality data. The availability of such data type seems to be limited and the LLM scaling of current AI systems seems to be constrained by this data type and the scarcity of AI raw material (Jürgens, 2024; Villalobos et al., 2024). This could result in insufficient data for training and, as Villalobos et al. (2024) conclude in a forecast, if current trends in LLM development continue, there is a 50% probability that the effective stock of human-generated public data will run out in 2024 and a 90% probability by 2026. Thus, the current trend in AI development is to attempt to compensate for the assumed data scarcity by applying data augmentation techniques (Xie et al., 2020).

Limitations

This study opted to use a melody continuation task to produce comparable stimuli between the two AIs and the music students. As a result, the choice of AI systems was very restricted. It is possible that the compositions could achieve higher ratings in a less restrictive task in which AI agents were allowed to compose polyphonic music with different instruments. Finally, the rapid and constant development of AI will lead to fast improvements in AI compositions. For example, a new generation of music-specific AI agents such as Suno AI Chirp (https://www.suno.ai) released in September 2023 integrates lyrics and human vocals as well as formal elements of song structure (e.g., verse and chorus) into the generation of popular music. Therefore, more studies like this should be conducted to constantly assess the development of the quality of AI compositions. By doing so, music research can make a valuable contribution to musicians and creatives in empirically investigating the progress being made by musical AI.

Statement of Ethics

The present study was conducted in accordance with ethical principles and standards according to the guidelines of the German Society for Psychology (Föderation Deutscher Psychologenvereinigungen, 2022) and with the principles outlined in the Declaration of Helsinki. According to German law, no ethics approval has been required. Written informed consent was attained by asking participants to continue only if they were willing to participate and if they had read and understood the instructions and information provided. Participants were told that participation was voluntary and that they had the right to withdraw from the study at any time. The data were anonymized and treated confidentially.

Funding

The authors have no funding to report.

Acknowledgments

The authors thank Prof. Dr. Raphael Thöne for making his composition available as material for this study.

Competing Interests

RK is Editor-in-Chief and KS is Editorial Assistant of the *Jahrbuch Musikpsychologie/Yearbook of Music Psychology*. They were not involved in the editorial process of this manuscript.

Data Availability

The research data for this article are available (see Schreiber et al., 2024).

Supplementary Materials

For this article, R scripts, data, codebook, and musical stimuli are available (see Schreiber et al., 2024).

Index of Supplementary Materials

Schreiber, A., Sander, K., Kopiez, R., & Thöne, R. (2024). *The creative performance of ChatGPT and Google Magenta compared to human-based solutions in a standardized melody continuation task* [Data, codebook, code, stimuli]. OSF. https://osf.io/qj8fp

References

Audacity Team. (2023). *Audacity* (Version 3.4.2) [Computer software]. https://audacityteam.org

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*(1), 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

Charyton, C., Jagacinski, R. J., & Merrill, J. A. (2008). CEDA: A research instrument for creative engineering design assessment. *Psychology of Aesthetics, Creativity, and the Arts, 2*(3), 147–154. https://doi.org/10.1037/1931-3896.2.3.147

Cope, D. (1991). *Computers and musical style*. A-R Editions.

Cope, D. (1996). *Experiments in musical intelligence*. A-R Editions.

Cope, D. (2000). *The algorithmic composer*. A-R Editions.

Cope, D. (2001). *Virtual music: Computer synthesis of musical style*. MIT Press.

Deruty, E., Grachten, M., Lattner, S., Nistal, J., & Aouameur, C. (2022). On the development and practice of AI technology for contemporary popular music production. *Transactions of the International Society for Music Information Retrieval, 5*(1), 35–49. https://doi.org/10.5334/tismir.100

Evanstein, M. (2023). *SCAMP (Suite for Computer-Assisted Music in Python)* (Version 0.9.2) [Computer software]. http://scamp.marcevanstein.com/

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Föderation Deutscher Psychologenvereinigungen. (2022). *Berufsethische Richtlinien* [Guidelines for professional ethics]. https://www.dgps.de/fileadmin/user_upload/PDF/Berufsetische_Richtlinien/BER-Foederation-20230426-Web-1.pdf

Frieler, K., & Zaddach, W.-G. (2022). Evaluating an analysis-by-synthesis model for jazz improvisation. *Transactions of the International Society for Music Information Retrieval, 5*(1), 20–34. https://doi.org/10.5334/tismir.87

GEMA. (2024). *AI and music: Generative artificial intelligence in the music sector*. https://www.gema.de/en/news/ai-study

Gioti, A.-M. (2021). Artificial intelligence for music composition. In E. R. Miranda (Ed.), *Handbook of artificial intelligence for music: Foundations, advanced approaches, and developments for creativity* (pp. 53–73). Springer International Publishing.

Gioti, A.-M., Einbond, A., & Born, G. (2022). Composing the assemblage: Probing aesthetic and technical dimensions of artistic creation with machine learning. *Computer Music Journal, 46*(4), 62–80. https://doi.org/10.1162/comj_a_00658

Google AI. (2023). *Magenta* (Version 2.0) [Computer software]. Google AI. https://magenta.tensorflow.org

Jürgens, J. (2024, May 8). Alles aufgesaugt [All suctioned]. *DIE ZEIT*. https://www.zeit.de/2024/21/kuenstliche-intelligenz-trainingsdaten-suche-google-meta-openai

Millet, K., Buehler, F., Du, G., & Kokkoris, M. D. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior, 143*, Article 107707. https://doi.org/10.1016/j.chb.2023.107707

Moffat, D. (2021). AI music mixing systems. In E. R. Miranda (Ed.), *Handbook of artificial intelligence for music: Foundations, advanced approaches, and developments for creativity* (pp. 345–375). Springer International Publishing. https://doi.org/10.1007/978-3-030-72116-9_13

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). *Levels of AGI: Operationalizing progress on the path to AGI.* arXiv. http://arxiv.org/pdf/2311.02462v2

MuseScore Team. (2023). *MuseScore* (Version 4.0) [Computer software]. https://musescore.com/about

Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y., & Savela, N. (2023). Artificial intelligence in fine arts: A systematic review of empirical research. *Computers in Human Behavior, 1*(2), Article 100004. https://doi.org/10.1016/j.chbah.2023.100004

OpenAI. (2023). *ChatGPT* (Version 3.5) [Computer software].

Pachet, F., Roy, P., & Carré, B. (2021). Assisted music creation with Flow Machines: Towards new categories of new. In E. R. Miranda (Ed.), *Handbook of artificial intelligence for music: Foundations, advanced approaches, and developments for creativity* (pp. 485–520). Springer International Publishing. https://doi.org/10.1007/978-3-030-72116-9_18

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rohrmeier, M. (2022). On creativity, music's AI completeness, and four challenges for artificial musical creativity. *Transactions of the International Society for Music Information Retrieval, 5*(1), 50–66. https://doi.org/10.5334/tismir.104

Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception, 7*(2), 109–149. https://doi.org/10.2307/40285454

Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., & Belfi, A. M. (2023). AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology. Applied, 29*(3), 676–692. https://doi.org/10.1037/xap0000447

Steinbeck, W. (2016). Würfelmusik [Dice music]. In *MGG Online.* https://www.mgg-online.com/mgg/stable/12552

Strawn, J., & Shockley, A. (2014). Computers and music. In *Grove Music Online.* Oxford University Press. https://doi.org/10.1093/gmo/9781561592630.article.A2256184

Tigre Moura, F., & Maw, C. (2021). Artificial intelligence became Beethoven: How do listeners and music professionals perceive artificially composed music? *Journal of Consumer Marketing, 38*(2), 137–146. https://doi.org/10.1108/JCM-02-2020-3671

Unyk, A. M., & Carlsen, J. C. (1987). Influence of expectation on melodic perception. *Psychomusicology: Music, Mind, and Brain, 7*(1), 3–23. https://doi.org/10.1037/h0094189

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data.* arXiv. http://arxiv.org/pdf/2211.04325v2

Vincent, J. (2016, September 26). This AI-written pop song is almost certainly a dire warning for humanity. *The Verge.* https://www.theverge.com/2016/9/26/13055938/ai-pop-song-daddys-car-sony

Webster, P. A. (1994). *Measure of creative thinking in music (MCTM): Administrative guidelines* [Unpublished manuscript].

Xie, Q., Dai, Z., & Hovy, E., Luong, M.-T., & Le, Q. V. (2020). *Unsupervised data augmentation for consistency training.* 34th Conference on Neural Information Processing Systems. https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html

Xiong, Z., Wang, W., Yu, J., Lin, Y., & Wang, Z. (2023). *A comprehensive survey for evaluation methodologies of AI-generated music.* arXiv. http://arxiv.org/pdf/2308.13736v1

Yin, Z., Reuben, F., Stepney, S., & Collins, T. (2023). Deep learning's shallow gains: A comparative evaluation of algorithms for automatic music generation. *Machine Learning, 112*(5), 1785–1822. https://doi.org/10.1007/s10994-023-06309-w

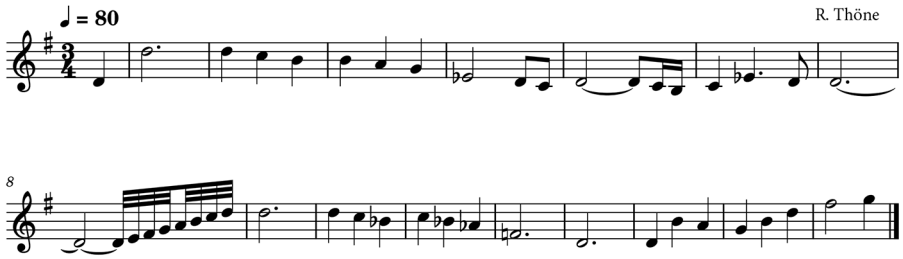
Zhang, J. D., & Schubert, E. (2019). A single item measure for identifying musician and nonmusician categories based on measures of musical sophistication. *Music Perception, 36*(5), 457–467. https://doi.org/10.1525/mp.2019.36.5.457

Appendix

Appendix 1: Melodic Material for Stimulus Generation

Figure A1

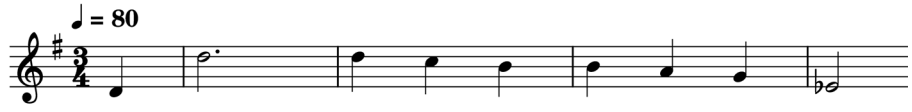
Complete First Phrase of the Original Melody "Aus meiner Feder" [From My Pen]



Note. Use of melodic material with the kind permission of Raphael Thöne. The full arrangement of this melody can be heard at https://www.youtube.com/watch?v=eYKdZBeY2eE

Figure A2

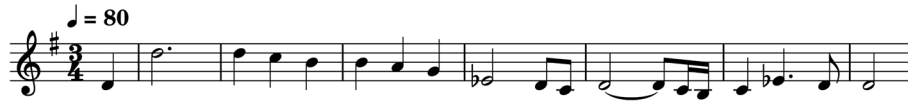
Original Melody Used for Probe Position 1



Note. Use of melodic material with kind permission of Raphael Thöne.

Figure A3

Original Melody Used for Probe Position 2



Note. Use of melodic material with kind permission of Raphael Thöne.

Appendix 2: Composition Instructions for Music Students

Your Task

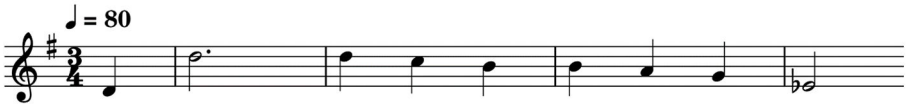
Please complete the melodic example below as you wish but in line with the following rules: the continuation should have:

- a length of 10–20 notes,
- a range from D3 (d) to D5 (d‘‘),
- different note durations (not only quarter notes),
- a clear melodic climax.

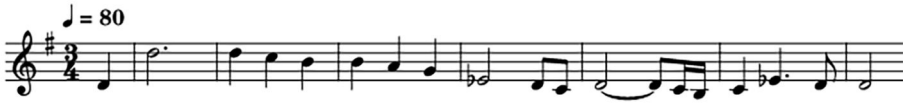
Each person should submit 5 melodic continuations.

You can use musical notation, play the solutions on an instrument of your choice, sing it and make a recording, or enter it directly into a notation program (e.g., MuseScore). Transposing instruments can be notated as fingered.

Melodic Beginning for Group A (PP1)



Melodic Beginning for Group B (PP2)



Appendix 3: Prompt for the AI Agent ChatGPT (V3.5)

Prompt (Syntax With Composition Instructions) for the AI Agent ChatGPT (V3.5)

Continue the given melody in the form of a list of (pitch, duration) pairs in Python syntax, where the pitch uses the MIDI pitch standard, and the duration represents the number of quarter notes. Use a pitch of None to represent a rest. Ensure the following:

- The continuation stays between MIDI pitch 52 and MIDI pitch 86
- The continuation is between 10 and 20 notes in length
- The melody should have a calm character and be in the style of film music
- The continuation should use a variety of note lengths
- The continuation should have a clear melodic peak

Option (a):

melody_pp1 = [(62, 1), (74, 3), (74, 1), (72, 1), (71, 1), (71, 1), (69, 1), (67, 1), (63, 2)]

Option (b):

melody_pp2 = [(62, 1), (74, 3), (74, 1), (72, 1), (71, 1), (71, 1), (69, 1), (67, 1), (63, 2), (62, 0.5), (60, 0.5), (62, 2.5), (60, 0.25), (59, 0.25), (60, 1), (63, 1.5), (62, 0.5), (62, 3)]

Note. Two melodic fragments of different lengths were used as inputs to be selected: (a) pp1 = probe position 1 (shorter melodic fragment, see Figure A2); (b) pp2 = probe position 2 (longer melodic fragment; see Figure A3).