# Searching for Human Bias Against AI-Composed Music

Dimiter Zlatkov[1(✉)], Jeff Ens[2], and Philippe Pasquier[2]

[1] University of British Columbia, Vancouver, BC V6T 1Z1, Canada
dzdimi14@gmail.com
[2] Metacreation Lab, Simon Fraser University, Burnaby V5A 1S6, Canada
https://metacreation.net/human-bias-against-ai-composed-music/

**Abstract.** With the popularization of musical AI in society comes the question of how it will be received by the public. We conducted an empirical study to investigate the hypotheses that human listeners hold a negative bias against computer-composed music. 163 participants were recruited from Amazon's MTurk to fill out a survey asking participants to rank 5 computer-composed and 5 human-composed musical excerpts based on subjective musical preference. Participants were split into two groups, one informed of correct authorship, the other deceived. The hypothesis, that those in the informed group would rank computer-composed excerpts as lower than human-composed excerpts, was not supported by significant results. We outline potential weaknesses in our design and present possible improvements for future work. A review of related studies on bias against AI-composed music and art is also included.

**Keywords:** Musical Metacreativity · Computational Creativity · Human-computer Interaction · Bias

## 1 Introduction

Neural Networks are becoming widely used to automate a number of tasks. They can act as our chauffeurs and personal assistants and they can even outperform humans at skilled tasks such as screening for melanoma (Fogel and Kvedar, 2018). While it is becoming accepted that humans can be outperformed by computers in various logical tasks, the general belief holds that this is not the case when it comes to more creative endeavors. However, AI systems, such as Google's DeepDream, are able to create novel visual artwork using their advanced neural networks and large databases (Fogel and Kvedar 2018; Marzano and Novembre 2017). Beyond visual artwork, computer systems have also created literary works and films (Hong and Curran 2019).

There have been computer-composed musical works as early as 1957 when professors of music, Hiller and Issacson, created the Illiac Suite for strings using the Illiac I computer at the University of Illinois Urbana-Champaign. Today, artificially composed music has come a long way to the point where programs such as audiometaphor.ca (Thorogood et al., 2022) will create a soundtrack


C. Johnson et al. (Eds.): EvoMUSART 2023, LNCS 13988, pp. 308–323, 2023.
https://doi.org/10.1007/978-3-031-29956-8_20

based on only a few user imputed key-words. There are even pop albums being released which use artificially composed music such as Hello World by SKYGGE (helloworldalbum.net 2021). This new field of creative music systems using AI has become known as Musical Metacreativity (https://musicalmetacreation.org). Musical Metacreativity (MuMe) systems range from tools meant to help users create music to more autonomous systems, which can both compose and play their own novel works.

Computational creativity presents a number of philosophical questions, particularly whether an AI system can be creative. Where general intelligence can often be measured by various tests, creative works do not have an optimal solution. The evaluation of creative works is fundamentally a subjective process. Critics argue that because many MuMe systems are trained using large databases of human-composed work, their compositions are not capable of being creative since they are imitations of past work (Jennings 2010). However, it is also true that artistic inspiration often comes from the work of others (Jackson 2017). Even though computer-composed works are indistinguishable from that of humans, there are still many people who argue these works are not "human-like" (McCarthy 2007). The question still remains whether people are willing to accept the creativity shown by MuMe systems. More specifically, do people hold a bias against computer created artwork such as music?

## 2 Background

### 2.1 Bias Against Computer Composed Artwork

Rather than looking at a bias against AI-composed artwork in the form of musical compositions, Norton et al. (2013) present a system called DARCI which was designed to render images to match a list of adjectives. Their system is based on Colton's creative tripod (2008) which posits there are three necessary behaviours for a system to be considered creative: skill, imagination, and appreciation. Colton defines skill as the ability to produce functional or quality artifacts which are recognized as members of their intended domain. Colton's definition of imagination adds that these artifacts must be original and meaningful in some way. Finally, Colton's appreciation is the ability of the system to evaluate its own works.

DARCI demonstrates creativity through Colton's three necessary behaviours. It demonstrates skill through creating original images that correlate with the appropriate adjectives it is given, imagination by generating these unpredictable yet non-random images, and appreciation by evaluating how strong of a match each image is to a database of adjective semantics. The images created by DARCI are created through an evolutionary process. In this process, a myriad of possible images is defined through a specific encoding called a genotype, and the rules used to transform the genotype into the desired output is called a phenotype. A population of random genotypes are evaluated based on the qualities of their respective phenotypes. This evaluation then goes through a fitness function

which determines which genotypes pass on to future generations of the evolutionary system. DARCI uses a fitness function which uses artificial neural networks to model user aesthetics and image features as the input to the neural network. DARCI is unique in that it focuses on computational creativity rather than pure evolutionary art. Therefore, DARCI's fitness function is made up of many neural networks each corresponding to a specific adjective, this way DARCI's fitness function does not measure a single kind of aesthetic sense but rather an overall sense of what an image means.

In addition to creating DARCI, Norton et al. developed a set of metrics for evaluating their system as well as an online survey to measure the novelty and quality of DARCI's work. Their survey had 6 questions all of which were answered using a five-point likert-type scale. The questions were specially designed to assess the creativity behind DARCI's creations. They asked participants whether they like the image, whether they think the image is novel, whether they would use the image as a desktop wallpaper, whether they had seen the image before, if they thought the image was hard to make, and if it was creative. The images chosen for the survey were created from the same source image and modified by one of ten adjectives: bright, cold, creepy, happy, luminous, peaceful, sad, scary, warm and weird.

The results of their survey suggest that people do in fact consider some of DARCI's creations to be creative. It was found that adjectives which describe emotion (peaceful, scary, happy, sad, and creepy) scored the highest on average in the survey. Adjectives which only described particular attributes (bright, warm and luminous) scored lowest. Interestingly, the highest scoring image was created off the adjective weird, which while not necessarily an emotion, is consistent with certain components of creativity such as novelty.

Similar to Norton et al. (2013), Ragot et al. (2020) conducted an experiment examining whether there is a negative perception towards AI-created paintings. They recruited 565 participants to evaluate paintings, produced by AI or humans, based on four dimensions: liking, perceived beauty, novelty, and meaning. Similar to previous studies, they used a priming effect stating whether each work was created by AI or a human. Participants were split into an AI-condition group and a human-condition group. Those in the AI condition were primed to believe that all the paintings they were presented were by AI, and those in the Human condition were told the paintings presented to them were created by human artists.

Participants in each group were presented with 8 different paintings randomly selected from 40 total paintings: 10 portraits created by AI, 10 landscapes created by AI, 10 landscapes created by humans, and 10 portraits created by humans. Participants in both groups were actually shown a mix of paintings created by AI and human artists. After rating each painting on the four dimensions, participants were asked if they remember whether the paintings they were presented were created by AI or humans as a manipulation check. Afterwards, participants were told the origin of the paintings were manipulated and asked to guess the origin of four randomly selected paintings. This was conducted as a modified Turing-test to avoid any bias in evaluation.

Ragot et al. (2020) found that participants did in fact rate art presented as AI-created to be significantly less liked and perceived as less beautiful, novel, and meaningful than those presented as being created by humans. Their results support those of Moffat and Kelly in the wider question of assessing a bias against computational creativity. They also found there was higher recognition of correct authorship for human paintings than AI paintings in their modified Turing-test.

Ragot et al. present these results to possibly be an effect of an inter-group bias. Inter-group bias generally refers to the tendency to evaluate one's own membership group (in-group) more favourably than a non-membership group (out-group) (Hewstone et al. 2002). In-group members show higher trust, cooperation, empathy, and positive regard to other members of their group, but not to out-group members. This discrimination towards those in the out-group could explain why participants rated human-created art (in-group) higher than AI-created (out-group) paintings. A key moderator of inter-group bias is threat. Threat can be defined in terms of the in-group's social identity, goals, values, position in the hierarchy, or even its existence (Hewstone et al. 2002). Ragot et al. propose that, in terms of their experiment, participants could view the existence of high quality computational creativity as a threat to their in-group (humans). Fear of AI systems often presents itself with the idea of people's jobs being displaced due to AI automation (McClure 2017). Perhaps it is due to a fear of human creativity being replaced that human artists are being rated higher than AI artists.

### 2.2 Bias and Musical Metacreativity

Multiple experimental studies have been conducted to test how people perceive AI-created artwork. Rather than testing the creativity of AI created art, Moffat and Kelly (2006) looked for a proposed bias against computer-composed music. Their experiment asked twenty participants to discern whether each composition was composed by a human or computer in a Turing test-like fashion. Participants were first tested on their music knowledge/experience and then subsequently divided into 'Musician' and 'Non-Musician' groups. The participants in both groups were given six one-minute long musical excerpts, three of these were human-composed and the other three composed by various computer systems. These pieces came in the form of three different styles: "Bach", "Strings", and "Free-form Jazz". After listening to each composition, participants indicated how much they "liked" a particular piece on a 5-point Likert scale. After this round, the authorship of each piece was revealed and the participants were then asked to rate each piece again, although in a disguised manner asking how willing they would be to buy, download, or recommend each piece to someone.

Moffat and Kelly found that their participants appeared to show prejudice against the AI-composed pieces, however, these findings were not strong enough to be deemed statistically significant due to the small sample size. They did, however, find significant data that participants seemed to consistently prefer human-composed music to computer-composed music. They also found that participants

were good at determining which pieces were computer-composed. Surprisingly, non-musicians outperformed musicians at this task.

Pasquier et al. (2016) built on the work of Moffat and Kelly (2006), conducting an empirical study investigating whether listeners hold a bias against computer composed music. Their study sought to improve upon previous studies, which had the issue of participants trying to outsmart the procedure by trying to pick up on "clues" to determine song authorship. By removing the Turing test-like condition from their study, Pasquier et al. tried to remove this difficulty from their study. They also attempted to remove any practice effects that may have happened from the non-randomized order of stimuli in previous studies.

Unlike Moffat and Kelly's (2006) study, Pasquier et al. (2016) divided their participants into three groups: Informed, Naive, and Revealed. In the Naive condition, participants were unaware of song authorship whereas in the informed condition the participants were explicitly told the author for each piece of music. In the Revealed group, participants first heard each piece without knowing the song authorship. Then, in the second round of listening, the authorship of each piece was explicitly told for each song. This condition of the experiment allowed them to check if there would be any "reaction" effect where the newfound knowledge of authorship could cause a drastic change in how people rate the music.

Because they did not include a Turing test-like section to their study, Pasquier et al. did not divide their participants into musician and non-musician groups. However, they did include a demographics questionnaire which included questions about age, gender, university major, country of birth and number of years living in Canada. They also asked each participant how much experience they had with computer programming languages as a measure for computer literacy. This measure was asked after the experiment to increase deception.

Where Moffat and Kelly decided that their stimuli should be of three different styles, Pasquier et al. chose to limit their musical selection to only one style. They created three unique computer-composed pieces in the style of "contemporary string quartet" and paired them with three other human-composed pieces that were of similar structural characteristics, taking into account tempo, polyphony, rhythm, and dynamics for pairing each piece. They believed that this pairing technique would offer better results than Moffat and Kelly who only paired pieces by style. All of the pieces used in Pasquier et al. were performed by the same string quartet who were given an equal amount of time to practice and perform each piece. The performers were unaware of which pieces were composed by humans and which ones were artificial. All of the artificial works were created by the software of Arne Eigenfeldt, a Canadian composer and long-term collaborator of Dr. Philippe Pasquier.

Participants were given a URL to an online survey where they were presented with video recordings of each piece one at a time. Between each musical piece, participants were shown a "palate-cleanse" and each participant was shown a random order of pieces in order to reduce practice effects. After listening to each piece, participants were then asked to rate each piece on four

different attributes using a 50-point bipolar scale. The four dimensions measured were: 'Good-Bad', 'Like-Dislike', 'Emotional-Unemotional', and 'Natural-Artificial'. They chose this rating method as a more sensitive measure to identify bias, which could be obscured in a more simple rating, such as simply "liking," which was used in previous experiments such as Moffat and Kelly.

Similar to Moffat and Kelly's study, Pasquier et al. did not find any significant results that show there is a bias against musical metacreativity, however, their results suggest that such a bias may exist. Their data showed that listeners were fairly uncertain in their ratings regardless of their knowledge of song authorship. There was a slight skew towards the "bad" and "artificial" dimensions for participants' ratings in the Revealed condition, but these were not significant enough to show any meaningful bias. The authors suggest that replication should be done with a larger sample size and recognize that their study is limited to only certain musical conditions, particularly style.

### 2.3 Influence of Context and Expectation When Searching for Bias Against AI-composed Music

Hong et al. (2020) conducted a study that examined the influence of people's met or unmet expectations about AI and their assessments of AI-composed music. Their experiment incorporated Expectancy Violation Theory (EVT) which explained individuals' reactions to met or unmet expectations in communication settings (Burgoon et al. 2016). This theory argues that when one's expectations are exceeded, they perceive the outcome as more favourable than if they made no expectations at all. The same effect happens in the negative direction, where if one's expectations were not met, the outcome is perceived as less favourable than if there were no expectations made. This holds that people's evaluation of artwork is biased by their belief of whether or not AI can be creative rather than the artwork itself. Hong et al. hypothesized that those participants who thought AI-composed music was better than expected will give higher ratings than people who think the music meets their expectations and vice versa.

They recruited 299 participants to complete their online study using Amazon's Mechanical Turk (MTurk). Four AI-composed musical songs were used for the study, two of which were in the electronic dance music (EDM) genre, the other two were of the classical genre. A pilot study was first conducted to make sure each piece would be rated similarly. After confirming each song would be rated equal in quality, each participant was presented with a randomly selected song out of the four and asked to listen to it. After listening to their given song, participants were told to report their evaluation of the musical piece using a 9-item scale based off of the "Rubric for assessing general criteria in a musical composition" (Hickey, 1999). This 9-item scale measured the aesthetic appeal, creativity, and craftsmanship of each song. After rating for musical quality, participants were asked how much the music's quality deviated from their expectations. This was measured using a 7-point Likert-type scale based off one used in a previous study by Burgoon et al. (2016). Finally, Participants filled

out another 7-point Likert type scale measuring participants' understanding of AI's creativity.

Their results showed a positive relationship between the perception of creative AI and the evaluation of AI-composed music. Their work shows an interesting implication when working with creative AI, namely that people's attitudes towards AI has a large impact on their attitudes towards AI-made creative works. Those who are able to be persuaded that AI can be creatively autonomous are more likely to rate MuMe works as having creative value. This holds in the other direction as well. Those with negative preconceptions of MuMe systems are shown to devalue MuMe compositions significantly.

Hong et al. (2020) found that one's prior expectations and attitudes towards AI have an impact on how they rate AI-composed music, showing that outside factors may play a part in how we perceive AI-composed music. Another important factor which may influence people's attitudes towards AI-composed music may be the cultural context of the music itself. Deguernel et al. (2022) looked to see what effects the culture and context of the music where AI is being applied has on their perception of a piece. They chose a genre where computer authorship is considered to be generally opposed to the music creating process, Irish Traditional Music (ITM). ITM has a strong emphasis on authenticity and etiquette (Hillhouse 2005), there are strong opinions on how tunes should be played and taught and which instruments should be used. Due to the values of ITM, there is a fear of a loss of authenticity of the genre due to commercialization and more modern music production techniques (Xuan & Ying 2022). Given the context of ITM is rooted heavily in human-centered tradition, Deguernel et al. hypothesize that ITM practitioners will show a bias against liking music they believe to be composed by an AI.

To test their hypothesis, Deguernel et al. recruited participants from traditional music programs at the University of Limerick, Ireland and had them complete an experiment where they would listen to 6 pieces of AI-composed Irish Traditional Music. Each piece was hand selected by a professional ITM musician, Padraig O'Connor. After selecting his 6 favorite pieces from a large corpus of 58,105 tunes, O'Connor recorded himself playing each composition on solo accordion, adding stylistic ornamentation and variations as he saw fit to ensure each recording sounded authentic. These 6 pieces were presented to participants during their completion of two tasks, the "Liking task" and the "Authorship task".

In the "Liking Task", participants were presented with each piece of music and asked "How much do you like the tune?" and recorded their answers on a 5-point Likert scale labeled: "Don't like it at all", "Don't like it", "Neutral", "Like it", and "Like it a lot". The scale was not shown until after the song finished playing and participants were encouraged to use the full range of the scale. After rating how much they liked each piece, participants completed the "Authorship Task". In The "Authorship task", participants were asked "How likely do you believe that the tune is composed by a computer?" and given another 5-point Likert scale. The two tasks were completed in this specific order to ensure that ratings in the

"Liking task" are not affected by a prior mention of AI. After completing both tasks, participants filled out a short questionnaire about demographics, musical practice, and familiarity with ITM.

Deguernel found a plausible bias amongst ITM practitioners for these six AI-composed pieces. Participants tended to like the tunes that they deemed likely to be composed by a human, and disliked pieces they believed to be composed by an AI. The difference in results with that of Moffat and Kelly (2006) and Pasqier et al. (2016) help validate the hypothesis of Deguernel et al. (2022), that the context in terms of musical culture and participants plays a role in observing a bias against AI-composed music. However, Deguernel et al. consider their work a pilot study and say that power analysis of their experiment must be conducted to determine the likelihood of a Type-I error.

When searching for a proposed bias against AI-composed music, the work of Moffat and Kelly (2006) and Pasquier et al. (2016) failed to find any significant bias. However when looking at whether people's prior conceptions about AI and musical AI compositions, Hong et al. (2020) and Deguernel et al. (2022) seem to find some effect. It seems that one's prior expectations as well as the cultural context of the music itself, comes to play when seeking for a negative human bias against AI musical compositions.

## 3 Methods

The present study aims to build off of the work conducted by Moffat and Kelly (2006), as well as Pasquier et al. (2016) in the aim to search for a general listener bias against computational creativity. By also asking participants their opinions on creative AI systems, we also build off the work of Hong et al. (2020) and investigate how people's predispositions towards computational creativity may affect their ratings towards musical metacreativity. We created a ranking survey based on the suggestions of Pasquier et al. (2016), this survey also included demographic questions including polls on musical and technological ability. According to, Yannakakis and Martinez (2015), rank-based questionnaires can help eliminate some of the problems associated with ratings-based questionnaires when evaluating subjective, psychological factors like emotional response, preference, or opinion.

### 3.1 Participants

Participants for both the pilot and main experiments were recruited using Amazon's Mechanical Turk (MTurk) service. Participants were compensated with a small monetary reward after completing the survey. Those who did not pass attention checks were excluded from the study. There were a total of 163 participants in the main experiment, two of which were excluded for not completing the attention check. The majority of participants identified as male (72%), with only one participant identifying as non-binary.

### 3.2 Musical Excerpts

The artificially composed pieces were created using the Multi-track Music Machine (Ens and Pasquier 2020). This system uses the transformer architecture to generate multi-track music by providing users with a fine degree of control over iterative re-sampling. The system is based on an auto-regressive model which is capable of generating novel music from scratch in a wide variety of genres and styles by using a multitude of preset instruments. Tracks can also be generated using MIDI track input as 'inspiration,' re-sampling the piece into further layers of musical composition.

Each computer-composed musical excerpt was paired with a similar human-composed work of the same style. These pairings were made based on rhythm, tempo, and tonality. These pieces were then later rated in a pilot study to ensure that each pair of human and computer-composed music would be rated equally when participants were blind to song authorship. All of the pieces used in the final experiment were of the same genre: contemporary pop.

### 3.3 Procedure

To test that the musical quality of the computer-generated pieces are on par with human-composed pieces of the same style, a short pilot study was conducted on Amazon MTurk. Participants were told to rate around 30 pieces to test for "equality". These 30 pieces were a mix of computer-composed and human-composed. For this survey, participants were not told about the authorship of each piece, they simply rated each song in terms of musical quality on a 7-point Likert scale and given no further information than what they could hear. This was done to ensure that computer composed pieces would be ranked equally as human-composed pieces. There was no significant difference in ratings between human-composed and computer-composed pieces during the pilot study.

In the primary study, participants were randomly divided into two groups: Deceived and Informed. After providing consent, participants completed a brief attention check. As seen in Fig. 1, participants in both groups were presented with ten different musical excerpts (around 10 s each) of mixed authorship for the first period of the experiment. In this period, participants were blind to song authorship and given the opportunity to play and pause each clip as they pleased. Playtime was tracked to ensure that each participant listened to each clip for an adequate amount of time. One of the audio clips presented to the participants was an audio message telling each participant to rate that clip at the bottom of the list, this was used to ensure participants were paying attention to the experiment. Participants were then told to rank each clip against one another by dragging their favourite to the top of the list and their least favourite to the bottom. After completing this first round of ranking, both groups entered the second period of the experiment where they would listen to the four musical excerpts again, however, this time they were told the authorship of each musical excerpt. The Informed group was told the actual composer for each piece, whether it was composed by a human or computer, whereas the Deceived group was told the

wrong composer for each song. In other words, the pieces that were composed by a human were presented as being composed by a computer system and vice versa in the Deceived group. For this second period, participants were asked how likely they would be to listen to each song on their own as a measure of how much each participant liked each piece. This measure was used to ensure that participants would not catch on to the purpose of the study.

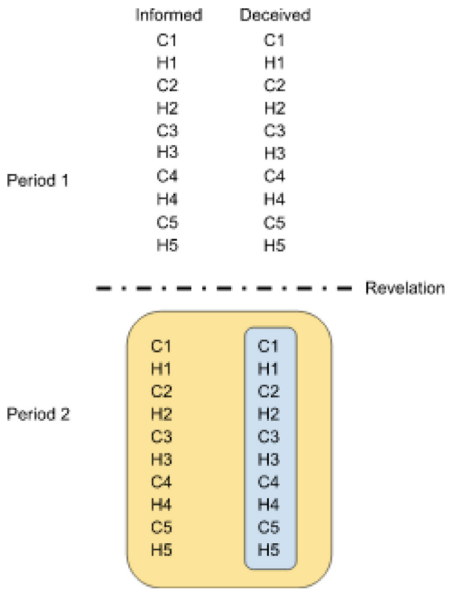


**Fig. 1.** Experiment design showing the two conditions, Informed and Naive. The yellow shaded area represents the period where authorship is told for each song. The blue shaded represents the songs where authorship was told deceptively. C1, C2, C3, C4, C5 refer to computer composed pieces where H1, H2, H3, H4, H5 refer to human-composed pieces respectively, the order of all pieces was randomized for each participant. (Color figure online)

After completing the second period of listening and ranking each musical piece, participants were asked to provide some general demographic information such as their age, gender, and country of residence. The questionnaire also included a number of questions asking participants how many years they have spent studying or playing music, as well as other questions to assess their level of computer literacy. After completing the questionnaire, each participant was debriefed on the purpose of the study if they chose to do so and were provided with contact information in case they had any further questions regarding the study.

We hypothesize that those in the Informed group will rank computer-composed music as lower than human-composed music. Secondly, we predict

that there will be a positive correlation between people's technological literacy and attitudes towards computational creativity and their overall rankings of computer-composed music.

## 4 Results

To analyze our data, a repeated measures or within-subjects ANOVA was conducted to compare the effects of perceived authorship on song ranking. This measure looked to see whether changing information about the clips will change the way participants will rank them. On visual inspection of the data, there did not seem to be any effects supporting either of our two hypotheses. This was confirmed by the ANOVA tests. When looking at computer-composed pieces ranked in the deceived group compared to the rankings from the first phase, there was no significant effect $F(1,4) = 0.6901$, $p = 0.4087$. Looking at computer-composed pieces in the Informed group, there was also no effect $F(1, 4) = 1.0684$, $p = 0.3044$. There was no difference when looking at human-composed pieces. Human-composed pieces in the deceived group showed no significant difference $F(1, 4) = 0.6901$, $p = 0.4087$. Human-composed pieces in the informed group were also not rated differently in the second phase showing no significant results $F(1, 4) = 1.0684$, $p = 0.3044$. Given the high p-values of our repeated measures ANOVA, our results can not be deemed statistically significant.

A Kendall-Tau rank correlation was performed to check intra-rater rank agreement. Human-composed pieces in the informed group had a median intra-rater rank correlation of $\tau = 0.6$. Human-composed pieces in the uninformed group had a median intra-rater rank correlation of $\tau = 0.399$. Computer-composed pieces in the informed group had a median intra-rank correlation of $\tau = 0.6$ as well as computer-composed pieces in the uninformed group $\tau = 0.6$. According to Akoglu (2018), a $\tau$ of 0.6 would be considered either moderate or strong whereas a $\tau$ of 0.39 would be on the boundary of weak and moderate. Given our results are in this range we can conclude that intra-rater ranking is fairly consistent in our experiment.

## 5 Discussion

Along with the past work of Moffat and Kelly (2006), as well as Paquier et al. (2016), the present study sought to find out if humans hold a negative bias against computer-created music. As with past studies, it seems like the anticipated bias seems to be mostly anecdotal and not as strong as we may have presupposed. Although there seems to be a slight skew towards ranking human-composed clips as higher, this effect is not strong enough to be deemed statistically significant. The first hypothesis tested, that those in the informed group would rate human-composed music as better than computer-composed music, was not supported by the repeated measures ANOVA analysis and subsequent T-tests.

The second hypothesis this experiment tested sought to examine whether people's attitudes towards technology and artificial intelligence, in general, would affect their ratings of music that were thought to be computer-composed. Unlike the previous findings of Hong et al. (2020), we did not find any significant correlation between participants' comfort level when using the computer and their rankings of computer-composed music. It must be mentioned, however, that Hong et al. (2020) were searching for an effect in people's expectation violation, not their direct ratings. In addition to not finding any significant evidence for the second hypothesis, there did not seem to be any effect between people's musical capability and their rankings of computer-composed music. This lack of an effect goes against the previous findings of Moffat and Kelly (2006), which found musicians rated computer-composed pieces as lower than non-musicians.

A contributing factor to the lack of support for the second hypothesis may be the population used for the study. While MTurk provides an easily accessible population of participants from all over the world, a large portion of this population is highly comfortable with using technology given the online nature of the platform, 65 out of 161 participants indicated they are "very proficient with computers" and 15 participants indicated they were "experienced programmers". This led the population in our current study to not be necessarily representative of the world population. The majority of participants in the current study reported their ages as under 40, with only 2 participants identifying as 65+, leaving particularly older age groups largely unrepresented. Attitude towards new technologies is often stable within generations, which may make it difficult to convince some, especially older, age groups that AI music is equal to human-composed music (Chung et al. 2010; Niehaves and Platfout 2014). As shown by Deguernel et al. (2022), the context and culture surrounding the music can have an impact on the way it is perceived. Future work could potentially look to see if those of a higher age group (50+) would have any differences on their outlook to computational creativity, as people in this age range are typically less comfortable with new technologies.

Although sample sizes for the current study were adequate, the high variability in the data, as well as the method of ranking clips, suggests that the design should be simplified. Perhaps a more direct comparison, using fewer clips that sound more distinct from one another, could provide different results. Another factor was that some participants reported that having to drag the 10 clips in their desired order was cumbersome and not intuitive. This could have caused listeners to become confused about which clip they were ranking. It may be possible the length of the audio clips were quite short, which in conjunction with the similarity in style between clips, may have led to the high variability in rankings among pieces regardless of their authorship. Finally, although there was no evidence in the data to support this, there is a slight chance that human and computer-composed pieces were distinguishable to listeners even while blind to authorship. There was a subtle characteristic in the computer-composed tracks that set the note velocity to a single value, which made these tracks sound slightly louder. Although this was later corrected by remastering the tracks, there is a very slight chance that astute listeners would have been able to pick up on this minute characteristic.

### 5.1 Future Work

Although the present study failed to find any significant bias against computer-composed music, this does not mean that such a bias does not exist. Our results are in line with that of Pasquier et al. (2016), however, they go against the findings of Moffat and Kelly (2006), Hong et al. (2020), and Ragot et al. (2020). There does not seem to be a clear consensus on whether there is a general bias against computationally creative artwork and more work must be done on the topic in order to answer our hypothesis.

Future studies could improve on the current design by creating a more direct comparison of musical clips. Rather than presenting 10 different clips to the participant at once, clips could be presented in pairs to keep the design simple. Given that many participants were overwhelmed when trying to rank 10 clips at once, this could prove to be a much simpler and easier to understand design. Further work in this field, one that does not choose to use a ranking-based approach, for example, could perhaps implement a more well-established music rating system such as the "Rubric for assessing general criteria in a composition assignment" (Hickey 1999), which was used by Hong et al. (2020). Their scale measured multiple factors such as aesthetic appeal, creativity, and craftsmanship. Using such a scale could provide more information into what exactly people perceive differently between human-composed and computer-composed works or whether there is any difference at all.

Perhaps instead of looking for a general perceived bias, future work should focus on finding a specific inter-group bias when evaluating AI-created art or music, as suggested by Ragot et al. (2020). Because of the widespread fear of losing one's employment due to AI automated systems (McClure 2017), it may be the case that this plays a significant role on how people perceive music created by such systems. Perhaps people are scared that their own creativity, something perceived to be innately human, could eventually be replaced by a non-human entity.

## 6 Conclusion

This study sought to search for a negative bias against music created by artificial intelligence systems. Similar to the previous study by Pasquier et al. (2016), we did not succeed in achieving our goal of finding such a bias. We outlined the reasoning behind our lack of significant results and presented a number of possible improvements for future research in this field of study.

Research like the present is part of a wider exploration of human-computer interaction. As artificial intelligence technologies continue to permeate through society, more studies that examine the interactions between humans and these computer systems should be conducted. The general consensus for a bias against computers and AI, especially in the field of creativity, is widely held anecdotally, yet there is still not much work done proving the evidence of such a bias.

Furthermore, music has been shown to provide many health benefits and is used for multiple clinical purposes (Alty et al. 1997; Cross 2014; Hargreaves et al. 2005). By starting to gain a more in-depth understanding of the way humans perceive and interact with metacreative systems, perhaps it may one day be possible for AI to generate creative works for a specific purpose, such as music therapy. In the same way that Open-AI's DALL-E 2 is meant to help artists with their creative process, MuMe systems can aid in the creative process and even help form new genres of music. In the same way that digital audio workstations (DAW), such as Logic and FL studio, have revolutionized music production, MuMe systems provide an excellent tool for artists to come up with new ideas and put them into practice.

This paper discusses the current state of knowledge in the search for a bias against computationally creative systems and paves the way for future work to increase the growing scope of knowledge on the subject. Although we did not find any significant results in our experiment, there does still seem to exist such a bias against computational creativity as presented by Moffat and Kelly (2006) and Ragot et al. (2020), and Deguernel et al. (2022). More work is needed to come to a consensus on whether or not there exists a bias against computer composed music and computational creativity as a whole.

## References

Akoglu, H.: User's guide to correlation coefficients. Turkish J. Emerg. Med. **18**(3), 91–93 (2018)

Alty, J.L., Rigas, D., Vickers, P.: Using music as a communication medium. In: CHI'97 extended abstracts on human factors in computing systems, Atlanta, GA, 22–27 March, pp. 30–31. ACM, New York (1997)

Burgoon, J.K., Bonito, J., Lowry, P., et al.: Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. Int. J. Hum. Comput Stud. **91**, 24–36 (2016)

Brewer, W., Treyens, J.: Role of schemata in memory for places. Cogn. Psychol. **13**(2), 207–230 (1981)

Chung, J.E., Park, N., Wang, H., et al.: Age differences in perceptions of online community participation among non-users: an extension of the Technology Acceptance Model. Comput. Hum. Behav. **26**(6), 1674–1684 (2010)

Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. J. Soc. Issues **56**(1), 81–103 (2000)

Cope, D.: Virtual music: computer synthesis of musical style. MIT Press (2004)

Cross, I.: Music and communication in music psychology. Psychol. Music **42**(6), 809–819 (2014)

Déguernel, K. Maruri-Aguilar, H., & ; Sturm, B. L. T. (2022). Ken Déguernel, Bob L. T. Sturm, Hugo Maruri-Aguilar. Investigating the relationship between liking and belief in AI authorship in the context of Irish traditional music. CREAI 2022 Workshop on Artificial Intelligence and Creativity

Miranda, E.R.: Artificial intelligence and music: an artificial intelligence approach to sound design. Comput. Music J. **19**(2), 59 (1995)

Ens, J., Pasquier, P.: MMM: exploring conditional multi-track music generation with the transformer. arXiv preprint arXiv:2008.06048 (2020)
Fogel, A.L., Kvedar, J.C.: Artificial intelligence powers digital medicine. Npj Digital Med. **1**(1) (2018)
Goldman, S.: Will openai's dall-e 2 kill creative careers? VentureBeat (2022, July 26). Retrieved October 2, 2022. https://venturebeat.com/ai/openai-will-dall-e-2-kill-creative-careers/
Hargreaves, D.J., MacDonald, R., Miell, D.: How do people communicate using music. In: Miell, D., MacDonald, R., Hargreaves, D.J. (eds.) Musical Communication, pp. 1–26. Oxford University Press, New York (2005)
Hillhouse, A.N.: Tradition and innovation in Irish instrumental folk music (T). University of British Columbia (2005). Retrieved from http://open.library.ubc.ca/collections/ubctheses/831/items/1.0092099
Hello World album the first album composed with an artificial intelligence. SKYGGE. (n.d.). https://www.helloworldalbum.net/
Hewstone, M., Rubin, M., Willis, H.: Intergroup bias. Annu. Rev. Psychol. **53**(1), 575–604 (2002)
Hong, J., Curran, N.M.: Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. ACM Trans. (2019)
Hong, J.W., Peng, Q., Williams, D.: Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. new media & society (2020)
Istok, E., Brattico, E., Jacobsen, T., et al.: "I love Rocken' Roll"-music genre preference modulates brain responses to music. Biological Psychol. **92**(2), 142–151 (2013)
Jackson, T.: Imitative identity, imitative art, and AI: artificial intelligence. Mosaic: Interdisciplinary Critical J. **50**(2), 47–63 (2017)
Jennings, K.E.: Developing creativity: artificial barriers in artificial intelligence. Mind. Mach. **20**(4), 489–501 (2010)
Norton, D., Heath, D., ; Ventura, D.: Finding creativity in an artificial artist. J. Creative Behav. **47**(2), 106–124 (2013)
Marzano, G., Novembre, A.: Machines that dream: a new challenge in behavioral-basic robotics. Procedia Comput. Sci. **104**, 146–151 (2017)
McCarthy, J.: From here to human-level AI. Artif. Intell. **171**(18), 1174–1182 (2007)
McClure, P.K.: "You're fired," says the Robot. Soc. Sci. Comput. Rev. **36**(2), 139–156 (2017)
Moffat, D.C., Kelly, M.: An investigation into people's bias against computational creativity in music composition (2006)
Niehaves, B., Plattfaut, R.: Internet adoption by the elderly: employing IS technology acceptance theories for understanding the age-related digital divide. Eur. J. Inf. Syst. **23**(6), 708–726 (2014)
Pasquier, P., Burnett, A., Maxwell, J.: Investigating listener bias against musical metacreativity. In: Proceedings of the Seventh International Conference on Computational Creativity, pp. 42–51 (2016, June)
Ragot, M., Martin, N., Cojean, S.: Ai-generated vs. human artworks. A perception bias towards artificial intelligence? Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (2020)
Thorogood, M., Pasquier, P., Eigenfeldt, A.: Audio Metaphor: Audio Information Retrieval for Soundscape Composition. In: Proceedings of the 9th Sound and Music Computing Conference, Copenhagen (2012)

An ai-generated artwork won first place at a state fair fine arts competition, and artists are pissed. VICE. (2022, August 31). Retrieved January 30, 2023, from https://www.vice.com/en/article/bvmvqm/an-ai-generated-artwork-won-first-place-at-a-state-fair-fine-arts-competition-and-artists-are-pissed

Yannakakis, G.N., Martınez, H.P.: Ratings are overrated! Frontiers in ICT **2**, 13 (2015)

Xuan, W.Z., Ying, L.F.: The development of Celtic Music Identity: Globalisation and media influences. Media Watch **13**(1), 34–48 (2022)