Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation

Ken Déguernel, Bob L. T. Sturm

Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation

Ken Déguernel[1,2] and **Bob L. T. Sturm**[1]
[1]Royal Institute of Technology KTH, Lindstedtsvägen 24 SE-100 44 Stockholm, Sweden
[2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Abstract

This paper surveys 27 published studies exploring bias in the evaluation of computational creativity. These studies look specifically at the involvement of AI, in generating music, images and graphics, poetry, and news articles. While some have found evidence of bias (43%), others find no bias (27%), or show modulation of bias through socio-cultural factors (30%) resulting in a lack of consensus on this issue. We argue for the importance of taking into account socio-cultural context when considering such biases in these creative pursuits. What styles do the artefacts belong to? Who are the participants involved in the study, and what are their relationships to the styles at hand? We discuss the implications of such considerations for future research in computational creativity. We propose some safeguards when conducting a study on bias in the evaluation of computational creativity, and propose directions to study more specifically when, and with whom it can be observed.

Introduction

Artificial intelligence (AI) has been applied in a wide variety of artistic fields such as poetry (Hämäläinen and Alnajjar, 2019), painting (Ramesh et al., 2021), comedy (Strapparava and Stock, 2011), and music (Herremans, Chuan, and Chew, 2018), and other more technical fields such as journalism (Broussard et al., 2019) or programming (Li et al., 2022). Applications of "computational creativity" have reached the attention of the general public through popular tools for generating free-form text (Brown et al., 2020), and generating images from textual descriptions (Rombach et al., 2021).

Human appreciation of creativity and its results is influenced by many factors, such as age, gender, personality, and expertise, but is also influenced by external factors regarding knowledge and context of production (Davies, 2003; Steinbeis and Koelsch, 2009) and socio-cultural factors, such as values and practice. Since knowledge of the production process is an important evaluative criterion (Lamb, Brown, and Clarke, 2018), a bias could exist when it comes to knowing or thinking that an artefact arises from computational creativity. Knowledge about such bias is moreover important when it comes to the evaluation of such creative systems as the appreciation of the artefacts they produce is sensitive to many subjective criteria. While the evaluation of computationally creative systems based on how close the artefacts they produce come to human-created ones can provide valuable insights, it encourages "superficial imitation" (Pease and Colton, 2011), and fails to take into consideration external factors that could trigger some bias in favour or against AI that could influence those results. As such, it is difficult to say whether an artefact generated (even partially) via computational creativity can be evaluated the same way as any other human-created artefact (Ariza, 2009).

One can see the bias discussed above in the frame of *algorithmic aversion*, a phenomenon where individuals have a negative attitude or mistrust towards AI systems (Dietvorst, Simmons, and Massey, 2015). This can manifest in various ways, such as resistance to using tools or services involving AI, scepticism about AI-generated decisions, and concerns about the impact of AI on society (Flick and Worrall, 2022). With a survey of 80 studies, Mahmud et al. (2022) identifies factors linked to algorithmic aversion: algorithmic factors such as the explainability of the algorithm, its presentation and accuracy; individual factors such as personality, psychological factors and familiarity with algorithms; high-level factors such as by whom algorithms are being used (e.g., banks, for-profit organizations) and social influences; and finally, task factors as in what the algorithms are used for. The opposite of this phenomenon is called *algorithmic appreciation* (Logg, Minson, and Moore, 2019).

In this paper, we review 27 papers describing studies in which quantitative analysis is applied to detect and measure bias for the task of music generation, image and graphic generation, and text generation. We propose potential explanations for when bias is (or is not) observed, such as the lack of accounting for contextual factors through the selection of study participants, or the presentation of artefacts with respect to their use. We discuss the implications of these results for future studies on bias against computational creativity, and on the evaluation of such systems.

A survey of contradicting results

This section surveys the results of all studies (to the best of our knowledge) explicitely attempting to measure bias in the evaluation of computational creativity for music generation, graphics and images, poetry, and journalism, in order to observe a variety of media, and both artistic and factual scopes. Table 1 summarizes these 27 publications.

Task	Paper	Style/Topic	N	Reported conclusion
Music	Dahlig and Schaffrath (1998)	German Folk songs	432	Varied
	Moffat and Kelly (2006)	Contemporary, free jazz, Bach	20	No bias
	Friedman and Taylor (2014)	Classical	58	No bias
	Pasquier et al. (2016)	Contemporary	122	No bias
	Jago (2019)	Song	200	Bias
	Hong, Peng, and Willians (2021)	EDM, classical	299	Bias
	Moura and Maw (2021)	Pop, classical	86	No bias
	Aljanaki (2022)	Classical	20	Bias
	Déguernel, Sturm, and Maruri-Aguilar (2022)	Irish traditional music	46	Bias
	Shank et al. (2022)	Classical	136	Bias
	Hong et al. (2022)	Rock, EDM, classical, country	222	No bias
Graphics and Images	Kirk et al. (2009)	Modern art	14	Bias
	Norton, Heath, and Ventura (2015)	Abstract	284	Bias
	Chamberlain et al. (2018)	Abstract, representational	65	Bias
	"	Portrait	349	Varied
	Hong (2018)	Abstract	28	Varied
	Hong and Curran (2019)	Abstract/Psychedelic	288	No bias
	Jago (2019)	Painting	201	Bias
	Ragot, Martin, and Cojean (2020)	Landscape/Portrait	565	Bias
	Wu et al. (2020)	Modern	544	Varied
	Gangadharbatla (2021)	Abstract, representational	530	Varied
Poetry	Wu et al. (2020)	Modern	544	Varied
	Hitsuwari et al. (2023)	Haiku	385	Varied
Journalism	Clerwall (2014)	Sport	46	No Bias
	van der Kaa and Krahmer (2014)	Sport, finance	252	Varied
	Graefe et al. (2018)	Sport, finance	986	Bias
	Waddell (2018)	Political news	311	Bias
	Liu and Wei (2019)	Spot news, interpretive news	355	Varied
	Longoni et al. (2022)	Headlines	3029	Bias
	Lermann Henestrosa, Greving, and Kimmerle (2023)	Popular science	469	No bias

Table 1: Summary of studies on bias in the evaluation of Computational Creativity with their respective topics studied, number of participants (N), and conclusions. "Varied" indicates that the existence of bias is modulated by socio-cultural factors.

Music Generation

One of the first experiments in this area is that of Dahlig and Schaffrath (1993, 1998). A participant listens to a melody and rates the degree to which it is "original" or "computer made" and whether they like it. As stimuli they use computer syntheses of eleven melodies: two German folk melodies and eleven melodies created by mixing up phrases of German folk melodies à la *Musikalisches Würfelspiel*. They report having 432 respondents, drawn from a variety of populations, including musicologists and young students from schools in Germany and China. From the results they conclude, "the biggest number of positive aesthetic evaluations was accorded to melodies regarded to be authentic. Contrariwise, melodies 'suspected' to be computer-made got the biggest number of negative evaluations".

Moffat and Kelly (2006) describes a two-stage listening experiment where participants assess six 1-minute excerpts of music. In the first stage, a participant listens to each excerpt and answers questions such as, "How much do you like this sample?" and "Do you think it was composed by a human or by a computer?" In the second stage, the participant is told the true origin of each excerpt (name of composer or computer system), and is asked questions such as, "Would you buy this piece of music?" Three of the excerpts are of computer-composed music in the styles of "Bach", "free-form jazz", and "pieces for strings", and three are of human-composed music in the same styles. They report from data collected from 20 participants that while they find "there is a common bias against computer-generated pieces", and that "[i]n almost every case, a piece of music is preferred when it is thought to be human-composed", they do not observe any significant differences between the rating of liking (stage 1) and enjoyment (stage 2) after a listener is told the origin of the music.

Pasquier et al. (2016) extends the study of Moffat and Kelly (2006), and builds upon past work in evaluating creative systems (Eigenfeldt, Burnett, and Pasquier, 2012). In the study, a participant listens to a music excerpt and rates their perception of it on four dimensions: "Good–Bad", "Like–Dislike", "Emotional–Unemotional", and "Natural–Artificial". Each participant listens to and rates each excerpt twice in the experiment, but in one of three different conditions. A participant in the "fully informed condition" is told about the origin of each excerpt. A participant in the "fully naïve condition" is never told about the origin of each excerpt. And a participant in the "revealed condition" is only told about the origin of each except after listening to and rating all excerpts once. They use 1-minute excerpts of six "contemporary string quartets", three composed by a human and three generated by an AI system. They report from 122 participants (university students) that "[w]hile our results do indicate a negative effect of the knowledge of computer authorship on listener judgements, this effect is not significant".

Friedman and Taylor (2014) describes a study where a participant listens to a music recording, and then rates several qualities, e.g., arousal, liking and quality. The participant then decides whether the piece was composed by computer or human, and whether it was played by computer or human. Each participant is assigned to one of two conditions: either the participant is explicitly told every music recording was composed and performed by a computer; or the participant is explicitly told every music recording was composed and performed by a human. The study uses synthesized recordings of four human-composed classical piano pieces of between 1.6 and 3.4 minutes duration. From an analysis of over 190 participants (undergraduate psychology students), they conclude, "the perception that the music

was computer-generated did not significantly alter participants' emotional responses or their judgments of the quality of what they had heard."

Jago (2019) presents a study where a participant listens to a 30-second recorded music excerpt and rates their perception of the "authenticity" of the work. In one condition, the participant is told the work is by a particular person. In another condition, the participant is told the work is by a particular AI. Four different music excerpts are used, each generated by the same AI system; but a participant only rates one excerpt. Based on the responses of 200 participants (from Amazon Mechanical Turk (Mturk) users in the USA), Jago (2019) concludes that the participants "believed that human work was more authentic, compared to an artificially intelligent algorithm's otherwise-identical work."

Moura and Maw (2021) describes a study where a participant reads a narrative about the music they will hear, then listens to two 1.5-minute music recording excerpts, and then answers questions about the experience. All participants listen to the same music, but each is assigned to one of two groups, corresponding to a particular narrative. One narrative states the music is composed by AI, and the other describes emotions and experiences reflected in the music, implying a human composed the music. The two excerpts are "pop-rock" and "classical" styles, each arising from human-AI collaboration. Based on the responses of 86 participants (German university students) they report no significant differences in responses between the two groups for either music excerpt, and conclude, "listeners' awareness of the nature of the composition process (human versus AI) posed no significant impact on participants' perceptions towards the songs [...] regardless of the different music genres."

Aljanaki (2022) discusses a study where a participant listens to recordings of two pieces for piano and rates each. All participants listen to the same music, but each participant is assigned to one of two groups: in one the real origin of each piece is given; in the other the origin is reversed. One piece is modern and composed by a human, and the other is composed by a machine, "reminiscent of romantic period in classical music". From the responses of 20 participants ("non-musicians"), Aljanaki (2022) concludes that the difference between responses of the groups was not significant.

Déguernel, Sturm, and Maruri-Aguilar (2022) describes a study where a participant listens to six music recordings, rates their liking of each, and then listens to them again rating their belief of each being composed by a computer. The six music recordings feature the same musician playing six different computer-generated "double jigs", (a form of Irish traditional dance music). Based on the responses of 46 participants (practitioners of Irish traditional music), they conclude that the practitioners "tend to like more the tunes they deem hardly likely to be composed by an AI. Alternatively, the more they report liking a tune the less they report believing the tune is AI-composed."

Hong, Peng, and Willians (2021) describes an experiment where a participant listens to a music recording and is told it is composed by either AI or a human, and is then asked to evaluate the music. Each participant is given only one of the four pieces and one of the possible origins. Four AI-composed pieces are used, two of the type "classical" and two of the type "EDM" (electronic dance music), each generated by the same AI system. Based on the responses of 299 participants (found using Mturk), they conclude, "accepting the creativity of AI is a prerequisite for a positive evaluation of its artistic merit ... [A]n unwillingness to accept AI products blocks appreciation."

Hong et al. (2022) presents a study where a participant reads a "mock" news article about an AI music generation system, then listens to a piece of music presented as composed by that system, and finally rates their experience of the piece. There are four different news articles and four different pieces. Each news article describes a different level of anthropomorphism and algorithmic independence of the AI music generation system. The four different pieces are composed by the same AI system, but in the styles "rock", "EDM", "classical" and "country." Each participant is randomly assigned a news article and a piece. Based on the responses of 222 participants (found using Mturk), they conclude that neither aspect have any significant impact on the ratings of the music.

Shank et al. (2022) presents three studies investigating the relationship between reported music liking and belief in AI authorship. In the first study a participant listens to twenty 15-second excerpts of human-composed music, and after each is asked whether it was composed by a human or AI and their confidence so, and is finally asked to rate how much they like the music. Each participant is given excerpts of either the music type "classical" or "electronic". Based on the responses of 295 participants (found using Prolific), they conclude that "music that was perceived as being composed by an AI was liked less than music that was perceived as being composed by a human". In a second study, a participant listens to eight 15-second excerpts of human-composed music, and after each is asked to rate their liking of it and its qualities. These specific excerpts were selected based on the responses to the previous study: four electronic music excerpts were selected as sounding the most "AI", and four electronic music excerpts were selected as sounding the most "human". The participant is assigned to one of three conditions. In the first, they are told all excerpts are composed by AI composing software. In the second, they are told all excerpts are composed by various composers. In the third, they are not told of the origin of the music. Based on the responses of 399 participants (found using Prolific), they do not find a significant effect of the purported origin on participant liking. They then present a third study where a participant listens to eight 15-second excerpts of human-composed music, and after each is asked to rate their liking of it and its qualities. These specific excerpts were selected based on the responses to the first study: the classical music excerpts sounding the most "human." The participant is told beforehand that some of the excerpts were composed by AI software. Each excerpt is presented as being composed by a specific person or a specific AI system. Based on the responses of 136 participants (found using Prolific), they conclude that "participants rated the music as both lower quality and liked it less if it were purportedly composed by an AI."

The conclusions from these 12 papers show a clear lack

of consensus on the existence of a bias in the evaluation of music generation systems. This could be explained by the use of different musical style (although different results have been found for classical music (Friedman and Taylor, 2014; Shank et al., 2022)), and the use of different criteria of evaluation and presentation of the algorithm (Hong, Peng, and Willians, 2021). Déguernel, Sturm, and Maruri-Aguilar (2022) also suggests a potential role of expertise and familiarity as a modulating factor of such a bias.

Graphics and images

Kirk et al. (2009) presents a study where a participant views a digital image of an abstract painting together with a text label showing its origin, and rates its pleasantness (aesthetic rating scale). Each participant is told they will see 200 abstract paintings, that half of them are from a famous gallery, and that half are generated by the experimenter using computer software. The 200 digital images were "selected from online sources" by the experimenters, and all appear to be human-created. Based on the responses of 14 participants (university students in Denmark), they conclude that "images under the gallery label were rated as having a significantly higher mean aesthetic value than those carrying the computer label."

Norton, Heath, and Ventura (2015) discusses a study where a participant views a pair of images (processed digital photographs) – one labeled as created by a human and the other labeled as created by a computer program – and selects the one they believe is a better image. All images for fifteen pairs were generated by the same computer program, and selected by the experimenters. From 330 responses collected online, they conclude that there was "a small but substantial bias either towards humans or against [the algorithm]."

Chamberlain et al. (2018) describes a study where a participant is shown in random order sixty digital images and is asked after each how much they like it, and then shown the images a second time and asked after each if they believe the image is man-made or computer-generated. A participant in a reversed condition is asked first if they believe an image is man-made or computer-generated, and then how much they like it. Half of the images were selected by the experimenters from online "computer art databases" being of types "abstract" or "representational", and the other half were of man-made artwork of the same types. Based on the responses of 65 participants (students and staff at KU Leuven), they conclude that for either condition "images that were categorized as computer-generated were rated as visually less pleasing." Chamberlain et al. (2018) describes a second experiment where a participant evaluates drawn human portraits made by a robot artist (a table-mounted animatronic arm holding a pen which makes marks on a piece of paper). Some participants see the robot and its artworks; some participants are just told about the robot and shown the artworks; and some participants are only shown the artworks and not told anything about them. Each participant answers a survey about their aesthetic responses. Based on the responses of 349 participants in the three conditions (attendees of the art gallery, and KU Leuven students and staff), they conclude the bias observed in the first experiment "can be moderated by interaction with the agents of the artwork. The presence of the robotic artists had a strong positive impact on aesthetic evaluations of the resulting artworks."

Hong (2018) describes a focus group in which participants view a digital image of an artwork and discuss questions about art and the involvement and relevance of AI. In one condition the group is told that the image they are viewing was produced by AI. In the other condition, the group is told it was produced by a human. Both groups view the same image, which was created by a human artist. From the discussion of the 14 people in each group (students at the University of Southern California), Hong (2018) concludes that the group being told the image they are viewing was produced by AI had "a stronger tendency toward the belief that AI cannot produce art," and that "one way to diminish a negative stereotype toward artificial intelligence being creative is to successfully persuade the public its autonomy" – which echoes the finding with the perception of robot artists in Chamberlain et al. (2018).

Hong and Curran (2019) presents a study where a participant views a digital image of an abstract artwork and then rates it along eight dimensions, e.g., originality, composition, and aesthetic value. There are four groups of participants, crossing factors of attribution knowledge (being told the images are created by AI, or not being told anything about human or AI authorship), and image source (images are generated by AI, or images are human created). Participants in the groups being told images are created by AI view the same set of six images; and the participants in the other groups view a different set. Six of the twelve images used are generated using three AI systems. The remaining images are of six human-made paintings, selected by the authors for sharing stylistic and thematic similarities with the AI-generated images. In each set of six images viewed by a group, half are from AI systems. From the responses of 288 participants (from Mturk) they conclude that "[the] evaluation of aesthetic value is done independently from bias related to the artwork and its artist."

Jago (2019) study, presented in the previous section, also have a participant sees and rate a digital image of a painting with the same procedure described above. Based on the responses of 200 participants (from Mturk users in the USA) Jago (2019) concludes again that "they believed that human work was more authentic, compared to an artificially intelligent algorithm's otherwise-identical work."

Wu et al. (2020) presents a study exploring the explicit and implicit attitude towards AI-generated paintings. Participants are shown a digital image of either a human- or AI-created painting, then asked to rate it on quality, imaginativeness, spatial presence, empathy, competence, and finally to rate their attitude towards AI. To take into account the implicit bias, participants are given an alleged human or AI origin for the piece they are evaluating. Based on the responses from 251 U.S. participants and 293 Chinese participants they report that U.S. participants were more critical to AI-generated art compared to human-generated content both explicitly and implicitly, whereas Chinese participants exhibited overtly positive attitudes towards AI-generated content, yet their implicit acceptance of it was lower than that

of human-generated content.

Ragot, Martin, and Cojean (2020) discusses a study in which a participant views a digital image of an artwork and rates it along four dimensions, e.g., liking and novelty. Participants in one group are "primed" with information that the artworks they will see were created by "some artificial intelligence", and in the other group that the artworks were created by "some artists". Each participant views 8 images, selected at random by the authors from 40 curated images: "10 portraits by AI, 10 landscapes by AI, 10 portraits by humans, and 10 landscapes by humans". Both human and AI artists were involved, Based on responses of 486 participants (from Mturk) they conclude "the artworks presented as AI-generated paintings were significantly less liked and were perceived as less beautiful, novel, and meaningful than paintings presented as drawn by a human."

Gangadharbatla (2021) describes a study where a participant views a digital image of an artwork and then rates nine characteristics of it, including creativity, aesthetic value and financial value. In one condition, a participant is given prior information that the images were generated by AI without human involvement. In another condition, the prior information relates to the human production of the artworks they will see. Each participant views the same four images of two types of art: "representational" and "abstract". One work in each type is human-created and the other is AI-generated. Based on responses of 530 participants (from Mturk) they conclude that "attribution knowledge [plays] a significant role in influencing individuals' evaluations of artwork."

The conclusions from these 9 papers similarly display a lack of consensus on the existence of a bias in the evaluation of artwork generation systems, with different results for the same types of artworks. Several factors of modulation of bias are found in those studies however. Personal factors such as culture, identified by Wu et al. (2020), algorithmic factors depending on how the system is presented or observed, identified by Hong (2018); Chamberlain et al. (2018), and contextual factors such as where the experiment is conducted, identified by Chamberlain et al. (2018).

Poetry

Wu et al. (2020) presents a study exploring the explicit and implicit attitude towards AI-generated poems, using the same procedure as for the graphics generation described above. Based on responses from 251 U.S. participants and 293 Chinese participants they conclude the same: that U.S. participants were more explicitly and implicitly critical to AI-generated poetry compared to human-generated content; and Chinese participants exhibited overtly positive attitudes towards AI-generated poetry, yet their implicit acceptance of it was lower than that of human-authored poetry.

Hitsuwari et al. (2023) describes a study consisting of two blocks: first, a rating block where a participant rates their liking of haikus according to 21 criteria such as beauty, valence, arousal, and novelty; and then a discriminating block where a participant is asked whether they think the haiku was created by AI or a human, and what criteria they use to make their decision. In one condition, a participant rates poems first and then predicts the author. In the other condition, these tasks are reversed. Stimuli are either human-made, AI-made, or made with a "Human in the loop". Based on the responses from 385 participants (Japanese recruited through CrowdWorks), they report that "task order (i.e., prior knowledge about whether the work was produced by AI) did not affect the evaluation of the beauty of haiku". However, the more beautiful a haiku was rated, the more likely it was believed to be created by a human.

Both those studies show a modulation of bias in the evaluation on poetry generation systems. On the one hand, Wu et al. (2020) identify culture as a personal factor and on the other hand, Hitsuwari et al. (2023) identify the presentation of the systems as an algorithmic factor modulating bias in evaluation.

Journalism

Clerwall (2014) describes a study where a participant reads a written account of a sports game, evaluates the article according to 12 descriptors (e.g., objective, trustworthy, and informative), and then is asked whether they think the text is human- or computer-written. Each account is either generated by a computer or written by a journalist. Based on the responses from 46 participants (undergraduate media students), Clerwall (2014) reports no significant differences on how the groups evaluated or perceived the articles.

van der Kaa and Krahmer (2014) replicates the study of Clerwall (2014) with news topics of sports and finance. In their study, however, participants are given an alleged source for the article: either a journalist or a computer. Participants rate the article according to the same 12 descriptors. Based on the responses from 188 Dutch news consumers and 64 professional journalists, they conclude there were "no differences in the perceptions of news consumers" depending on authorship attribution, and that "news consumers have no strong negative or positive feelings toward computer-written news". On the other hand, "[j]ournalists perceive themselves as more trustworthy compared to their 'computer colleagues"', showing an impact of expertise. They also note a difference in the perceived level of trustworthiness depending of the topic of the article.

Graefe et al. (2018) replicates the study of van der Kaa and Krahmer (2014). Based on the responses of 986 German-speaking participants (recruited through SoSci Panel), they report that "articles are consistently perceived more favorably if they are declared as written by a human journalist, regardless of the actual source".

Waddell (2018) discusses two studies in which participants rate the accuracy and credibility of news article. In the first study, participants read a data-driven news article about politics attributed to a known news source. Participants are randomly assigned a condition in which the article is attributed to a specific journalist or to a "robot reporter". The second study replicates the first one but participants read articles about the weather, stock market, and science, and are also asked to fill in a "robot recall" questionnaire in which they are asked to recall a film or a show which involve a robot as a main or supporting character, and answer questions about how "good or bad", "human-like", this character is. Based on the responses from 129 in the first study and

182 participants in the second study (all recruited through Mturk), Waddell (2018) reports that "news attributed to a machine is perceived as less credible than news attributed to a human journalist". This effect is still present after the "robot recall", although it is slightly modulated by it.

Liu and Wei (2019) describes a study with two news organisations two news types (sport and interpretive news), and two alleged writers (AI or human). Participants first indicate their political values according to a questionnaire, and then are asked to read one randomly selected news article using the templates from the newspapers' websites and with the alleged identity of the writer indicated at three places in the article. Participants then rate their emotional involvement and perception of the news article. Based on the responses from 355 U.S. participants (recruited through Mturk), they report that AI author attribution induces less emotional involvement, is perceived of less expertise, but is also perceived as more objective. Moreover, "for a media organization whose news was more trusted, utilizing news-writing bots enhanced perceived news objectivity. Otherwise, employing bots further reduced perception of the writer's trustworthiness and expertise", showing an effect of the context and of the participants' political opinions.

Longoni et al. (2022) describes two studies in which participants rate the trustworthiness and accuracy of news headlines. In the first, participants are randomly assigned to a condition where they see the news items tagged as written by AI, or a condition where they see the news items tagged as written by human. In the second study, participants see news items tagged both ways. Some headlines are true and some are false. Based on the responses from 3029 participants in the first study, and 1005 participants in the second study (all recruited through Lucid), they report that the effect of source attribution has a significant effect on trust and on perceived accuracy.

Lermann Henestrosa, Greving, and Kimmerle (2023) presents a study testing if the credibility and trustworthiness of popular science articles can be influenced by human or AI attribution of authorship. Participants are given a popular science article with either a neutral or evaluatively positive tone, and are asked to rate it on 19 criteria judging message credibility and 12 criteria judging perceived trustworthiness. In one condition, participants are told the article was written by a journalist; in the other, they are told the article was written by a computer algorithm. In both cases, alleged sources are provided and participants are told that the articles had been published in a reputable newspaper. Based on the responses from 469 participants (recruited using Prolific), they report that although the tone of the article had a significant impact on the way articles were perceived, this effect is independent of authorship attribution, which has no significant impact on the evaluation of credibility and trustworthiness. They note, "Although participants made a clear difference in how they perceived the alleged authors, this difference was not at all reflected in their evaluation of the message"

The conclusions from these 7 papers show a lack of consensus on the existence of a bias in the evaluation of news article generation systems. Several factors of modulation of bias are found: Contextual factor such as the topic of the article, as found by van der Kaa and Krahmer (2014), and personal factors such as expertise, identified by van der Kaa and Krahmer (2014), and political stance, identified by Liu and Wei (2019).

Discussion: The importance of the socio-cultural context

The previous section refers to 27 papers related to measuring bias in the evaluation of Computational Creativity. It is apparent that there is no clear consensus. Although it is possible that the differences in the outcomes is only due to differences in the methodologies, we believe that these discrepancies are better explained by the different socio-cultural contexts in which the studies were conducted. What styles/topics do the evaluated artefacts belong to? Who are the participants involved in the study? And what are their relation to the styles/topics at hand? In this section, we describe how socio-cultural factors influence art appreciation and our relation to creativity, and how this can be a dominant factor in the evaluation of Computational Creativity.

Contextual factors

Art appreciation is influenced by many properties of the artefact itself (Koelsch, Vuust, and Friston, 2019; Obermeier et al., 2013; Hagtvedt, Patrick, and Hagtvedt, 2008), and can be modulated by personal individual factors (Orr and Ohlsson, 2005; Dubnov, Burns, and Kiyoki, 2016; Hitsuwari and Nomura, 2022). There is empirical proof that knowledge of extra-artistic factors influences the way we perceive art (Leder and Nadal, 2014; Greasley and Lamont, 2016). For instance, Brieber et al. (2014) shows that the setting in which art is experienced influences one's appreciation of it. Art is found more interesting and viewed longer in a museum than in a laboratory setting. Similarly, North, Hargreaves, and Hargreaves (2004) observes that people change their listening habits depending on the time, the activity they are doing, or their location. Flôres and Ginsburgh (1996) shows that the order of music performances had a significant correlation with the ranking of the professional juries in a competition, given performances occurring at the end of the competition an advantage. These kinds of contextual factors also have an impact even when it is only based on belief. For instance, Lauring et al. (2016) shows that for 'art-naïve' students, social priming, i.e., saying that a group of other students or art professionals rated positively or negatively an artwork, or giving alleged price information about an artwork, has a significant impact in their liking rating. Similarly, belief that a piece of music is composed by a well-established artist (Fischinger, Kaufmann, and Scholtz, 2018) or performed by a renowned musician (Kroger and Margulis, 2016) bias a listener's reported appreciation.

Culture and expertise

As described by Lubart (2010), the definition and conceptual boundaries of creativity is dependent on culture, which define on the one hand what and who can be considered creative, but also the "why and how" of creativity: "Culture is

omnipresent, and for this very reason its impact is often underestimated." The impact of this have been raised recently in the scope of AI ethics by Huang, Sturm, and Holzapfel (2021) regarding applications of AI to music, showing once again the importance of culture in the "why and how" of Computational Creativity.

Cultural familiarity has been shown to have an impact on the perception and appreciation of specific characteristics of art. For instance, Maher (1976) shows that musical intervals that would be considered very dissonant in Western culture appear in Indian classical music, and trigger different responses depending on the familiarity of the participants. A more extreme case of this phenomenon has been shown by McDermott et al. (2016), who observe that "consonant" or "dissonant" harmony is not a characteristic that matters for the music appreciation of native Amazonians. Lahdelma and Eerola (2020) recommend controlling for cultural familiarity and musical expertise for studies involving the perception of music dissonance.

Expertise is another factor that influences art perception and appreciation. Winston and Cupchik (1992) studies the aesthetic assessments of art-naïve and experienced students when shown "popular art" and "high-art paintings". They show that while art-naïve students prefer popular art, and experienced students prefer high-art paintings, the evaluation criteria used by each group are different: art-naïve students report more their emotional responses to artworks, while experienced students focus more on objective and structural properties of the artworks. Pearce (2015) describes a similar phenomenon for music, showing that although musical expertise is not significantly correlated to emotional experience, it has an impact in the processing of long-term musical structure, and on the aesthetic judgement of consonance and dissonance, and of musical complexity.

Darda and Cross (2022) studies the impact of cultural familiarity on the evaluation of Indian and Western visual art (painting and dance). Indian participants (21 experts, 24 non-experts) and Western participants (21 experts, 26 non-experts) are shown abstract and representational paintings and dance videos belonging either to Indian culture or Western culture. Participants are asked to rate the stimuli according to familiarity, complexity, evocativeness, abstractness, technical competency, beauty and liking. They report that cultural familiarity creates a in-group bias for dance (although, the same is not found for painting) and that there is a preference for representational art. However, the in-group bias is modulated by expertise as it is only found in art-naïve participants. Similarly, the preference for representational art is also modulated by expertise but only for Western participants. This study shows the intertwined relationship and disparity between cultural familiarity and expertise, which creates a complex system to consider when evaluating bias.

In regards to the evaluation of Computational Creativity, our survey shows that cultural familiarity and expertise has an impact. Wu et al. (2020) shows the impact of cultural background on the explicit and implicit attitude towards AI-generated poems and paintings, reporting a difference between U.S. and Chinese participants regarding their explicit acceptance of AI-generated contents and general attitude towards it. van der Kaa and Krahmer (2014) does not find differences in the perceptions of 188 news consumers depending on AI or human authorship attribution, but that differences appear in the perceptions of a group of 64 professional journalists.

The Product and the Process

The four P's of creativity, (Rhodes, 1961; Jordanous, 2014) distinguish between the 'Product' (the artefact produced by creativity), and the 'Process' (the set of actions taken leading to the production of an artefact). Although, we should keep in mind that different cultures and practices will focus more or less on the Product or the Process (Lubart, 2010), there is evidence that knowledge of, or even belief in, the production process of an artefact and the context in which it is produced is an important evaluative criterion for the resulting artefact (Chamberlain et al., 2018).

Compelling evidence of such a phenomenon is provided by Davies (2003), showing that one would not appreciate or value in the same way an original piece of art, truly novel for its time, reaching new frontiers of craftsmanship, and a newly made replica of it (of whatever fidelity), as has been shown by numerous cases of forgery (Bowden, 1999). Wolz and Carbon (2014) test this by showing participants artworks labelled as original or copies, showing that the alleged authenticity has a major impact on art appreciation. Another example is provided for music by Canonne (2018) who conducts a qualitative study where musicians listen to the same audio recording of a duet, but are either told they are listening to a composition or an improvisation. The interviews show that musicians' experience of the piece is very different in each condition, focusing on different aspects of the music, listening more to the acoustical features and overall structure when they believe they are listening to a composition, and listening more to the relational process and the interactions between instruments when they believe they are listening to an improvisation.

Related more directly to bias about the evaluation of Computational Creativity, research in neuropsychology (Steinbeis and Koelsch, 2009) shows that believing that an artefact is human-made (as opposed to AI-generated) activates areas of the cortex reported for mental state attribution, indicating that participants are engaging with the process and intentions of the alleged human artist. Moreover, as described in our survey, Chamberlain et al. (2018) discusses a study where participants observing the drawing process of a robot in a museum setting for as long as they want show a change in the parity of their assessments. This raises questions on the importance of audience engagement (Candy and Bilda, 2009) with the process and the product in this kind of study.

Safeguards and future directions

Considering the importance of socio-cultural contexts, we propose reframing the question of bias and Computational Creativity in order to better take context into account. Inspired by Lincoln and Guba (1985); Li (2004), we propose safeguards for future studies about bias in evaluating Computation Creativity:

- *Use thick descriptions:* In order to compare the research context of a study with those of other studies, we recommend using thick descriptive information regarding the methodology and the context. In particular, as many detailed information should be given regarding the stimuli used, the participants of the study, their relationship to the stimuli's style/topic, where and how the study was conducted, and other relevant contextual information (Ponterotto, 2006).
- *Refrain from generalizing:* As social and behavioural phenomena are bound by their specific contexts, we advise refraining from making generalizations about the results of a study about bias outside of the context studied even when results appear to be "statistically significant".
- *(Near-)Natural situation:* We strongly advise to closely align the research context with the artistically-relevant environment. A study should minimize external interference or changes that could be introduced as a result of the research. This applies to both the content of stimuli – which should be as 'natural' as possible regarding the style/topic – and the environment in which the stimuli are observed.
- *Triangulation:* We advise using triangulation as a mean of verifying both their data and interpretations. This involves using multiple sources of data, as well as different evaluation methods. Various data collection methods could be used such as surveys and interviews. One way of doing this is asking participants about their strategies for discriminating between human and AI authorship, as done by Déguernel, Sturm, and Maruri-Aguilar (2022); Chamberlain et al. (2018); Hitsuwari et al. (2023)

What does this entail for future research in Computational Creativity? First, regarding the evaluation of creative systems (Lamb, Brown, and Clarke, 2018), if biases need to be taken into account, then it means that a specific study regarding the specific context (or as close as possible) should be conducted, as studies conducted in different socio-cultural contexts may yield results that are irrelevant to the target domain. Therefore, Computational Creativity evaluation could actually become a great experimental ground, like in Norton, Heath, and Ventura (2015) or Hitsuwari et al. (2023), for better understanding the whys and wherefores of *algorithmic aversion* and where *algorithmic appreciation* arises in the scope of creativity, as they offer a large variety of applications with their respective socio-cultural contexts. Second, the results from all the studies presented in our survey show that questions around the existence of *algorithmic aversion* as a general direction of research in computational creativity, might not be the correct framing for future work. Instead, the field might focus more of the "what? where? when? who? and how?" of such biases, as this will lead to a better understanding of the impact of creative systems in society and help to lead more informed discussions in regard to AI ethics (Holzapfel, Jääskeläinen, and Kaila, 2022).

Acknowledgments

This work was supported by the grant ERC-2019-COG No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism.

References

Aljanaki, A. 2022. Attitude towards and evaluation of computer-generated music in music listeners and musicians. In *Proc. of the Conference on AI Music Creativity*.

Ariza, C. 2009. The interrogator as critic: The Turing test and the evaluation of generative music systems. *Computer Music Journal* 33(2):48–70.

Bowden, R. 1999. What is wrong with an art forgery?: An anthropological perspective. *The Journal of aesthetics and art criticism* 57(3):333–343.

Brieber, D.; Nadal, M.; Leder, H.; and Rosenberg, R. 2014. Art in time and space: Context modulates the relation between art experience and viewing time. *PloS one* 9(6).

Broussard, M.; Diakopoulos, N.; Guzman, A. L.; Abebe, R.; Dupagne, M.; and Chuan, C.-H. 2019. Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly* 96(3):673–695.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; et al. 2020. Language models are few-shot learners. *CoRR* abs/2005.14165.

Candy, L., and Bilda, Z. 2009. Understanding and evaluating creativity. In *Proc. of the 7th ACM Conference on Creativity and Cognition*, 497–498.

Canonne, C. 2018. Listening to improvisation. *Empirical Musicology Review* 13(1–2).

Chamberlain, R.; Mullin, C.; Scheerlinck, B.; and Wagemans, J. 2018. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity and the Arts* 12(2):177–192.

Clerwall, C. 2014. Enter the robot journalist: users' perceptions of automated content. *Journalism Practice* 8(5).

Dahlig, E., and Schaffrath, H. 1993. Komputerowa symulacja melodii ludowych. *Eksperyment Muzyka* 38(3–4).

Dahlig, E., and Schaffrath, H. "1998". Judgments of human and machine authorship in real and artificial folksongs. *Computing in Musicology*.

Darda, K. M., and Cross, E. S. 2022. The role of expertise and culture in visual art appreciation. *Scientific Reports* 12(1):10666.

Davies, S. 2003. Ontology of art. In Levinson, J., ed., *The Oxford Handbook of Aesthetics*. Oxford University Press.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. of Exp. Psychology: General* 144(1).

Dubnov, S.; Burns, K.; and Kiyoki, Y. 2016. Cross-cultural aesthetics: analyses and experiments in verbal and visual arts. In *Cross-Cultural Multimedia Computing: Semantic and Aesthetic Modeling*. Springer. 21–41.

Déguernel, K.; Sturm, B. L. T.; and Maruri-Aguilar, H. 2022. Investigating the relationship between liking and belief in AI authorship in the context of Irish traditional music. In *Proc. of CREAI: Workshop on Artificial Intelligence and Creativity*.

Eigenfeldt, A.; Burnett, A.; and Pasquier, P. 2012. Evaluating musical metacreation in a live performance context. In *Proc. of ICCC*.

Fischinger, T.; Kaufmann, M.; and Scholtz, W. 2018. If it's mozart, it must be good? the influence of textual information and age on musical appreciation. *Psychology of Music* 48(4):579–597.

Flick, C., and Worrall, K. 2022. The ethics of creative AI. In Vear, C., and Poltronieri, F., eds., *The Language of Creative AI: Practices, Aesthetics and Structures*. Springer International Publishing. 73–91.

Flôres, R. G., and Ginsburgh, V. A. 1996. The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician* 45(1):97–104.

Friedman, R. S., and Taylor, C. L. 2014. Exploring emotional responses to computationally-created music. *Psychology of Aesthetics, Creativity and the Arts* 8(1):87–95.

Gangadharbatla, H. 2021. The role of AI attribution knowledge in the evaluation of artwork. *Empirical Studies of the Arts* 40(2).

Graefe, A.; Haim, M.; Haarmann, B.; and Brosius, H.-B. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* 19(5):595–610.

Greasley, A., and Lamont, A. 2016. Musical preferences. In Hallam, S.; Cross, I.; and Thaut, M., eds., *The Oxford Handbook of Music Psychology*. Oxford University Press.

Hagtvedt, H.; Patrick, V. M.; and Hagtvedt, R. 2008. The perception and evaluation of visual art. *Empirical studies of the arts* 26(2):197–218.

Herremans, D.; Chuan, C.-H.; and Chew, E. 2018. A functional taxonomy of music generation systems. *ACM Computing Surveys* 50(5).

Hitsuwari, J., and Nomura, M. 2022. How individual states and traits predict aesthetic appreciation of haiku poetry. *Empirical Studies of the Arts* 40(1):81–99.

Hitsuwari, J.; Ueda, Y.; Yun, W.; and Nomura, M. 2023. Does human–AI collaboration lead to more creative art? aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* 139(107502).

Holzapfel, A.; Jääskeläinen, P.; and Kaila, A.-K. 2022. Environmental and social sustainability of creative AI. In *Generative AI and CHI Workshop*.

Hong, J.-W., and Curran, N. M. 2019. Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15(2).

Hong, J.; Fischer, K.; Ha, Y.; and Zeng, Y. 2022. Human, I wrote a song for you: An experiment testing the influence of machines' attributes on the AI-composed music evaluation. *Computers in Human Behavior* 131(107239).

Hong, J.-W.; Peng, Q.; and Willians, D. 2021. Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society* 23(7):1920–1935.

Hong, J.-W. 2018. Bias in perception of art produced by artificial intelligence. In *Human-Computer Interaction. Interaction in Context*, 290–303.

Huang, R.; Sturm, B. L. T.; and Holzapfel, A. 2021. Decentering the West: East Asian philosophies and the ethics of applying artificial intelligence to music. In *Proc. of ISMIR*.

Hämäläinen, M., and Alnajjar, K. 2019. Let's FACE it. Finnish poetry generation with aesthetics and framing. In *Proc. of the 12th International Conference on Natural Language Generation*, 290–300.

Jago, A. S. 2019. Algorithms and authenticity. *Academy of Management Discoveries* 5(1).

Jordanous, A. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proc. of ICCC*.

Kirk, U.; Skov, M.; Hulme, O.; Christensen, M. S.; and Zeki, S. 2009. Modulation of aesthetic value by semantic context: An fMRI study. *NeuroImage* 44(3):1125–1132.

Koelsch, S.; Vuust, P.; and Friston, K. 2019. Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences* 23(1):63–77.

Kroger, C., and Margulis, E. H. 2016. "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music* 45(1):49–64.

Lahdelma, I., and Eerola, T. 2020. Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension. *Scientific Reports* 10:8693.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys* 51(2).

Lauring, J. O.; Pelowski, M.; Forster, M.; Gondan, M.; Ptito, M.; and Kupers, R. 2016. Well, if they like it... effects of social groups' ratings and price information on the appreciation of art. *Psychology of Aesthetics, Creativity, and the Arts* 10(3).

Leder, H., and Nadal, M. 2014. Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–developments and challenges in empirical aesthetics. *British journal of psychology* 105(4):443–464.

Lermann Henestrosa, A.; Greving, H.; and Kimmerle, J. 2023. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Computers in Human Behavior* 138(107445).

Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; et al. 2022. Competition-level code generation with alphacode. *Science* 378(6624):1092–1097.

Li, D. 2004. Trustworthiness of think-aloud protocols in the study of translation processes. *International Journal of Applied Linguistics* 14(3):301–313.

Lincoln, Y., and Guba, E. G. 1985. *Naturalistic inquiry*. Sage.

Liu, B., and Wei, L. 2019. Machine authorship in situ: effect of news organization and news genre on news credibility. *Digital Journalism* 7(5):635–657.

Logg, L. M.; Minson, J. A.; and Moore, D. A. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Longoni, C.; Fradkin, A.; Cian, L.; and Pennycook, G. 2022. News from generative artificial intelligence is believed less. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106.

Lubart, T. I. 2010. Cross-cultural perspectives on creativity. In *The Cambridge Handbook of Creativity*. Cambridge University Press. 265–278.

Maher, T. F. 1976. "need for resolution" ratings for harmonic musical intervals: A comparison between indians and canadians. *J. of Cross-Cultural Psychology* 7(3).

Mahmud, H.; Islam, A. N.; Ahmed, S. I.; and Smolander, K. 2022. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change* 175(121390).

McDermott, J. H.; Schultz, A. F.; Undurraga, E. A.; and Godoy, R. A. 2016. Indifference to dissonance in native amazonians reveals cultural variation in music perception. *Nature* 535(7613):547–550.

Moffat, D. C., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. *Assessment* 13(11).

Moura, F. T., and 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.

North, A.; Hargreaves, D. J.; and Hargreaves, J. J. 2004. Uses of music in everyday life. *Music perception* 22(1).

Norton, D.; Heath, D.; and Ventura, D. 2015. Accounting for bias in the evaluation of creative computational systems: an assessment of DARCI. In *Proc. of the 6th International Conference on Computational Creativity*, 31–38.

Obermeier, C.; Menninghaus, W.; Von Koppenfels, M.; Raettig, T.; Schmidt-Kassow, M.; Otterbein, S.; and Kotz, S. A. 2013. Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in psychology* 4.

Orr, M. G., and Ohlsson, S. 2005. Relationship between complexity and liking as a function of expertise. *Music perception* 22(4):583–611.

Pasquier, P.; Burnett, A.; Gonzalez Thomas, N.; Maxwell, J. B.; Eigenfeldt, A.; and Loughin, T. 2016. Investigating listener bias against musical creativity. In *Proc. of the 7th International Conference on Computational Creativity*.

Pearce, M. T. 2015. Effects of expertise on the cognitive and neural processes involved in musical appreciation. In Huston, J. P.; Nadal, M.; Mora, F.; Agnati, L. F.; and Cela-Conde, C. J., eds., *Art, aesthetics and the brain*. Oxford University Press. 319–338.

Pease, A., and Colton, S. 2011. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *Proc. of the AISB Symposium on AI and Philosophy.*, 15–22.

Ponterotto, J. G. 2006. Brief note on the origins, evolution, and meaning of the qualitative research concept 'thick description'. *The Qualitative Report* 11(3):538–549.

Ragot, M.; Martin, N.; and Cojean, S. 2020. AI-generated vs. human artworks. a perception bias towards artificial intelligence? In *CHI Conference on Human Factors in Computing Systems*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proc. of the 37th International Conference on Machine Learning*, 8821–8831.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models. *CoRR* abs/2112.10752.

Shank, D. B.; Stefanik, C.; Stuhlsatz, C.; Kacirek, K.; and Belfi, A. M. 2022. AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied* 28.

Steinbeis, N., and Koelsch, S. 2009. Understanding the intentions behind man-made products elicits neural activity in areas dedicated to mental state attribution. *Cerebral Cortex* 19(3):619–623.

Strapparava, C., and Stock, O. 2011. Computational humour. In Cowie, R.; Pelachaud, C.; and Petta, P., eds., *Emotion-Oriented Systems*. Springer. 609–634.

van der Kaa, H. A. J., and Krahmer, E. J. 2014. Journalist versus news consumer: The perceived credibility of machine written news. In *Proc. of the Computation+Journalism conference*.

Waddell, T. 2018. A robot wrote this?: how perceived machine authorship affects news credibility. *Digital Journalism* 6(2):236–255.

Winston, A. S., and Cupchik, G. C. 1992. The evaluation of high art and popular art by naive and experienced viewers. *Visual Arts Research* 18(1).

Wolz, S. H., and Carbon, C. C. 2014. What's wrong with an art fake? cognitive and emotional variables influenced by authenticity status of artworks. *Leonardo* 47(5):467–473.

Wu, Y.; Mou, Y.; Li, Z.; and Xu, K. 2020. Investigating american and chinese subjects' explicit and implicit perceptions of AI-generated artistic work. *Computers in Human Behavior* 104(106186).