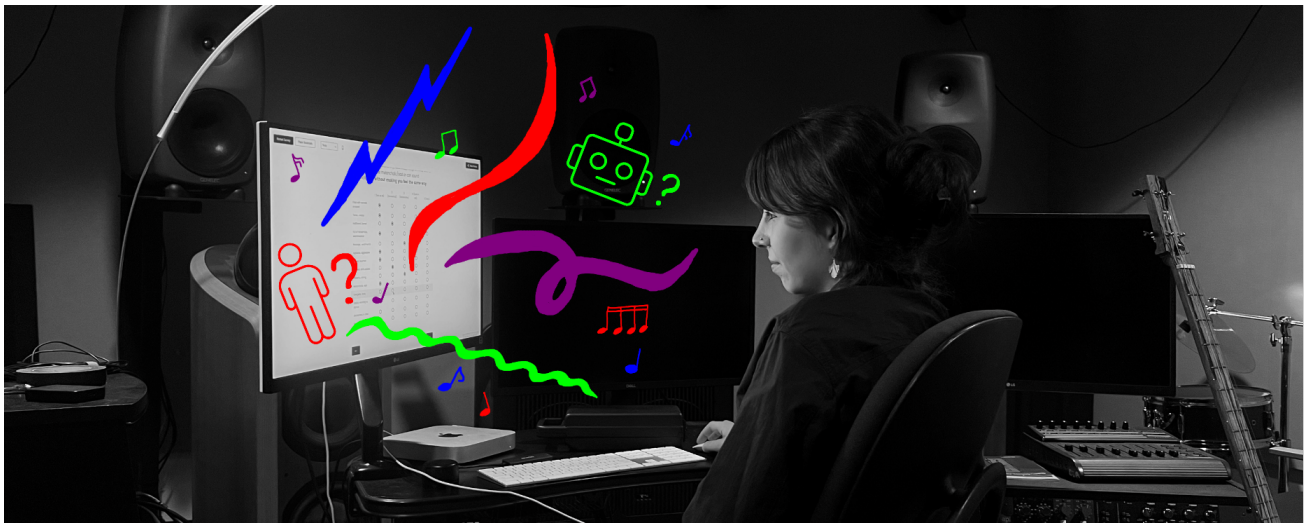
Understanding Listener Perceptions of AI and Human-Composed Music in Emotional Applications

Kimaya Lecamwasam
MIT Media Lab
Cambridge, USA
klecamwa@media.mit.edu

Tishya Ray Chaudhuri
Myndstream
London, UK



Abstract

Designing music-based affective technologies requires understanding of how perceptions of AI versus human authorship shape trust and authenticity. We investigate how listener perception of AI-generated versus human-composed music affects emotional resonance and regulation. Drawing on affective computing and human-computer interaction frameworks, participants listened to AI- and human-composed music across labeling conditions (Correct, Incorrect, or Unlabeled) and emotion cases (Calm and Upbeat). Participants rated preference, efficacy of target emotion elicitation, and emotional impact. Results showed participants found human-composed music more effective in eliciting their target affective states and linked humanness to imperfection, flow, and "soul," underscoring authenticity as central to appraisal and ultimately leading to design implications relevant to music-based HCI. These findings challenge the assumption that preference alone defines system success, highlighting design implications for affective and wellness technologies that foreground authenticity, transparency, and human creativity.

Keywords

Generative Music, Emotion Efficacy, Perception, Authenticity, Preference

1 Introduction

Generative AI is changing the shape of music creation. Despite its increasing integration, it is unclear whether the emotional impact of AI-generated music matches that of human-composed. To assess this, we conducted a study (N = 152) exploring how listeners classify and perceive AI and human music (Fig. 1). Participants listened to four one-minute-long, instrumental-only audio clips (two human and two AI) from two emotion cases: "Calming/Soothing" (Calm) and "Upbeat/Invigorating" (Upbeat). Participants were split into three groups: Unlabeled, Correctly labeled, and Incorrectly labeled (Figure 1-I). All participants completed demographics and validated pre/post-task questionnaires assessing state anxiety [46] (exploratory measure) and musical emotional resonance [10], indicated which songs they preferred and thought most effectively conveyed the target emotion, and provided reasoning. This study was preregistered prior to data collection on the Open Science Framework [25]. We hypothesized[1]:

H1 Without origin labels, participants will prefer human-composed music and find it more effective.
- **H1.a** For Upbeat music, unlabeled human-composed music will yield higher GEMIAC scores in high arousal/high valence emotion classes [29].
- **H1.b** For Calm music, unlabeled human-composed music will yield higher GEMIAC scores in low arousal/high valence emotion classes [29].

H2 Participants will prefer, and find more effective, songs labeled human, regardless of actual origin.
- **H2.a** For Upbeat music, songs labeled human, regardless of actual origin, will yield higher GEMIAC scores in high arousal/high valence emotion classes [29].
- **H2.b** For Calm music, songs labeled human, regardless of actual origin, will yield higher GEMIAC scores in low arousal/high valence emotion classes [29].

NIME '26, June 23–26, 2026, London, UK

[1] Hypotheses condensed for clarity.

H3 Interaction with musical stimuli will impact momentary anxiety measures, though this requires further study.

These hypotheses are consistent with previous work showing that musicians communicate discrete emotions through acoustic cue patterns (e.g. tempo, dynamics, and articulation) and that expectation management and expressive microtiming shape listeners' affective responses [15, 18, 23, 41]. By contrast, text-to-music systems have been limited by a lack of ability to "precisely control musical features so that the resulting music exhibits the desired affect" [11] due to interdependencies, dataset limitations, and lack of transparency [5, 6, 11]. Our findings suggest participants tend to find human music more effective in eliciting target emotions but skew towards preference for AI. Additionally, though GEMIAC *category* representation was not impacted by origin source, GEMIAC *scores* were significantly higher for human music in specific cases. Qualitative data also revealed persistent associations between human composition and emotional authenticity.

2 Related Work

2.1 Emotion-Sensitive AI and Ethics

Given music's well-established impact on human perception, cognition, and emotion [32, 39], there is growing interest in the use of *generative* music in similar contexts, including health interventions [35, 44], adaptive soundtracks [19, 20], and personalized recommendations [42, 43]. This interest reflects broader trends in AI research towards developing systems that better understand and respond to human emotions. However, ethical considerations remain, including debates over the universality of emotional expression and the implications culture-dependent and nature-nurture nuance hold for emotion-centered algorithms [12, 21, 40]. In emotion regulation contexts, clinical algorithms must (1) minimize bias, (2) promote user autonomy and confidentiality, and (3) be clinically effective [14, 47]. We seek to contribute to this larger conversation by investigating the musical features that are highlighted as "human" by human listeners, to assess how authenticity, emotional resonance, and cultural context shape human-computer interactions.

2.2 Ethics and Generative Music

Current perspectives on AI in music creation/production raise questions regarding copyright law [48], training bias [28], and practical application [2]. In a multi-subject case study of working musicians from a variety of genres and disciplines, researchers highlighted musicians' focus on the threat AI poses to employment and artistic integrity, in part due to concerns about a lack of both transparency and creative control within AI tools [31]. Recent advances in music generation have intensified concerns. SunoAI [49], Udio [54], and Lyria [51] have demonstrated unprecedented ability to generate music across genres, as evidenced by the 2025 Deezer and Ipsos survey ($N = 9000$) which revealed 97% of participants found AI-generated music perceptually indistinguishable from human-composed [56]. Results such as these raise urgent questions about the future of artistic integrity, creator attribution, and economic sustainability in music creation [36]. Furthermore, many existing systems rely heavily on Western tonal music, reducing cross-cultural applicability [1]. This parallels current emotion-sensitive systems' tendencies to oversimplify emotion states, mapping them to culturally limited dimensions, thereby failing to capture the complexity and nuance of human emotion [17, 28].

2.3 Existing Comparisons of AI-Generated and Human-Composed Music

Tigre Moura's and Maw's (2021) participants reported negative attitudes and low purchase intentions for AI music, though awareness of AI did not impact actual music evaluations, suggesting a disconnect between opinions and perceptual experiences [53]. Hong et al. (2022) found that, while anthropomorphizing generative AI systems did not affect music quality scores, accepting AI as a "musician" led to higher ratings, suggesting that role perception drives evaluations more than sonic attributes [16]. Zenieris (2023) demonstrated this labeling effect directly. Participants who knew whether songs were AI-generated or human-composed rated human music 80% higher, while uninformed participants preferred AI 66.7% of the time [58]. Sun et al. (2023) found that anthropomorphic features enhanced AI music's perceived competence and warmth [50], while Liu (2025) found that emotional attachment drove continued use of AI music platforms [27]. Chu et al. (2022) compared music generation models, linking melodiousness to highest satisfaction, though token representation methods and model characteristics also had significant impact [8]. Fernando et al. (2024)'s review of studies of emotional responses to AI music found that, while some systems show promise, ongoing listener skepticism calls for more research on emotional authenticity [13].

However, critical gaps remain, particularly related to the impact of AI on emotion regulation. While Hong et al. and Zenieris showed labeling that affects preference, neither examined whether this also impacts emotional efficacy [16, 58]. Additionally, while Tigre Moura and Maw and Zenieris examined responses when labels were experimenter-provided, few studies have systematically assessed how listeners' identification accuracy of unlabeled AI-versus-human music affects subsequent judgments [53, 58]. We aim to address these gaps by investigating: (1) systematically distinguishing preference (aesthetic appeal) from efficacy (perceived emotional/therapeutic effectiveness) using within-participant comparisons and validated emotional measures; (2) labeling accuracy in an Unlabeled condition where participants identify music origin; (3) emotion-specific effects for Calm versus Upbeat music; (4) directional preference-efficacy consistency to reveal whether these constructs align or diverge; and (5) how listener mislabeling rates vary and whether this predicts preference.

3 Methods

This study was conducted on Prolific via Qualtrics survey ($N = 152$) in September 2024. Participants listened to Calm and Upbeat music, chosen from the most requested use-cases identified by Myndstream, a wellness music company[2]. There were two one-minute-long songs within both cases, one generated by SunoAI and the other composed in-house by Myndstream using digital instruments (four total). SunoAI and the producer received the same two, 200-word prompts written by Myndstream's Head of Music based on briefs given to human composers to compose for the two emotion cases[3]:

- Calm: "Compose a relaxing instrumental piece to eliminate anxiety. Use soothing melodies, gentle harmonies, and calm tempo. Include soft piano, ambient synths and ethereal effects to evoke peace and relaxation."

[2]More information about Myndstream can be found at: myndstream.com
[3]Please see Supplementary Material A for descriptions of and links to all songs used in the protocol.

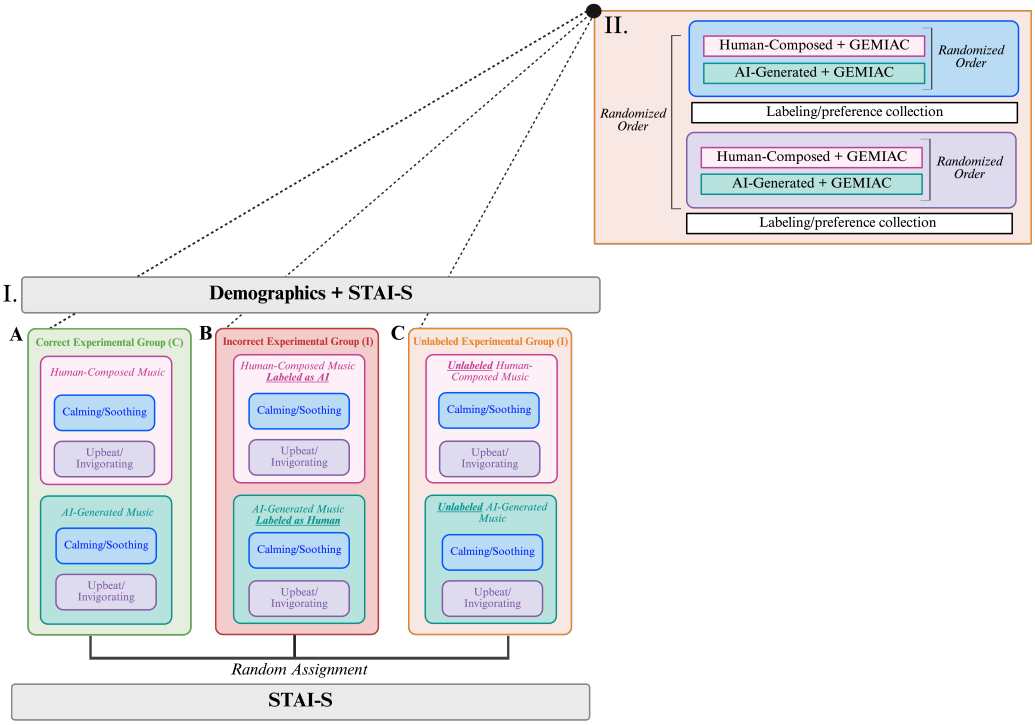


Figure 1: (*I*) – Participants completed demographics/STAI-S questionnaires, and were randomly assigned to one experimental group: (A) Correctly labeled AI-generated and human-composed music; (B) Incorrectly labeled music (AI labeled as human and vice versa); and (C) Unlabeled music, where participants labeled which songs they thought were human and which they thought were AI. They completed the STAI-S again at the end of the study. (*II*) – All listened to four songs: two Calm (one AI, one human) and two Upbeat (one AI, one human) in randomized order. Following each song, participants completed a labeling/preference task. Figure created in https://BioRender.com

- Upbeat: "Compose an invigorating instrumental piece to boost mood. Use uplifting melodies, dynamic rhythms, and a lively tempo. Include electronic drums, vibrant synths, and ethereal effects to evoke joy."

Both the producer and the researcher who generated the music did not have access to the other case to prevent bias. The AI-generated song was chosen after one round of generation without post-production.

Participants completed a demographic questionnaire (music training/preference, opinions on AI in music, experience with generative AI, etc.) and the State Anxiety Scale from the State-Trait Anxiety Inventory (STAI-S) [46], due to its sensitivity to short-term interventions, though STAI-S data were purely exploratory. Participants were randomly assigned to one experimental group: (1) Correctly labeled AI-generated and human-composed music; (2) Incorrectly labeled music, where AI-generated music was labeled as human-composed and vice versa; and (3) Unlabeled music, where participants labeled which music they thought was AI-generated and which they thought was human-composed (Fig. 1-I). At the end of the protocol, Incorrect participants were informed of the mislabeling and were asked to reflect. All participants received a disclaimer about potential mislabeling to meet ethical guidelines (MIT COUHES protocol E-6149).

Participants were randomly assigned to begin with either Calm or Upbeat music as part of a fully factorial design (Fig. 1-II). Within each emotion block, all participants listened to one song at a time (randomly assigned to either AI or human first) and, immediately after listening to each song, completed the Geneva Music-Induced Affect Checklist (GEMIAC) to assess the emotional resonance of each piece [10]. The GEMIAC survey builds upon the Geneva Music Emotion Scale (GEMS) by incorporating a broader range of genres and positive/negative emotions [10]. We used the intensity rating[4] due to the short duration and "relatively homogeneous emotional tonality" of the stimuli, following Coutinho's and Scherer's recommendation [10].

Participants also compared each song pair, reporting song which they preferred and which more effectively conveyed the target emotion ("efficacy") (Fig. 1-II). Unlabeled participants labeled which song they believed was human and which was AI. All participants reflected on the musical features that influenced their choices. Finally, participants completed the STAI-S again. This study was randomized and included repeated measures, using a mixed design to compare both between- and within-subject data. We embedded two attention checks into the survey.

3.1 Data Analysis

3.1.1 Quantitative Analysis. We analyzed preference and efficacy ratings using multinomial logistic regression (`nnet` R package [38]), since both outcomes were categorical with three unordered levels (AI-Generated, Human-Composed, Neither/Unclear). Models included experimental group, emotion case, and their interaction, with Type III (Wald) chi-square tests for overall effects and Bonferroni-corrected pairwise comparisons via the `emmeans` package [26]. We assessed within-participant preference-efficacy

[4]Please see Supplementary Material B for a mockup of the GEMIAC intensity scale.

agreement using generalized linear mixed-effects models (GLMMs) with binomial distributions (lme4 package [3]), including random intercepts for participants. We conducted directional consistency analyses using Fisher's Exact Tests to determine whether AI-preferrers also rated AI as more effective, and compared consistency rates between AI- and human-preferrers. We assessed Unlabeled group labeling accuracy by comparing participants' labels against actual music origin and used Fisher's Exact Tests to examine whether accuracy predicted preference/efficacy. Demographic effects were examined using GLMMs predicting (1) preference-efficacy agreement, (2) binomial AI preference, and (3) Unlabeled participants labeling accuracy. Age ranges were converted to numeric midpoints. All models included experimental group and emotion as covariates. GEMIAC scores were analyzed using linear mixed-effects models to assess how age, prior experience with AI-generated music, general music experience, and opinions on AI in the music industry impacted scores, with fixed effects for experimental group, emotion condition, music origin, GEMIAC category, and their interactions and random intercepts for participants to account for repeated measures across observations. All analyses used $\alpha = 0.05$ with Bonferroni correction for multiple comparisons, conducted in R version 4.5.1 [34].

3.1.2 Qualitative Analysis. Free responses were analyzed using Braun and Clarke's guidelines [9]. We generated initial codes inductively, ranging from surface-level observations (e.g., "mentions rhythm", "describes vocals") to latent meanings (e.g., "prefers human imperfection", "distrusts AI"). Themes were grouped by shared meaning: "organic" and "alive" were clustered under *Naturalness and Humanity*, while "emotionally moving" and "soothing" supported *Emotional Resonance*. We examined internal coherence and refined themes to ensure distinction. We note that free response data were analyzed thematically by a single rater. Given the absence of inter-rater reliability assessment, these findings should be considered exploratory and interpreted with appropriate caution. Future work should incorporate multiple raters and formal reliability metrics to validate identified patterns.

4 Results

4.1 Preference, Efficacy, and Labeling

Overall, participants rated human-composed music as more effective (58.9%) than AI (33.9%), while preferences were more balanced (51.2% AI vs 43.1% human across both emotions). Multinomial logistic regression revealed a significant difference between preference and efficacy ($\chi^2(2) = 19.03, p < 0.001$). Specifically, participants were 52% less likely to rate AI music as effective compared to their preference ratings ($OR = 0.48, 95\%CI[0.36, 0.65]$, $z = -4.25, p < 0.001$), indicating a substantial preference-efficacy disconnect (Hypotheses **H1**, **H2**).

4.1.1 Impact of Experimental Group. Experimental group significantly predicted music preference ($\chi^2(4) = 16.51, p = 0.002$). The Incorrect group (72% AI preference for Calm, 50% for Upbeat) showed significantly higher AI preference (though the music was labeled as human during the task) than the Correct group (41% AI for Calm, 37% for Upbeat; $OR = 3.00, 95\%CI[1.21, 7.43], z = 2.42, p\_adj = 0.047$) (Hypotheses **H2.a**, **H2.b**). The Unlabeled group showed similar patterns (69% AI for Calm, 39% for Upbeat; $OR = 2.33, z = 1.94, p_adj = 0.157$), though this did not reach significance after correction (Hypotheses **H1.a**, **H1.b**). No significant difference emerged between Incorrect and Unlabeled groups ($p\_adj = 1.00$). Experimental group did not significantly predict efficacy ratings ($\chi^2(4) = 5.29, p = 0.259$). All groups rated human music as more effective overall (Correct: 61% human, Incorrect: 53% human, Unlabeled: 63% human), with no significant pairwise differences (all $p\_adj > 0.72$) (Hypotheses **H1**, **H2**).

4.1.2 Impact of Emotion Condition. Emotion condition significantly influenced preference ($\chi^2(2) = 9.40, p = 0.009$). Participants showed higher AI preference for Calm music (60.5%) compared to Upbeat music (42.1%; $OR = 0.61, z = -1.15, p = 0.249$). Within-group analyses revealed significant emotion effects for the Correct ($\chi^2(2) = 8.42, p = 0.015$) and Unlabeled groups ($\chi^2(2) = 9.09, p = 0.011$), with marginal effects in the Incorrect group ($\chi^2(2) = 5.18, p = 0.075$). The Unlabeled group showed the most dramatic shift: 68.6% AI preference for Calm music dropped to 39.2% for Upbeat music (29.4% decrease) (Hypotheses **H1.b**, **H1.a**, **H2.b**, **H2.a**).

Emotion condition significantly influenced efficacy ratings ($\chi^2(2) = 6.40, p = 0.041$), with participants 58.7% less likely to rate AI as emotionally effective for Upbeat music ($OR = 0.41, z = -1.91, p = 0.056$). Within-group analyses revealed significant emotion effects for the Correct group ($\chi^2(2) = 6.29, p = 0.043$) and highly significant effects for the Unlabeled group ($\chi^2(2) = 18.75, p < 0.001$). AI efficacy ratings in the Unlabeled group dropped from 51% for Calm to 11.8% for Upbeat music (39.2% decrease) (Hypotheses **H1.b**, **H1.a**, **H2.b**, **H2.a**).

4.1.3 Preference-Efficacy Directional Consistency. Preference and efficacy agreement was approximately 71% overall, with no significant differences by experimental group ($\chi^2(2) = 0.44, p = 0.804$) or emotion ($\chi^2(1) = 0.43, p = 0.511$). However, critical asymmetries emerged in directional agreement. 89% of participants who preferred human music also rated it more effective, demonstrating strong internal consistency. However, only 54.8% of participants who preferred AI music rated it more effective (34.2% difference). This asymmetry was most pronounced in the Unlabeled group for Upbeat music, where only 25% of AI-preferrers also rated AI as more effective. 65% (13/20) of participants who preferred AI-generated Upbeat music rated human-composed music as more effective (OR = 0.027, p < 0.001). Significant consistency differences between AI and human preferrers emerged for Upbeat music across all groups (Correct: $p = 0.002$, Incorrect: $p = 0.041$, Unlabeled: p < 0.001).

4.1.4 Unlabeled Group Labeling Accuracy. Emotion condition was the strongest predictor of labeling accuracy ($\chi^2(1) = 28.37, p < 0.001$), with participants performing substantially better on Upbeat (56.9% correct) than Calm music (13.7% correct; $\beta = 2.92, p < 0.001$). For Calm music, 80.4% of participants mislabeled the human song as AI, while 76.5% mislabeled AI as human. Labeling accuracy significantly predicted preference for both Calm (Fisher's exact test, p = 0.003) and Upbeat music ($p = 0.009$): 79.1% of incorrect labelers preferred Calm AI music, compared to 14.3% of correct labelers (61.9% versus 24.1% for Upbeat music). Critically, 94.1% (32/34) of participants who preferred AI-generated Calm music and 63.2% (12/19) of those who preferred Upbeat AI mislabeled the tracks as human-composed, suggesting "AI preference" could reflect preference for music participants believed to be human-composed. Labeling accuracy showed a marginally significant relationship with efficacy ratings for Calm music ($p = 0.108$) and no relationship for Upbeat music ($p = 1.00$). Among participants who rated AI as more effective for Calm music, 92% (23/25) mislabeled the music as human. However, for Upbeat music, only

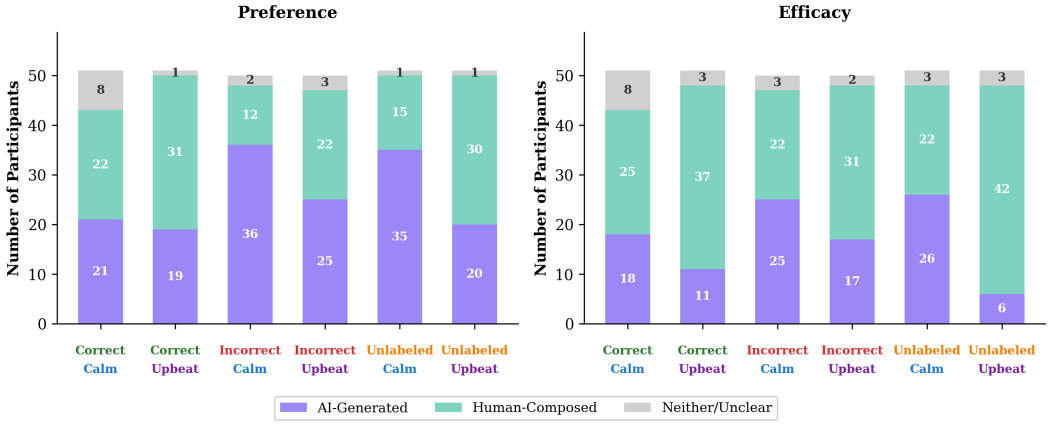


Figure 2: Stacked bar plots showing the number of participants in each experimental group (Correct, Incorrect, Unlabeled) who indicated (left) preference for and (right) perceived efficacy of AI-generated music, human-composed music, or neither/unclear, separated by emotion (Calm, Upbeat). Numbers indicate participant counts within each category.

50% (3/6) of participants who thought AI as more effective mislabeled the song.

4.1.5 Impact of Demographics. No demographic variables significantly predicted preference-efficacy agreement (all $p > 0.1$), indicating that the preference-efficacy disconnect occurs consistently across demographic groups. Age significantly predicted AI preference ($\chi^2(7) = 17.42, p = 0.015$), with participants aged 34-41 showing dramatically elevated AI preference (78.7%) compared to other age groups ($range : 34.6 - 66.7\%; \beta = 2.23, p = 0.014$). Emotion condition remained a strong predictor ($\chi^2(1) = 14.00, p < 0.001$), with significantly lower AI preference for Upbeat music across demographics. Counterintuitively, prior AI experience ($\chi^2(4) = 5.31, p = 0.257$) and opinions about AI in music ($\chi^2(5) = 4.19, p = 0.523$) did not predict AI preference. Notably, participants who strongly opposed AI in music still preferred AI-generated music 66.7% of the time, consistent with a mislabeling-driven preference pattern. Prior AI experience significantly predicted labeling accuracy ($\chi^2(3) = 16.94, p < 0.001$) but, interestingly, participants with average prior experience with AI were significantly worse at correctly identifying AI music ($\beta = -4.31, p = 0.018$). Age ($p = 0.201$) and musical experience ($p = 0.683$) did not improve labeling accuracy, suggesting that difficulty distinguishing AI from human music may transcend both technical familiarity and musical expertise.

4.2 STAI

Exploratory STAI-S results indicate that brief music exposure may impact momentary anxiety. Following the listening task, STAI-S scores decreased in 74 participants, increased in 50, and remained stable in 28, with clinically meaningful changes (≥ 8 points) observed in both directions (Hypothesis **H3**). However, these results remain inconclusive. While we cannot isolate specific effects due to the pre-post design, these findings underscore the need for further research into the roles of trust, perceived authenticity, and listener context in shaping health outcomes.[5]

[5]Please see Supplementary Material D for a more in depth discussion of STAI-S results

4.3 GEMIAC

Estimated marginal means were calculated to examine differences in participants' emotional responses. In the Calm condition, participants reported significantly higher ratings for low arousal/high valence emotions such as "Relaxed, peaceful" ($\beta = 2.309, p < 0.001, M = 3.32, SE = 0.068$), "Full of tenderness, warmhearted" ($\beta = 1.510, p < 0.001, M = 2.70, SE = 0.068$), and "Moved, touched" ($\beta = 1.333, p < 0.001, M = 2.48, SE = 0.068$) (Hypotheses **H1.a**, **H2.a**). Conversely, the lowest-rated emotions in this condition were "Agitated, aggressive" ($M = 1.13, SE = 0.068$) and "Tense, uneasy" ($M = 1.17, SE = 0.068$), suggesting that calming music effectively minimizes arousing negative emotions. In contrast, in the Upbeat case, participants reported significantly higher ratings for high arousal/high valence emotions, such as "Energetic, lively" ($\beta = 1.118, p < 0.001, M = 2.98, SE = 0.068$), "Joyful, wanting to dance" ($\beta = 0.804, p = 0.001, M = 2.66, SE = 0.068$), and "Inspired, enthusiastic" ($\beta = 0.373, p = 0.142, M = 2.51, SE = 0.068$) (Hypotheses **H1.b**, **H2.b**). Notably, the lowest-rated emotions were "Melancholic, sad" ($M = 1.18, SE = 0.068$) and "Agitated, aggressive" ($M = 1.25, SE = 0.068$). Pairwise comparisons between GEMIAC categories revealed several significant contrasts. In the Calm condition, "Relaxed, peaceful" was rated significantly higher than nearly all other emotions, particularly "Tense, uneasy" ($z = 29.29, p < 0.0001$). In the Upbeat condition, "Energetic, lively" was significantly higher than "Melancholic, sad" ($z = 24.49, p < 0.0001$).

These findings suggest that music categorized as Calm reliably elicited feelings of tranquility, tenderness, and awe, while Upbeat music was associated with increased energy, enthusiasm, and movement. Overall, it appears that the specific GEMIAC *categories* identified were not influenced by labeled origin, since human and AI music elicited similar responses. However, GEMIAC *scores* showed some sensitivity to music origin. In the Upbeat condition, the highest-rated GEMIAC category across conditions for both AI and human music was "Energetic, lively." In the Unlabeled ($p = 0.015$) and Correct ($p = 0.0014$) groups, scores were significantly higher for human music than AI (Hypotheses **H1.b**, **H2.b**). By contrast, no significant differences between scores for AI and human music were observed for "Relaxed, peaceful" in the Calm condition (Hypotheses **H1.a**, **H2.a**) or within the Incorrect group (Hypotheses **H2.a**, **H2.b**). These results suggest listeners

may find human-composed music more effective in eliciting emotional responses than AI-generated music in some circumstances, though further assessment is needed.

4.3.1 Impact of Demographics. Participants who reported excellent prior experience with generative music showed significantly higher overall GEMIAC scores ($\beta = 1.58, SE = 0.56, p = 0.005$), while other levels of experience (e.g., average, good, poor) were not significantly different from the reference, which could serve as an indication of confirmation bias due to prior positive exposure [24], though further investigation is needed. This effect was moderated by age, though age on its own did not yield significant effect. For example, those aged 34–41 with "good" prior AI music experience reported significantly higher ratings ($\beta = 1.963, SE = 0.895, p = 0.031$) than those aged 18–25 with no prior AI music experience. Additionally, participants who actively sought wellness music reported higher emotional responses ($\beta = 0.55, SE = 0.24, p = 0.026$) relative to those who did not. No significant main effects were found for general music experience or opinions on the role of AI in music. Post-hoc comparisons using the `emmeans` package confirmed robust differences in emotion intensity ratings between GEMIAC categories across both emotion case and participant-level covariates, suggesting that these emotional responses are consistent regardless of prior musical experience or AI attitudes. Overall, the results indicate that both participant backgrounds and specific emotional targets shape emotional impact. Random effects indicated modest variability across participants ($\sigma^2 = 0.24$), suggesting that individual differences contributed to emotional response patterns beyond fixed demographic predictors.

4.4 Free Response Data

4.4.1 Thematic Trends by Emotion Case. In the Calm case, participants frequently described music with references to "peace," "gentle[ness]," and "slower pace." Correct participants often articulated clear emotional connections, describing music that "takes me away to a more peaceful place" or "gave an overall calm feeling." In contrast, some Incorrect participants showed signs of cognitive dissonance, expressing confusion or second-guessing, though most indicated their preference/efficacy determinations did not change post-debrief, since they were shaped by personal taste, not origin labels. Unlabeled participants often used generic descriptors like "relaxing" or "not really different," potentially reflecting less confident appraisals.

In the Upbeat case, participants focused on physical/energetic reactions, referencing tempo, rhythm, and movement ("danceable," "energized," and "joyful"). Correct participants shared, "It just made me feel better than the other song," and, "[the human music] feels like an actual beat a person made." The Incorrect group often questioned whether high energy alone signaled human composition ("The [AI] song felt like it was trying too hard to make me feel something.") Unlabeled participants frequently cited technical details like beat and melody ("I couldn't tell a difference, but the rhythm stood out more.")

4.4.2 Thematic Frequencies and Cues. From approximately 1,700 analyzed responses, we constructed nine main themes:

- Naturalness and Humanity (252 responses)
- Perceptual Ambiguity or Uncertainty (226)
- Emotional Resonance (225)
- Mechanical/Synthetic Quality (177)
- Technical Features (176)
- Listener Agency or Subjectivity (121)
- Genre or Context Association (54)[6]
- Creativity, Novelty, or Usefulness(51)[6]
- Indifference or Detachment (50)[6]

Naturalness and Humanity appeared most frequently in the Correct and Incorrect groups, relating "flow", "emotion", "realness", "naturalness", "organic quality", "soul", and "feel[ing] human" to melody, beat, and rhythm. Interestingly, only one participant, in the Correct group, specifically mentioned the impact of auditory artifacts, stating that "the presence of more imperfections or audial artifacts" would make human-composed Upbeat music appear more AI. Interestingly, in the Incorrect and Unlabeled groups, some assigned human characteristics to AI music, with one noting:

> "...the human-composed piece [actually AI] blows the AI-generated piece [actually human] totally away, hands down. The AI piece [actually human] doesn't have nearly as much emotion..."

Mechanical/Synthetic Quality appeared frequently in the Incorrect group, possibly reflecting label-driven confusion. Phrases like "too perfect", "robotic", and "artificial" were common. One participant stated, "it sounded like a fake performance by a human." ***Perceptual Ambiguity or Uncertainty*** was most common in the Incorrect group. Participants often admitted confusion ("not sure", "couldn't tell", "hard to say"). One explained, "I would not have known it was AI unless you told me, it felt like a guess." Another noted, "I had to just choose based on a hunch." In some cases, participants felt the pieces lacked distinguishing features. ***Emotional Resonance*** appeared across groups via affective language such as "calming," "uplifting," "enchanting," or "joyful." While some noted, "[they] just liked how it made [them] feel", others included technical descriptors, indicating emotional interpretation coexists with analytical reasoning. ***Technical Features*** were particularly mentioned in the Unlabeled group, with focus on "beat" ("The beat felt natural..."), melody ("The melody...had a nicer flow."), rhythm ('The rhythm helped build the emotion."), tone ("The tone was softer...more organic."), and tempo ("It had a slower tempo that helped me relax."). Participants mentioned that they were, "...not sure about the composer, but the instrumental was well layered." However, participants generally did not provide detailed insight into *how* the specific elements they mentioned impacted preference or efficacy, echoing calls in prior work [57] for further investigation into the impact of musical vocabulary on identification tasks. ***Listener Agency or Subjectivity*** emphasized personal interpretation and taste ("for me", "I think", "depends on the person"), indicating awareness of subjectivity in musical evaluation. Participants situated their responses within personal frameworks rather than making universal claims ("It just felt better to me. I know others might disagree.")

These findings raise broader questions about the evaluative frameworks used for determining preference and emotional efficacy. If listeners gravitate to perceived humanness or authenticity, this suggests social/relational dimensions remain central to emotion and value. Judgments guided by technical or aesthetic qualities indicate more evaluative or formal approaches. Understanding which features listeners prioritize helps explain variance in responses to AI versus human music. These results have implications for the design and training of future models

[6]For analyses of themes with under 60 responses, please see Supplementary Material C.

and interfaces, which should account for the importance of human imperfection, emotional expressivity, and cultural context for meaningful emotional resonance. This insight is also critical for efforts to protect human musical creativity in an era where automated systems are increasingly capable of producing emotionally persuasive outputs.

4.4.3 Incorrect Group Reflections on Mislabeling. Post-debrief, Incorrect participants reflected on whether their preference/efficacy designations changed. A chi-square goodness-of-fit test (Yes ($n = 17$), No ($n = 29$), and Unclear/Mixed ($n = 4$)) indicated participants were more likely to maintain their initial judgments ($\chi^2 = 18.76, p < 0.001$) despite the mislabeling, although many reflected on their assumptions or acknowledged surprising capabilities of AI. Additionally, many who reported a perspective change still emphasized the primacy of personal taste. Responses also included resistance to machine-authored art or a sense of validation regarding intuition. Individuals who felt Unclear/Mixed, however, typically showed partial reconsideration, such as surprise at songs' origin sources or acknowledgment of the quality of AI-generated music, without fully changing perspective.

5 Discussion

Using the GEMIAC, the STAI-S, and comparison tasks, we measured aesthetic emotional responses and momentary emotion regulation in reaction to AI-generated and human-composed music across Correctly labeled, Incorrectly labeled, and Unlabeled conditions for both Calm and Upbeat music. Our findings revealed a systematic preference-efficacy disconnect: participants tended to find human music more effective in eliciting target emotions but preferred AI music. This disconnect, driven in part by inability to accurately identify music origin, has critical implications for AI music system design, the role of perceived authenticity in emotional regulation, and the continued importance of preserving human musical creativity in an era of increasing automation. Though GEMIAC *category* representation did not differ significantly by origin, GEMIAC *scores* were significantly higher for human-composed music in specific cases, and qualitative data revealed persistent associations between perceived "humanness" and authenticity.

5.1 Perceptions of Humanness and the Role of Labeling

Exploratory thematic analysis emphasized qualities like "flow", "realness", "organic[ness]", "soul", and "imperfection" as indicators of humanity, though these patterns require validation through multi-rater coding. These descriptors echo prior work studying music performance that highlights the importance of micro-expressive timing, interpretive variability, and dynamic shaping to human musicality [7, 33, 37, 45]. However, participants attributed human characteristics to AI-generated music when they believed it was human-composed. This pattern reveals a critical finding: perceived origin, rather than actual music origin, appeared to influence perception. In the Unlabeled group, labeling accuracy was poor for Calm music (13.7% correct) but substantially better for Upbeat music (56.9% correct). For Calm music, 80.4% mislabeled the human song as AI and 76.5% mislabeled AI as human, effectively reversing attributions. Most critically, 94.1% of Unlabeled participants who preferred AI-generated Calm and 63.2% who preferred AI in the Upbeat cases mislabeled the music as human-composed. Participant labeling accuracy significantly predicted preference for both Calm ($p = 0.003$) and Upbeat music ($p = 0.009$). Among participants who mislabeled the music, 79.1% preferred AI for Calm compared to only 14.3% among those who correctly labeled. These findings suggest observed "AI preference" reflected preference for music believed to be human-composed. Listeners also seemingly interpreted irregularities/flaws as hallmarks of authenticity. One participant remarked, "...it felt like a real person was playing. There were tiny flaws that made it feel alive," suggesting listeners intuitively link imperfection with expressiveness. These findings underscore the importance of preserving traces of human imperfection while also affirming the ongoing cultural and emotional significance of human-made music.

5.2 Preference, Efficacy, and Labeling Effects

Three interrelated patterns emerged: (1) widespread mislabeling of AI music as human-composed, particularly for Calm music; (2) preference ratings following participants' *beliefs* about music origin; and (3) efficacy ratings more accurately reflecting *actual* music origin, creating a systematic preference-efficacy disconnect. Overall, participants rated human-composed music as more effective (58.9%) than AI-generated music (33.9%), while preferences were more balanced (51.2% AI vs. 43.1% human). Multinomial logistic regression revealed participants were 52% less likely to rate AI music as effective compared to their preference ratings ($OR = 0.48, p < 0.001$). Critically, this disconnect was asymmetric. Among participants who preferred human music, 89.0% also rated it as more effective, demonstrating strong internal consistency. In contrast, only 54.8% of AI-preferrers also rated it as more efficacious (a 34.2% difference). This was most pronounced in the U group for Upbeat music, where only 25% of AI-preferrers also rated AI as more effective. Remarkably, 65% of participants who preferred AI-generated Upbeat music rated human-composed music as more efficacious ($p < 0.001$).

Experimental group significantly predicted preference ($\chi^2(4) = 16.51, p = 0.002$) but not efficacy ($\chi^2(4) = 5.29, p = 0.259$). The Incorrect group showed significantly higher AI preference (which they thought was human music) than the Correct group ($p = 0.047$). All groups consistently rated human music as more effective, regardless of labeling condition and across demographic groups, suggesting that perceived origin drove the preference-efficacy asymmetry rather than musical qualities, age, AI experience, or musical expertise. Interestingly, participants with average prior experience with AI performed significantly worse at identifying AI music ($\beta = -4.31, p = 0.018$), and musical experience did not improve labeling accuracy ($p = 0.683$). Even participants who strongly opposed AI in music preferred AI-generated music 66.7% of the time when they believed it was human-composed.

We identified a complex interplay between musical preference and emotional efficacy, since participants did not always prefer the music they found most emotionally effective. These results echo findings from Thompson (2006) where, during live performances of classical music, audiences were able to decouple perceived musical quality (which could be equated to the preference metric in this work) and emotion elicitation, ultimately suggesting that enjoyment was better predicted by emotional engagement than perceived quality [22, 52]. These findings challenge the use of preference as a primary metric in relaying user feedback. If participants prefer music that is not optimally emotionally effective, preference alone may not capture functional success for therapeutics or mood-modulation. This disconnect

suggests aesthetic appreciation operates separately from perceived emotional effectiveness, with implications for how generative systems are evaluated, though further study is necessary.

5.3 Design Implications

Our results offer several actionable design insights for novel, emotionally intelligent musical interfaces that incorporate AI. Most notably, we find that listener preference does not always align with emotional efficacy. This challenges assumptions that preference sufficiently proxies affective success. Practically, these findings call for a deeper consideration of principles from music perception and psychology when designing novel musical interfaces. GEMIAC results demonstrate the need to calibrate system goals to the arousal level of the target emotion. In the Calm condition, AI and human music were comparably effective in evoking low arousal/high valence emotions. However, in the Upbeat condition, human music significantly outperformed AI on measures of "energetic, lively" in both Unlabeled ($p = 0.015$) and Correct ($p = 0.0014$) groups. This asymmetry suggests generative models may be more effective for low-arousal use cases (meditation, background music, wellness applications) while high-arousal contexts (motivation, celebration, live performance) necessitate human oversight or collaboration. Designers should carefully consider where AI-generated music is appropriate and ethically aligned, ensuring these tools augment rather than displace musicians.

Exploratory analyses of free response data suggest "flow", "realness", "imperfection", and "soul" signal human intentionality and creativity. Interestingly, participants used similar terms to described AI-generated music misidentified as human, indicating sonic cues can simulate human musicality. These findings highlight the enduring value of human expression and performance subtleties and caution against attempts to wholly replace/replicate them or strive solely for technical perfection.

Even if Incorrect group participants mentioned reevaluating the music after discovering true authorship, the majority did not change their preference/efficacy designations. This suggests that, while framing influences appraisal, emotional impressions are not entirely overwritten by authorship knowledge. Design strategies promoting transparency may balance openness with interpretive flexibility. Labeling should educate, helping listeners distinguish artistic collaboration from replacement, fostering support for human musicians. One such example is the *Jordan and the jam_bot* performance (2024), where Naseck et al. constructed a kinetic sculpture to help audiences understand whether musical output was human or AI-generated [4, 30]. Many participants also emphasized the need for interpretive agency and subjectivity. Designers might support open-ended interpretation through interfaces that allow users to describe music's affective impact without requiring definitive labels, preserving space for personal meaning-making while pulling design principles from complimentary HCI work, including journaling tools [55] .

5.4 Limitations and Future Directions

We note that this study is limited by stimulus variety and sample size. Free response data were coded by a single rater without inter-rater reliability assessment, limiting generalizability. While this work offers insights into listener interpretations of ambient and mood-based music, further research should use an expanded set of stimuli, exploring genre-specific reactions and cultural variability, given how participants with prior preference for wellness music reported higher GEMIAC emotional responses ($\beta = 0.55, SE = 0.24, p = .026$). Beyond this, even within our target genres, there are countless potential prompts and songs that could elicit completely different sets of labels, preferences, and efficacies. We also note that it is difficult to perfectly match stimuli on all possible covariates. However, we believe that this work serves as a meaningful first step.

The lack of concrete discussion of auditory artifacts warrants future studies to assess the impact of musical and AI/ML expertise on emotional resonance and reflection. The preference-efficacy dichotomy deserves continued exploration through fine-grained musicological analysis and alternative assessment methods, such as confidence ratings and biometric assessment. Additionally, although STAI-S data were exploratory, they provide foundation for future work examining how human- versus AI-generated music affects anxiety and emotional regulation, potentially incorporating clinical populations, real-time physiological markers, or longitudinal exposure to evaluate therapeutic impact.

5.5 Conclusion

Our results call for a shift in how emotional generative music systems are evaluated and designed, challenging common metrics for evaluation and highlighting the need to distinguish aesthetic appreciation from functional outcomes. Instead of privileging surface-level preference or formal coherence, future systems should prioritize emotional credibility, interpretive openness, and relational trust. Designing to support, not supplant, human artistry is vital for affective outcomes and honoring musical experiences' social dimensions. Protecting space for human musicianship, emotional nuance, and creative labor must remain a central ethical commitment as generative music evolves.

6 Ethical Standards

This study met the ethical standards set by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (MIT COUHES) protocol E-6149. This work was supported by discretionary funding from Opera of the Future at the MIT Media Lab. Author TRC was employed by Myndstream during the study period. To mitigate potential conflicts of interest, all data collection and analyses were conducted independently by KL in accordance with IRB-approved protocols.

Acknowledgments

Many thanks to Prof. Tod Machover and the Opera of the Future group; Freddie Moross, Adrienne O'Brien, Jordan Galvan, and the Myndstream team; and Dr. Anna Huang and the Human-AI Resonance Lab for their guidance and support. Thanks also to Dr. Jin Ha Lee, Dr. William Brannon, Dr. Nikhil Singh, Stephen Brade, Jessie Mindel and Lancelot Blanchard for their feedback.

References

[1] George Athanasopoulos, Tuomas Eerola, Imre Lahdelma, and Maximos Kaliakatsos-Papakostas. 2021. Harmonic organisation conveys both universal and culture-specific cues for emotional expression in music. *PLoS One* 16, 1 (2021), e0244964.

[2] Julia Barnett. 2023. The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 146–161.

[3] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence* 12, 1 (2015), 2.

[4] Lancelot Blanchard, Perry Naseck, Stephen Brade, Kimaya Lecamwasam, Jordan Rudess, Cheng-Zhi Anna Huang, and Joseph Paradiso. 2025. The jam_bot, a real-time system for collaborative free improvisation with music language models. In *Ismir 2025 Hybrid Conference*.

[5] Jean-Pierre Briot and François Pachet. 2020. Deep learning for music generation: challenges and directions. *Neural Computing and Applications* 32, 4 (2020), 981–993.

[6] Filippo Carnovalini and Antonio Rodà. 2020. Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence* 3 (2020), 14.

[7] Elaine Chew and Andrew McPherson. 2017. Performing music: Humans, computers, and electronics. In *The Routledge Companion to Music Cognition*. Routledge, 301–312.

[8] Hyeshin Chu, Joohee Kim, Seongouk Kim, Hongkyu Lim, Hyunwook Lee, Seungmin Jin, Jongeun Lee, Taehwan Kim, and Sungahn Ko. 2022. An empirical study on how people perceive AI-generated music. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 304–314.

[9] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.

[10] Eduardo Coutinho and Klaus R Scherer. 2017. Introducing the GEneva Music-Induced Affect Checklist (GEMIAC) a brief instrument for the rapid assessment of musically induced emotions. *Music Perception: An Interdisciplinary Journal* 34, 4 (2017), 371–386.

[11] Adyasha Dash and Kathleen Agres. 2024. Ai-based affective music generation systems: A review of methods and challenges. *Comput. Surveys* 56, 11 (2024), 1–34.

[12] Paul Ekman and Dacher Keltner. 1970. Universal facial expressions of emotion. *California mental health research digest* 8, 4 (1970), 151–158.

[13] Poorna Fernando, Thilini V Mahanama, and Manya Wickramasinghe. 2024. Assessment of human emotional responses to ai–composed music: a systematic literature review. In *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Vol. 7. IEEE, 1–6.

[14] Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research* 21, 5 (2019), e13216.

[15] Alf Gabrielsson and Patrik N Juslin. 1996. Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of music* 24, 1 (1996), 68–91.

[16] Joo-Wha Hong, Katrin Fischer, Yul Ha, and Yilei Zeng. 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 (2022), 107239.

[17] Rujing Huang, Andre Holzapfel, Bob Sturm, and Anna-Kaisa Kaila. 2023. Beyond diverse datasets: Responsible MIR, interdisciplinarity, and the fractured worlds of music. *Transactions of the International Society for Music Information Retrieval* 6, 1 (2023), 43–59.

[18] David Huron. 2008. *Sweet anticipation: Music and the psychology of expectation*. MIT press.

[19] Patrick Edward Hutchings and Jon McCormack. 2019. Adaptive music composition for games. *IEEE Transactions on Games* 12, 3 (2019), 270–280.

[20] Manuel López Ibáñez, Maximiliano Miranda, Nahum Alvarez, and Federico Peinado. 2021. Using gestural emotions recognised through a neural network as input for an adaptive music system in virtual reality. *Entertainment Computing* 38 (2021), 100404.

[21] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 19 (2012), 7241–7244.

[22] Patrik N Juslin. 2013. From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of life reviews* 10, 3 (2013), 235–266.

[23] Patrik N Juslin and Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin* 129, 5 (2003), 770.

[24] Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation* 32 (1995), 385–418.

[25] Kimaya Lecamwasam and Freddie Moross. 2024. Assessing the Emotional Resonance of Human-Composed Music As Compared to AI-Generated Music. https://doi.org/10.17605/OSF.IO/62T9E

[26] Russell V. Lenth and Julia Piaskowski. 2025. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://rvlenth.github.io/emmeans/ R package version 2.0.1.

[27] Zhixin Liu. 2025. The Impact of User Experience on Continuous Usage Intention for AI-Generated Digital Music Platforms: Examining the Mediating Roles of Emotional Attachment and Technology Self-Efficacy. *Sage Open* 15, 4 (2025), 21582440251406713.

[28] Atharva Mehta, Shivam Chauhan, and Monojit Choudhury. 2024. Missing Melodies: AI Music Generation and its "Nearly" Complete Omission of the Global South. *arXiv preprint arXiv:2412.04100* (2024).

[29] Bruno Mesz, Sebastián Tedesco, Felipe Reinoso-Carvalho, Enrique Ter Horst, German Molina, Laura H Gunn, and Mats B Küssner. 2023. Marble melancholy: using crossmodal correspondences of shapes, materials, and music to predict music-induced emotions. *Frontiers in psychology* 14 (2023), 1168258.

[30] Perry Naseck, Lancelot Blanchard, Madhav Lavakare, Kimaya Lecamwasam, and Joseph A Paradiso. 2025. Physical Manifestation of Generative AI Music Systems for Live Performance. In *Proceedings of the SIGGRAPH Asia 2025 Art Papers*. 1–6.

[31] Michele Newman, Lidia Morris, and Jin Ha Lee. 2023. Human-AI Music Creation: Understanding the Perceptions and Experiences of Music Creators for Ethical and Productive Collaboration.. In *ISMIR*. 80–88.

[32] Adrian C North, David J Hargreaves, and Jon J Hargreaves. 2004. Uses of music in everyday life. *Music perception* 22, 1 (2004), 41–77.

[33] Caroline Palmer. 1997. Music performance. *Annual review of psychology* 48, 1 (1997), 115–138.

[34] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[35] Jessica Sharmin Rahman, Tom Gedeon, Sabrina Caldwell, Richard Jones, and Zi Jin. 2021. Towards effective music therapy for mental health care using machine learning tools: human affective reasoning and music genres. *Journal of Artificial Intelligence and Soft Computing Research* 11, 1 (2021), 5–20.

[36] Md Awsafur Rahman, Zaber Ibn Abdul Hakim, Najibul Haque Sarker, Bishmoy Paul, and Shaikh Anowarul Fattah. 2024. SONICS: Synthetic Or Not–Identifying Counterfeit Songs. *arXiv preprint arXiv:2408.14080* (2024).

[37] Bruno H Repp. 1998. Variations on a theme by Chopin: Relations between perception and production of timing in music. *Journal of Experimental Psychology: Human Perception and Performance* 24, 3 (1998), 791.

[38] Brian Ripley, William Venables, and Maintainer Brian Ripley. 2016. Package 'nnet'. *R package version* 7, 3-12 (2016), 700.

[39] Teppo Särkämö, Mari Tervaniemi, and Minna Huotilainen. 2013. Music perception and cognition: development, neural basis, and rehabilitative use of music. *Wiley Interdisciplinary Reviews: Cognitive Science* 4, 4 (2013), 441–451.

[40] Klaus R Scherer, Elizabeth Clark-Polner, and Marcello Mortillaro. 2011. In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology* 46, 6 (2011), 401–435.

[41] Olivier Senn, Claudia Bullerjahn, Lorenz Kilchenmann, and Richard von Georgi. 2017. Rhythmic density affects listeners' emotional response to microtiming. *Frontiers in Psychology* 8 (2017), 1709.

[42] Anjali Sharma, Shubham Vishwakarma, and Liya T Mathew. 2024. Feel good ai: Voice-enabled emotion-based music recommendation system. In *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 1–6.

[43] Vijay Prakash Sharma, Azeem Saleem Gaded, Deevesh Chaudhary, Sunil Kumar, and Shikha Sharma. 2021. Emotion-based music recommendation system. In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 1–5.

[44] Lin Shen, Haojie Zhang, Cuiping Zhu, Ruobing Li, Kun Qian, Wei Meng, Fuze Tian, Bin Hu, Björn W Schuller, and Yoshiharu Yamamoto. 2024. A First Look at Generative Artificial Intelligence Based Music Therapy for Mental Disorders. *IEEE Transactions on Consumer Electronics* (2024).

[45] Olga Solomonova, Olha Ohanezova-Hryhorenko, Viola Demydova, Yurii Kuchurivskyi, and Volodymyr Bordonyuk. 2023. Interpretive Content of a Musical Work: The Performing Aspect. *Convergences-Journal of Research and Arts Education* 16, 32 (2023), 125–138.

[46] Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz FS Natalicio, and Diana S Natalicio. 1971. The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology* 5, 3 & 4 (1971).

[47] Isabel Straw. 2021. Ethical implications of emotion mining in medicine. *Health Policy and Technology* 10, 1 (2021), 191–195.

[48] Bob LT Sturm, Maria Iglesias, Oded Ben-Tal, Marius Miron, and Emilia Gómez. 2019. Artificial intelligence and music: open questions of copyright law and engineering praxis. In *Arts*, Vol. 8. MDPI, 115.

[49] PM Suhailudheen and Ms Sheena Km. 2025. Suno AI: Advancing AI-Generated Music with Deep Learning. *Authorea Preprints* (2025).

[50] Daoyin Sun, Haodong Wang, and Jie Xiong. 2024. Would you like to listen to my music, my friend? An experiment on AI musicians. *International Journal of Human–Computer Interaction* 40, 12 (2024), 3133–3143.

[51] Lyria Team, Antoine Caillon, Brian McWilliams, Cassie Tarakajian, Ian Simon, Ilaria Manco, Jesse Engel, Noah Constant, Yunpeng Li, Timo I Denk, et al. 2025. Live music models. *arXiv preprint arXiv:2508.04651* (2025).

[52] Sam Thompson. 2006. Audience responses to a live orchestral concert. *Musicae Scientiae* 10, 2 (2006), 215–244.

[53] Francisco Tigre Moura and Charlotte Maw. 2021. Artificial intelligence became Beethoven: how do listeners and music professionals perceive artificially composed music? *Journal of Consumer Marketing* 38, 2 (2021), 137–146.

[54] Udio. 2024. Udio: AI Music Generation. https://www.udio.com. AI music generation platform. Accessed: 2026-02-10.

[55] Kevin Wang, Samantha Barg, and Ramon Lawrence. 2025. Designing for an AI-Augmented Journaling Experience: Balancing Guidance and Autonomy for Deeper Emotional Insight. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[56] Jesper Wendel. 2025. Deezer and Ipsos study: AI fools 97% of music fans. https://newsroom-deezer.com/2025/11/deezer-ipsos-survey-ai-music/. Accessed: 2026-02-10.

[57] Simin Yang, Courtney N Reed, Elaine Chew, and Mathieu Barthet. 2021. Examining emotion perception agreement in live music performance. *IEEE transactions on affective computing* 14, 2 (2021), 1442–1460.

[58] Rafael Zenieris. 2023. *Perception and Bias towards AI-Music*. B.S. thesis. University of Twente.