Title

Artificial Intelligence Versus Human Focus Group Rating of Facial Attractiveness

Permalink

https://escholarship.org/uc/item/9573t2xh

Authors

Goshtasbi, Khodayar

Hakimi, Amir A

Wong, Brian JF

Publication Date

2024-02-19

DOI

10.1089/fpsam.2023.0281

Facial Plastic Surgery & Aesthetic Medicine
Volume 0, Number 0, 2023
© American Academy of Facial Plastic and Reconstructive Surgery, Inc.
DOI: 10.1089/fpsam.2023.0281

Facial Plastic Surgery
&Aesthetic Medicine

Artificial Intelligence Versus Human Focus Group Rating of Facial Attractiveness

Khodayar Goshtasbi, MD,[1] Amir A. Hakimi, MD,[2] and Brian J.F. Wong, MD, PhD[1,3,4*]

Abstract

Background: Many open-access artificial intelligence (AI)-based websites that rate facial attractiveness are available, but none have been compared with human focus group outcomes.

Objective: To compare human and AI-based websites scoring of facial attractiveness of adult female white faces.

Methods: A 40-photograph database of AI-generated adult, white, female, expressionless, and frontal-view facial images were scored by otolaryngology residents and five AI-based facial rating websites: prettyscale.com, attractivenesstest.com, face-score.com/en, hotchat3000.com, and beautyscoretest.com. Sample *t*-test and bivariate correlation were performed for statistical analyses.

Results: The focus group of 24 otolaryngology residents consisted of 62.5% males and 58.3% white participants. There was a strong positive correlation between average human score and average AI score for each photo (Pearson's correlation 0.84, $p < 0.01$). The average human raters' scores were significantly lower than the average AI scores (5.0 ± 1.8 vs. 6.9 ± 0.9, $p < 0.01$). Thirty images (75.0%) had statistically higher scores from the AI websites versus the focus group. On correlation analysis, all AI-based websites individually had scores that positively correlate with the human scores (all $p < 0.05$).

Conclusion: AI-based websites and human focus-group scoring of facial attractiveness of adult white female faces were significantly correlated with the AI ratings biased toward higher values, encouraging their cautious utilization in future research.

Introduction

Facial attractiveness is an intrinsically difficult quality to measure. Modern studies have used focus groups, which may include trained/expert evaluators or lay participants, to score facial attractiveness.[1–6] The attractiveness scores are used in the context of research across a wide range of subjects and academic disciplines.[1–7] Despite the established validity of using large focus groups, obtaining a high number of participants can be difficult with limitations, including selection bias, observer expectancy effect, or participation exhaustion. These limit the scope and breadth of certain facial attractiveness studies or become a barrier to researchers who have limited access to large-scale and high-quality focus groups, which are generally drawn from students at undergraduate universities.

[1] *Department of Otolaryngology—Head and Neck Surgery, University of California, Irvine, Irvine, California, USA.*
[2] *Department of Otolaryngology—Head and Neck Surgery, MedStar Georgetown University Hospital, Washington, District of Columbia, USA.*
[3] *School of Biomedical Engineering, University of California Irvine, Irvine, California, USA.*
[4] *Beckman Laser Institute and Medical Clinic, Irvine, California, USA.*

Portions of this study have been submitted as an abstract to the AAFPRS Society at the 2023 COSM meeting, May 15–19, Chicago, Illinois.

*Address correspondence to: *Brian J.F. Wong, MD, PhD, Department of Otolaryngology—Head and Neck Surgery, University of California, Irvine, Medical Center, The Beckman Laser Institute and Medical Clinic, 1002 Health Sciences Road, Irvine, CA 92617, USA*, Email: bjwong@uci.edu

KEY POINTS

Question: Are human and artificial intelligence (AI)-based websites comparable in rating facial attractiveness of adult white female faces?

Findings: Average AI-based websites and human focus-group scores of facial attractiveness of adult white female faces were significantly correlated, although the average human raters' scores were consistently lower than the average AI-based website scores. All five AI-based websites individually had positively correlative scores to the human scores.

Meaning: The AI-based websites' facial attractiveness scores of white female faces were a good correspondent of ratings from a human focus group, encouraging their cautious utilization in future research.

Artificial intelligence (AI) algorithms may offer a new approach to evaluate facial attractiveness[8,9] and can theoretically optimize the process by improving speed, access, and objectivity with the added benefit of continuous improvement over time with training. The availability of AI-based facial attractiveness websites has increased and also gained popularity among the public and in the media.[10–12] The objective of this study is to compare facial attractiveness scores generated by AI-based websites with a human focus group in evaluating adult white female faces. We hypothesize that among faces being rated for attractiveness, scores between available AI-based websites will correlate with those of a human focus group.

Methods

This study did not require institutional review board approval due to meeting exemption criteria. Images were selected from an online library by www.generated.photos (Generated Media, Inc.), which archives a large database of high-quality AI-generated facial images. These images were created by the company's generative adversarial networks trained on tens of thousands of photographed images.[13] The database was filtered for adult, white, female, front-facing images with neutral facial expression.

Forty photographs were selected with the objective of identifying a wide range of attractiveness. Pilot studies involved preliminary rating of this data set by three research associates. The mean attractiveness score was 6.3 ± 1.6 on a 10-point Likert scale, which suggested a good range of attractiveness. This data set was then used for both human and AI evaluation.

For focus group evaluation, an anonymous online survey was created using www.typeform.com. The survey comprised 42 questions: gender (male vs. female), race (white vs. nonwhite), followed by the 40 images with star-rating formatted responses. Three different versions of the survey were created with different randomized order of the images to account for response fatigue, and these different survey versions were randomly distributed to the participants.

Responders were asked to rate the attractiveness of the photographs on a scale of 1–10, with 1 being the least attractive and 10 being the most attractive. A focus group of otolaryngology—head and neck surgery residents were chosen for participation. This focus group was chosen given the residents' experience with facial analysis and cosmetic standards. The survey was sent to 30 residents from three different U.S. residency programs.

For the AI rating portion, we searched the internet with various combinations of terms ''artificial intelligence,'' ''AI,'' ''attractiveness,'' ''attractive,'' ''beauty,'' ''scoring,'' and ''rating.'' Five publicly available, free-of-charge, and popular websites that offer AI-based facial attractiveness ratings were used: 1, prettyscale.com; 2, attractivenesstest.com; 3, face-score.com/en; 4, hotchat3000.com; and 5, beautyscoretest.com.[14–18]

Three of the websites, prettyscale, beautyscoretest, and face-score, gave scores in percentages, and thus their scores were scaled down to a 1–10 range score. The PASW Statistics 18.0 software (SPSS, Inc., Chicago, IL) was used for statistical analyses with a p-value threshold of <0.05 designated for statistical significance.

Results

A total of 24 participants from the resident focus group completed the surveys for an 80.0% response rate. The focus group consisted of 15 males (62.5%) and 14 white (58.3%) participants. The average time to complete the survey was 5.3 min. The average attractiveness score for the photographs by the focus group was 5.0 ± 1.8 (median 4.7, range 2.3–8.3). The standard deviation (SD) range of 0.9–1.8 and mean of 1.3 suggested no significant scoring disagreements between the raters for any of the photographs.

Figure 1 demonstrates several examples of the surveyed photographs. On paired sample t-test, there was no statistically significant difference in the average scoring between male and female raters (4.9 ± 1.8 vs. 5.0 ± 1.9, $p = 0.31$). However, there was a statistical difference depending on the raters' race, with nonwhite raters giving a higher average attractiveness score than white raters (5.3 ± 1.7 vs. 4.7 ± 1.8, $p < 0.01$).

The 40 images were then independently rated by the five AI-rating websites, which resulted in an average attractiveness score of 6.9 ± 0.9. The average SD of 1.2 (range 0.7–1.9) was comparable with the average SD of human raters ($p = 0.17$), suggesting a similar level of intergroup rating agreement. There was a strong positive correlation between average human score and average AI score for each photo (Pearson's correlation 0.84, $p < 0.01$), and the linear correlation is also depicted in Figure 2.

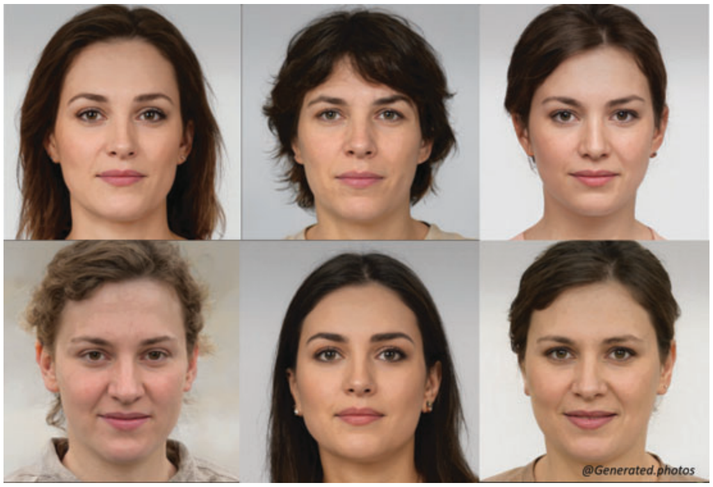

Fig. 1. Examples of the AI-generated faces: 40 frontal-view, expressionless, adult white female faces. AI, artificial intelligence.

This correlation was consistent for both genders and races (Pearson's correlation range 0.81–0.85, all $p < 0.01$). However, on paired sample *t*-test, average human raters' scores were significantly lower than the average AI scores (5.0 ± 1.8 vs. 6.9 ± 0.9, $p < 0.01$). Even when only looking at the 19 photos with a human averaged score of >5.0, the human scores were still significantly lower than the AI scores (6.61 vs. 7.60, $p < 0.001$). Table 1 compares the average scores from the resident focus group and AI websites for the 40 photographs.

Among the 40 photographs, 30 (75.0%) had significantly different scores between the human and AI raters, all of which had consistently higher scores from the AI compared with the human raters. Figure 3 shows three examples of photographs with statistically similar scores, and three examples of photographs with statistically different scores between human and AI raters.

Paired sample *t*-test showed that the human-rated scores were significantly lower than scores by each of the five AI-based websites when compared individually (all $p < 0.01$). Table 2 compares the five AI-rating websites' scores with each other. This demonstrates that prettyscale website and attractivenesstest website had statistically similar scores ($p = 0.206$), although their average scores were on the highest range (8.15 and 8.01, respectively) compared with human scores' 4.9 average. Beautyscoretest and Hotchat3000 also had statistically similar scores ($p = 0.274$) but their average scores (6.45 and 6.27, respectively) were on the lower range and closer to human scores' average.

Face-score had the lowest average score (5.49) and closest to the human scores' average. On correlation analysis, all AI-based websites had statistically significant positive correlation with the human scores, with Pearson's correlations from an order of highest to lowest corresponding to attractivenesstest (Pearson's 0.869), beautyscoretest (Pearson's 0.816), face-score (Pearson's 760), hotchat3000 (Pearson's 0.644), and lastly prettyscale (Pearson's 0.315).

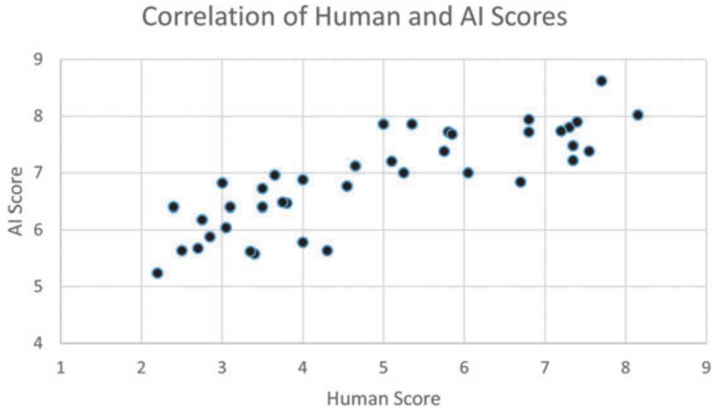

Fig. 2. A strong positive correlation between average human score and average AI website scores for the 40 photos has been demonstrated (Pearson's correlation 0.84, $p < 0.01$). The linear correlation signifies a reliably correlative relationship between the two scoring groups.

Table 1. The average scores (from 1–10) from focus group and artificial intelligence websites of each photograph (numbered 1–40)

Picture	Focus group	AI group	*p*	Picture	Focus group	AI group	*p*-Value
1	7.22	7.29	0.457	21	7.21	7.48	0.284
2	4.29	5.64	**0.046**	22	6.75	6.84	0.452
3	7.70	8.62	**0.043**	23	5.37	7.00	**0.019**
4	3.50	6.40	**0.005**	24	3.21	6.40	**<0.001**
5	5.83	7.38	**0.042**	25	2.58	6.40	**<0.001**
6	6.83	7.72	0.124	26	7.63	7.38	0.321
7	3.63	6.72	**<0.001**	27	2.71	5.68	**0.017**
8	4.04	5.78	**0.006**	28	4.67	6.76	**0.004**
9	7.21	7.80	0.162	29	5.29	7.86	**<0.001**
10	5.96	7.72	**0.016**	30	2.58	5.64	**<0.001**
11	5.75	7.68	**<0.001**	31	7.54	7.90	0.310
12	3.50	5.58	**0.001**	32	5.33	7.20	**0.008**
13	3.00	6.82	**<0.001**	33	3.83	6.46	**<0.001**
14	4.75	7.12	**<0.001**	34	8.25	8.02	0.305
15	5.46	7.86	**<0.001**	35	2.25	5.24	**<0.001**
16	7.29	7.74	0.565	36	3.50	5.62	**0.002**
17	4.13	6.88	**<0.001**	37	3.96	6.48	**<0.001**
18	2.75	6.18	**<0.001**	38	3.08	6.04	**<0.001**
19	2.83	5.88	**<0.001**	39	6.01	7.00	0.085
20	6.87	7.94	**0.041**	40	3.79	6.96	**<0.001**

Bolded values denote statistical significance ($p < 0.05$).

For each photo, the average human score and average AI-based website score are compared through independent sample *t*-test. Thirty of the 40 photographs (75.0%) had significantly different scores ($p < 0.05$), with a higher value from the AI websites in all of these instances.

AI, artificial intelligence.

Discussion

AI is emerging as a tool to evaluate facial attractiveness, and this is the first study that compares attractiveness scores from the most prominent and trafficked AI-based websites[14–18] to a focus group. The utilized database comprised AI-generated images of young white females with a wide range of attractiveness. A strong positive and linear correlation between the AI and expert human scores was observed, which were independent of the gender and race of the human experts.

The results suggest that these AI-based ratings can be used for future research in facial attractiveness. It is important to consider that despite the good linear correlation, the AI ratings are biased toward higher values than the human-based scores. Given the linear correlation, AI-based websites can be a reliable way to compare the attractiveness of two or several images to each other on a relative basis.

In as much as facial attractiveness is a mature field of study and has relied upon focus groups for assesment,[1–7] no study to date has compared AI-based facial rating websites with their human counterpart. The results of this study herein suggest reasonable correlation between the two metrics in our subject database. With further investigation, AI-based facial attractiveness rating websites can potentially be used for research, education, and clinical applications.

Their use may accelerate research in this field and serve as a proxy for large-cohort human focus groups, which can be challenging to implement. In a clinical setting and specifically for facial plastic surgeons, these algorithms maybe be a useful method to evaluate morph preoperative images and provide guidance in planning surgery.

Although the websites used here do not disclose their code/algorithm, some rudimentary information on the analytic process can be garnered. Prettyscale.com attempts to analyze facial symmetry, placement and sizes of various features, shape and size of the subunits, as well as the distance between these subunits.[19]

Attractiveness.com's attractiveness score is based on a deep neural network model (based on ResNet architecture) trained on 100,000 photographs with human-based scores for training.[20] Hotchat3000.com uses CLIP, a machine learning model trained by OpenAI, which was then

Fig. 3. Six photographic examples of the adult white female faces used in the study. The top row photographs had statistically similar scores ($p > 0.05$) between the AI-based website and human focus group, whereas the bottom row photographs had statistically different scores ($p < 0.05$) between the AI-based website and human focus group.

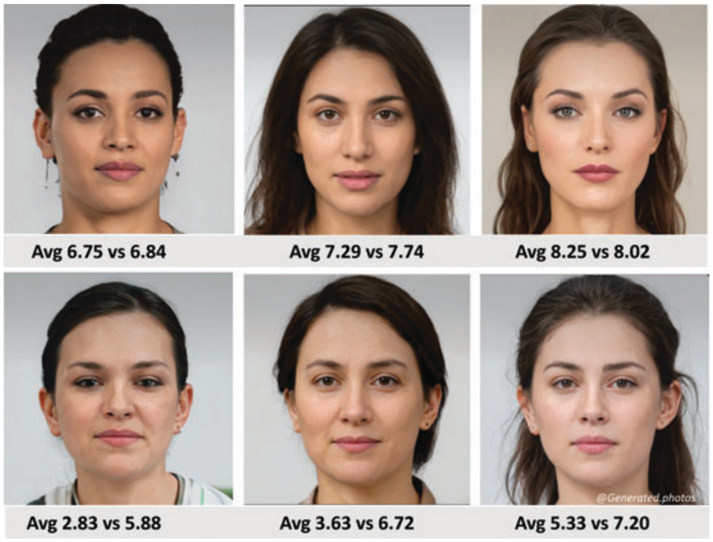

Table 2. Comparison of the five artificial intelligence-rating websites' scores to each other through paired sample *t*-test

Website	Pretty scale	Attractivenesstest	Face-score	Hotchat3000	Beautyscoretest
Prettyscale		**8.15 vs. 8.01** ***p* = 0.206**	8.15 vs. 5.49 *p* < 0.001	8.15 vs. 6.27 *p* < 0.001	8.15 vs. 6.45 *p* < 0.001
Attractivenesstest	**8.01 vs. 8.15** ***p* = 0.206**		8.01 vs. 5.49 *p* < 0.001	8.01 vs. 6.27 *p* < 0.001	8.01 vs. 6.45 *p* < 0.001
Face-score	5.49 vs. 8.15 *p* < 0.001	5.49 vs. 8.01 *p* < 0.001		5.49 vs. 6.27 *p* < 0.001	5.49 vs. 6.45 *p* < 0.001
Hotchat3000	6.27 vs. 8.15 *p* < 0.001	6.27 vs. 8.01 *p* < 0.001	6.27 vs. 5.49 *p* < 0.001		**6.27 vs. 6.45** ***p* = 0.274**
Beautyscoretest	6.45 vs. 8.15 *p* < 0.001	6.45 vs. 8.01 *p* < 0.001	6.45 vs. 5.49 *p* < 0.001	**6.45 vs. 6.27** ***p* = 0.274**	

Bold values show the insignificant *p*-values ($p > 0.05$) suggesting similar scores by those two websites. The table demonstrates that prettyscale website and attractivenesstest website had statistically similar scores ($p = 0.206$). Likewise, Beautyscoretest and Hotchat3000 also had statistically similar scores ($p = 0.274$).

trained on two data sets of human faces (SCUT-FBP5500 and Hotornot.com), which were labeled with human ratings.[21] Face-score.com uses the Face++ application programming interface that also accounts for lighting, pose, symmetry, and expressions.[22]

The major drawback of all these models is that they reflect the data sets that they are trained on, without novel decision-making capabilities. As these online resources continue to evolve, and more such websites become available, they should become more precise in rating facial attractiveness.

There are important limitations that should be considered. First, the database consisted only of adult white females (appearing in their 20s–30s) with neutral facial expressions. This was by design to decrease the variability of the photographs and control for the effect of gender, race, and age on attractiveness perception. Regardless, this implies that the presented results may not be extrapolated to other characteristics such as male gender, nonwhite race, or faces with expression.

Broader investigations are warranted with different race and ethnicities, age groups, and expressions, and must evaluate whether the correlation we observed remains consistent. Since beauty perception also relies on race,[23] future AI-based websites may allow subcategorizing training data by the race/ethnicity of the evaluators to more accurately reflect the outcomes. The emotion of the faces were also not evaluated in this study even though emotion can have a significant influence on the perceived attractiveness.[24,25]

Given that in real life we observe most human faces in motion and with emotion, future studies should address this issue. Finally, the methodology may not be extrapolated to other photographic views such as lateral and profile view. In fact, some of the websites (e.g., prettyscale.com) that rely heavily on facial symmetry are unable to provide results for nonfrontal views. This is a limitation of current software and will be a challenge to address as it implicitly requires three-dimensional information.

Lastly, granular demographic data from the human focus group (age, ethnicity, etc.) were not collected for anonymity reasons, and the focus group race was self-assigned.[26] Within the context of this study design, a strong linear correlation between AI and human focus group scores was observed. Future studies should be cognitive of the caveat that despite the good correlation, the AI scores may be inflated compared with realistic human measurements, and must be appropriately scaled and calibrated.

Conclusion

This study supported our hypothesis that there was a strongly positive and linear correlation between facial attractiveness scores of AI-based websites and a human expert focus group. Despite the significant correlation, the AI scores were significantly higher than human scores on a consistent basis. This suggests that AI-based websites may provide an efficient means to gauge facial attractiveness and may be cautiously incorporated in further research, while acknowledging that the AI-based scores are biased toward higher values. Further studies are warranted to evaluate the comparability of these AI-based scores with human focus-group-based scores in other attractiveness contexts or as the AI websites continue to improve in the future.

Authors' Contributions

Conceptualization, data collection, data analysis, statistical modeling, writing and editing the drafts, and final approval of the version to be published by K.G. Data collection, data analysis, writing and editing the drafts, and final approval of the version to be published by A.A.H. Conceptualization, data interpretation, editing the drafts, and final approval of the version to be published by B.J.F.W.

Author Disclosure Statement

No competing financial interests exist.

Funding information

No funding was received for this article.

References

1. Sinko K, Cede J, Jagsch R, Strohmayr AL, McKay A, Mosgoeller W, et al. Facial aesthetics in young adults after cleft lip and palate treatment over five decades. *Sci Rep*. 2017;7(1):15864.
2. Popenko NA, Devcic Z, Karimi K, Wong BJF. The virtual focus group: a modern methodology for facial attractiveness rating. *Plast Reconstr Surg*. 2012;130(3):455e–461e.
3. Wong BJF, Karimi K, Devcic Z, McLaren CE, Chen W-P. Evolving attractive faces using morphing technology and a genetic algorithm: a new approach to determining ideal facial aesthetics. *Laryngoscope*. 2008;118(6):962–974.
4. Popenko NA, Tripathi PB, Devcic Z, Karimi K, Osann K, Wong BJF. A quantitative approach to determining the ideal female lip aesthetic and its effect on facial attractiveness. *JAMA Facial Plast Surg*. 2017;19(4):261–267.
5. Gu JT, Avilla D, Devcic Z, Karimi K, Wong BJF. Association of frontal and lateral facial attractiveness. *JAMA Facial Plast Surg*. 2018;20(1):19–23.
6. Hu AC, Hong EM, Dunn BS, Gu JT, Wong BJF. The effect of a consumer nose reshaper on nasal tip projection and the perceived attractiveness of Asian females. *Facial Plast Surg Aesthetic Med.* 2021;23(4):314–315.
7. Richer V, Berkowitz J, Trindade de Almeida A. Eyebrow shape preference across age, gender, and self-reported ethnic group. *Dermatologic Surg*. 2023;49(2):171–176.
8. Hebel NSD, Boonipat T, Lin J, Shapiro D, Bite U. Artificial intelligence in surgical evaluation: A study of facial rejuvenation techniques. *Aesthetic Surg J Open Forum*. 2023;5:ojad032.
9. Liu AS, Salinas CA, Sharaf BA. Using artificial intelligence to quantify sexual dimorphism in aesthetic faces: Analysis of 100 facial points in 42 Caucasian celebrities. *Aesthetic Surg J Open Forum*. 2023;5:ojad046.
10. Manandhar-Richardson T. An attractiveness researcher puts the internet's most popular "hotness algorithms" to the test. Medium. 2018. https://thomas-richie-richardson.medium.com/an-attractiveness-researcher-puts-the-internets-most-popular-hotness-algorithms-to-the-test-3278dbcb03b2 [Last accessed: September 14, 2023].
11. Mahdawi A. This AI-powered app will tell you if you're beautiful–and reinforce biases, too. The Guardian. March 6, 2021. Available from: https://www.theguardian.com/commentisfree/2021/mar/06/ai-powered-app-tell-you-beautiful-reinforce-biases [Last accessed: September 14, 2023].
12. Kato B. AI 'hot or not' tool rates celeb faces: Sydney Sweeney and more shocking results. The New York Post. April 28, 2023. Available from: https://nypost.com/2023/04/28/ai-hot-or-not-tool-rates-celeb-faces-sydney-sweeney-and-more/ [Last accessed: September 14, 2023].
13. GeneratedPhotos. Frequently Asked Questions Generated Photos. 2023. https://generated.photos/faq [Last accessed: September 14, 2023].
14. prettyscale. https://www.prettyscale.com/ [Last accessed: September 14, 2023].
15. attractivenesstest. https://attractivenesstest.com/ [Last accessed: September 14, 2023].
16. hotchat3000. https://hotchat3000.com/ [Last accessed: September 14, 2023].
17. beautyscoretest. https://www.beautyscoretest.com/ [Last accessed: September 14, 2023].
18. face-score. https://face-score.com/en [Last accessed: September 14, 2023].
19. Prettyscale Help & FAQs. https://www.prettyscale.com/help/ [Last accessed: September 14, 2023].
20. How it works Attractivenesstest. 2023. https://attractivenesstest.com/blog/How the app works [Last accessed: September 14, 2023].
21. Whiddington R. How Hot Are You? Art Collective MSCHF's New Chat Site Lets A.I. Do the Evaluating. Artnet News. April 27, 2023. Available from: https://news.artnet.com/news/mschf-hot-chat-3000-ai-chatbot-2291171 [Last accessed: September 14, 2023].
22. Eden AI. Best Face Recognition APIs in 2023. https://www.edenai.co/post/best-face-recognition-apis#:~:text=Face%2B%2B [Last accessed: September 14, 2023].
23. Dimitrov D, Kroumpouzos G. Beauty perception: a historical and contemporary review. *Clin Dermatol*. 2023;41(1):33–40. doi: 10.1016/j.clindermatol.2023.02.006
24. Tracy JL, Beall AT. Happy guys finish last: the impact of emotion expressions on sexual attraction. *Emotion*. 2011;11(6):1379–1387. doi: 10.1037/a0022902
25. Hester N. Perceived negative emotion in neutral faces: gender-dependent effects on attractiveness and threat. *Emotion.* 2019;19(8):1490–1494. doi: 10.1037/emo0000525
26. Flanagin A, Frey T, Christiansen SL, Bauchner H. The reporting of race and ethnicity in medical and science journals: comments invited. *JAMA*. 2021;325(11):1049–1052.