The Impact of Facial Characteristics on the Face-Naming Effect Through Human and AI-Generated Faces

Levi Cottrell

Elphinstone Secondary School

Gibsons, BC, Canada

January 15, 2024

Abstract

Previous research has shown a ‘face-naming’ effect, where humans and machine learning algorithms are able to guess the name of a face given only the face and 4-5 name options at an accuracy greater than random chance. This effect was previously studied in French and Israeli contexts, whereas this study was carried out within a mid-sized high school in a small town near Vancouver, BC, Canada, with 23 participants. Two sections comprised this study. Participants were tested with human faces in the first section, with their real names, and 3 other names of similar popularity. An accuracy of approximately 19.48% was found in this experiment, where random chance is 25%. Participants were tested with AI-generated faces in the second experiment, each with similar characteristics to a corresponding human face in the first experiment. These faces were tested with the same names as their similar human counterpart. An accuracy of approximately 17.8% was found in this experiment. This suggested that one: in this context, the face-naming effect did not emerge; and two: the face naming effect is barely influenced by, if not hindered by, the similarity of facial characteristics in these AI-generated faces.

Background

Names are an integral part of the human experience. The action of naming another person originates from needing to refer to people as individuals instead of groups. Although it is essential to us, humans are not the only species that name each other; dolphins and elephants have complex social relationships, including naming systems. Elephants, for example, use specialized calls to address individuals. Dolphins use a similar system, using signature whistles instead of calls to identify each other. Humans, however, have an obsession over names. We name or find the name of every item, idea, and person we meet. Our sociality is the root of this, as we often give names to the many individuals and groups we contact. To converse, we need communal names for items and ideas. Because of our community, sociality, and need for survival, humans have personified different words and their sounds, usually based on safety and danger. Softer phonemes (l, s, r, m, etc…) are usually common in the names for 'softer' items and ideas, whereas harsher phonemes (k, t, z, b, etc…) are generally found in relatively 'harsher' words.

Similarly, people can guess others' names at a rate much higher than random (Zwebner et al., 2017). In this study, I want to discover whether the face-naming effect is present in differing contexts to those already explored and to find if facial characteristics impact said face-naming effect. This study will have two groups of participants, both given a demographic survey. I will give the first group, or control group, a collection of photographs of humans, as well as a list of four names, one being the person's real name. They will match a name to the picture. In the second group, I will give the same list of names, and a set of pictures of similar, albeit different pictures generated by a GAN (generative adversarial network) AI model. After I collect all of the data, I will compile the information and extract possible biases and effects of naming on facial appearance.

Review of Literature

Zwebner, Sellier, Rosenfield, Goldenberg, and Mayo (2017) researched the effects of social perceptions on facial appearance, how well people can guess the names of faces, and the accuracy of computer algorithms guessing the names of faces. They tested their hypothesis of a *face-naming effect,* where a person's name shapes their facial appearance. Of their eight total studies, three are largely relevant to this study, including 1A, 1B, and 2.

In study 1A, a hypothesis-blind assistant obtained photos of Israeli young adults (20-30) from a large social network platform. They gathered 'filler' names from local websites containing name catalogues. They added four filler names for each of the photos and true names, in the manner of multiple-choice. No filler name matched any true name and they repeated zero names. In total, they selected 80 filler names and 20 pictures with true names. The order of images and name choices was random for each participant. Participants consisted of one hundred twenty-one local Israeli students. (60 women, 61 men, all native speakers, M_{age} = 27.77 years [SD 4.08]) They asked each participant to choose from the five names for the true name of the person depicted in the picture for each of the twenty faces. If a participant was familiar with a face, they discounted the results of this test, to limit artificial inflation of the face-name matching rate. The result was an average of 28.21% accuracy, where chance is 20%. This study established the existence of a possible face-naming effect.

Study 1B replicated study 1A with a larger sample of stimuli to refute the possibility that the face-naming effect is due to name popularity. In this study, they selected filler names from other true names, instead of zero repeated names. In studies 1B and 2, they gathered photos through university campus announcements instead of online images and names. They took these pictures front-on with a neutral expression at a resolution of 400x400 pixels. The facilitators also limited

cosmetic effects like makeup and headwear. This set of target photos and names consisted of 50 local Israeli students. They split the fifty photos into two groups of 25 each and given to participants evenly to keep set sizes manageable. Participants of study 1B consisted of sixty-seven Israeli students (32 women, 35 men, all native speakers, M_{age} = 25.59 years [SD 3.36]) The procedure of study 1B was similar to 1A, although instead of five possible name choices per picture, they used four. They discounted eight of 1600 data points due to participant familiarity. The result of this study was an average success rate of 29.91%, where chance was 25%. Study 1B shows a replication of the face-naming effect while addressing the possible complication of external filler names in Study 1A. They theorize that this replication suggests that the face-name matching is true, and not solely a product of filler name features that cause them to stand out. (Ex. Name popularity or uniqueness.)

Study 2 aimed to replicate the face-naming effect in a different culture to determine if the effect transcended culture. Although it has a different method to 1B, study 2 also aimed to discount the theory that the effect appeared due to differences in target and filler names. Because of this, they limited names related to religion and ethnicity. They limited target faces and participants to French Caucasians to keep ethnicity constant. In this study, individuals emailed pictures due to a lack of target faces. Participants of Study 2 included one hundred sixteen French participants (M_{age} = 23 years [SD 2.90]) Their set of target names and photos consisted of French Caucasian given names that were neither popular nor rare. They conducted this study similarly to the first two studies, where they showed participants a list of four names, with one being the true given name. They discounted zero data points due to zero participant familiarity with target faces. Participants matched names successfully at an average rate of 40.52%, where chance was 25%. This success rate suggests that the face-naming effect is present in separate cultures.

Miller et al (2023) studied AI-Hyper-realism, specifically in AI-generated faces. In Experiment 1, participants comprised 124 White US residents (62 women, neurotypical, M_{age} = 34.4 years, [*SD* = 8]). In their testing, they used an assortment of one hundred AI-generated faces and one hundred human faces, all Caucasian (half male, half female). They created the AI-generated faces using a StyleGAN2 model and matched them with a similar human face. They gave participants approximately 100 assorted faces (half AI, half human), and asked them to decide if a picture was human or AI-generated. After this, they asked participants to rate their confidence in their answer from 0-100, where 0 is not at all confident and 100 is completely confident. They found that participants were more likely to select artificially generated faces as human than human faces, and human faces were more likely to be judged as AI than AI faces. Participants most confident in their choice of 'human face' were more likely to choose an AI-generated face. In Experiment 2, participants consisted of 610 people (312 women, M_{age} = 35.3 years, [*SD* = 8.6 years]). They asked participants to rate AI and human faces on one of 14 attributes, although the researchers did not inform them that AI faces were present. Participant screening was otherwise identical to Experiment 1. Attributes tested included distinctiveness, attractiveness, familiarity, memorability, age, and nine attributes commonly mentioned in reasoning during Experiment 1. They found that participants described AI faces as average, familiar, attractive, and less memorable than actual human faces.

These two studies provide a groundwork for this study. Zwebner et al.'s (2017) study analyzed the face-name matching effect, where participants had a significantly higher chance of guessing someone's true name when provided a photo of their face. This face-naming effect is instrumental in finding biases in participants and AI face generation. They also found that the

face-naming effect transcended culture, appearing in studies in both Israel and France. With my control group, I will confirm the existence and accuracy of the face-naming effect in a Western Canadian High School. Miller et al.'s (2023) study on AI Hyper-realism provides evidence to support that AI-generated faces are nearly indistinguishable from human faces, especially those of Caucasian ancestry. This supports the assumption that participants generally cannot tell between AI and human faces.

Methods

Materials

Materials for this study included filler names, AI-generated target faces, and target human faces and names. First, I collected target names and matching human faces. A study-blind assistant collected these faces with directions to choose 20 faces from a popular social media platform along with their matching first names. Since humans see AI-generated Caucasian faces as more human than other ethnicities (Miller et al, 2023), the assistant subjectively perceived the chosen faces as Caucasian or White. Since humans perceive Caucasian faces as more human, they should therefore be more likely to have a matching human name. Although the pictures could not be exactly and perfectly formatted with consistent lighting, the assistant collected photos where the target face: takes up most of the image, faces the camera, and is generally well-lit. Subsequently, the assistant cropped these images to a square and assigned the target names to the corresponding faces. If possible, they removed the background. Next, they collected filler names. Filler names are non-target names shown beside the target name and photo to provide a more diverse name selection. We needed sixty filler names for the 20 target names because we presented each target name alongside three additional ‘filler’ names. The assistant gathered filler names from the closest three names on a name frequency chart, namely the Statistics Canada Baby Name Observatory. I instructed the assistant to enter the target names into the field labelled ‘Select Name at Birth,’ looking at the name’s frequency based on an approximate age and binary sex. We used a different name frequency chart if the target was born prior to 1993. After finding the frequency of the name, the assistant selected three closely frequent names. For example, if the target was a male born in 2000 named ‘Zachary,’ sample filler names would be ‘Tyler,’ ‘Ethan,’ and ‘Benjamin.’ We excluded and replaced a name if it was unusual or matched a separate target name. Lastly, we

collected AI-generated faces. A second study-blind assistant gathered these faces from a StyleGAN2 model and free-to-access website, named “thispersondoesnotexist.com.” The assistant found Caucasian faces that matched the basic categorization of the target human faces. The first assistant described the categorization of the human faces and gave these descriptions to the second assistant. Although we found some deviation in collecting the faces, the photos collected all faced the ‘camera’ and were lit. We based the selection guidelines on age, gender, and hair. The age categories were: young adult (~18-30), adult (~30-45), middle age (45-60), and senior (~60+). The gender categories were basic, only generalized and subjectively judged as ‘male’ and ‘female.’ Hair categories consisted of ‘head-hair:’ fully bald, half-bald, short, medium, long; and facial hair: clean-shaven, mustache, beard. The assistant assigned the StyleGAN2 photos the same target name and filler names as the similar target human face.

Procedure

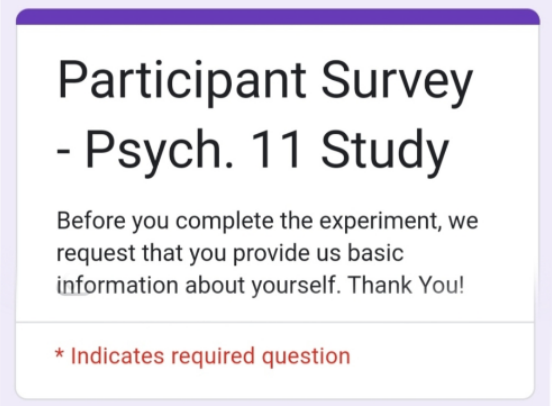
We split this study’s procedures into two testing sections. The first section was a control group, to establish the face-naming effect in a Canadian context. This section was a close recreation of Zwebner et al.'s (2017) 1A study, as described in the Review of Literature above. This control group used human faces, target names, and filler names. I printed each photo onto a piece of paper, with the target and filler names ordered randomly beneath it, labelled a, b, c, and d. We gave each participant a stapled booklet containing the pictures and corresponding names (Example to the Left) and a separate paper to input their name selection. We instructed participants to

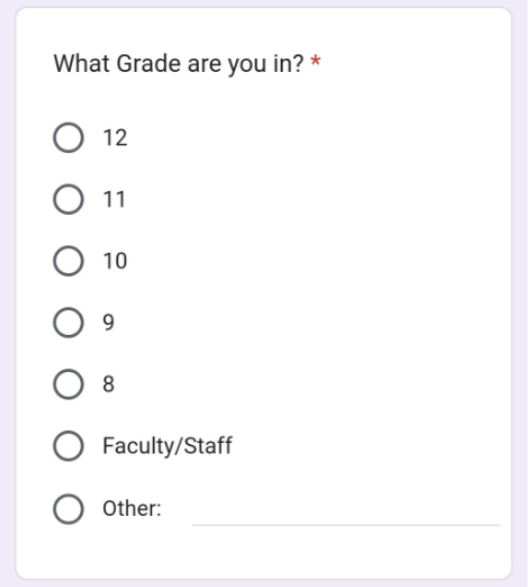
a. Richard
b. Hugh
c. John
d. Dan

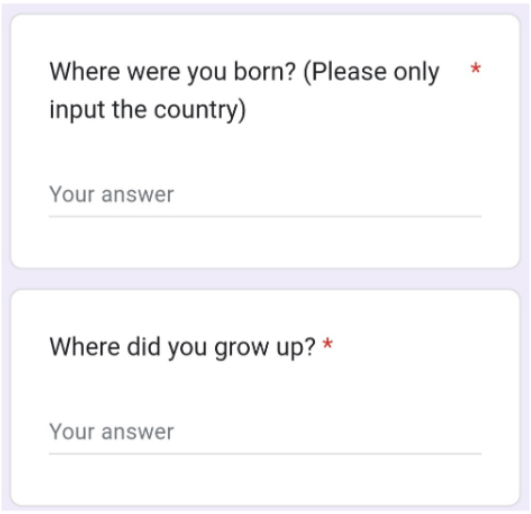
look at the photo on the paper and select which of the four names was the true name of that person. We also instructed them to tell the facilitator if they were familiar with or recognized any of the faces, to reduce artificial inflation of the face-naming effect. The second section was when StyleGAN2 faces became a factor. This section aimed to test if AI-generated headshots have the same name as their similar counterpart, and if the face-naming effect is due to basic features rather than some other factor. The main procedure was the same as the first section, however, we made minor changes. We ordered the AI-generated faces randomly in the booklet given to participants. We instructed participants to look at the photo on the paper and select which of the four names was the true name of that person. We also instructed the participants to tell the facilitator if they were familiar with or recognized each face.

Participants

Before the experiment started, participants completed a demographic survey. We designed this survey to find basic information that may skew their answers, such as culture, ethnicity, and identity. The survey is below.







After completing this survey, the participants completed the booklet described above in "Procedures." When they completed the experiment, we compensated participants with assorted candy and chocolate.

Data

Participants

Age

Grade	8	9	10	11	12	Staff
Participants	1	4	1	5	11	1

Nationality

Country	Canada	USA	Germany	Thailand	Philippines
Participants	18	2	1	1	1

Gender

Gender	Male	Female
Participants	15	7

Religion

Religion	None	Christian	Agnostic	Buddhist	Unique
Participants	10	7	4	1	1

Overall Results

Each percentage/score is based on the participants' score on the experiment, out of 20. For example, if a participant scored 4/20, their accuracy would be 20%.

Varied (AI) Experiment (%)	Control (Human) Experiment (%)
Mean = 17.81%, Median = 20%	Mean = 19.48%, Median = 17.5%

Trial 1 Results

Varied (AI) Experiment (/20)	Control (Human) Experiment (/20)
2	1
3	3
3	4
4	4
4	4
5	5
6	6
6	6
Mean = 4.125 (20.625%), Median = 4 (20%)	Mean = 4.125 (20.625%), Median = 4 (20%)

Trial 2 Results

Varied (AI) Experiment (/20)	Control (Human) Experiment (/20)
1	4
3	4
3	3
5	
Mean = 3 (15%), Median = 3 (15%)	Mean = 3.67 (18.33%), Median = 4 (20%)

Discussion

Participants

Overall, participants comprised a sampling of students from Grades 8 to 12, as well as a single data point from a school faculty member. This study had 23 participants, including one faculty member and 22 students from Elphinstone Secondary. Of these students, the mean score was approximately 11 (10.96) however, the median grade was 12 (11.5 rounded), with half the student data points coming from grade 12. Distribution of student participants' ages is presented in the table below. Within these grades, there was some variance for Nationality, Religion, and Gender. The majority of participants consider themselves to be Canadian, at 78.3% self-reporting Canadians. Along with this, 95.6% of participants self-reported that they grew up in or in combination with Canada. For Religion, 43% stated they have no religion, 30.4% stated they were Christian, 17.4% stated they were Agnostic, and 8.6% reported another religion/belief. Of those who reported a Gender, 31.81% were Female, and 68.18% were Male.

Grade	8	9	10	11	12
Participants	1	4	1	5	11

Trial 1

In the first trial, there were 16 participants (Grade distribution below). Because both the control and and varied experiments were conducted at the same time, half of the participants completed the control (human-face experiment), and half completed the varied (AI-faces experiment). In this Trial, Participants were first instructed to complete the demographic survey, and complete the experiment two at a time.

Grade	8	9	10	11	12	Staff
Participants	0	4	0	2	9	1

Trial 2

In the second trial, there were seven participants (grade distribution in Table below). Nearly half the participants completed the human-face experiment, and over half completed the AI-face experiment (3:4, respectively). In this trial, participants were first instructed to complete the demographic survey, and completed the experiment four at a time.

Grade	8	9	10	11	12
Participants	1	0	1	3	2

Results

Overall, participants in the human-face experiment scored an average of 19.48%, whereas in the AI-face experiment participants scored an average of 17.81%. In both experiments, the chance of correctly guessing the true name of a face was 25%, meaning in both the human-face experiment and the AI-face experiment, participants on average scored worse than random chance. In the human-face experiment, participants scored a median of 20%, and in the AI-face experiment, participants scored a median of 17.5%. In the human-face experiments, the most common accuracy was 20%, and in the AI-face experiments, it was 15%. In each experiment, two participants scored at an accuracy above random chance, all with an accuracy of 30%.

Grade Based

Within the Grade 8 results, there was only one data point, that being 15%. Within the Grade 9 results, there were 4 data points, equally distributed between human and AI-face experiments. The accuracy in the human-face experiment was 20% and 22.5% in the AI-face experiment. For grade 10 results, there was only one data point, that being 15%. In Grade 11 results, the accuracy was 17.5% for the human-face experiment, and 13.34% for the AI-face experiment (Median = 15%). For the Grade 12 group, they achieved an accuracy of 20% in the

human-face experiment, and 20.83% in the AI-face experiment. In this group, the median accuracy is 20% and 22.5%, respectively. For the Staff group, there was only one data point, that being 30% accuracy in the human-face experiment.

Religion Based

Of the five religion-based groups present in this experiment, the most numerous was the 'No religion' group. This group had an accuracy of 23% in the human-face experiment, and 21% in the AI-face experiment. In this group, the median accuracy is 20% and 25%, respectively. Next is the Christian group, with an accuracy of 20% in the human-face experiment, and 17% in the AI-face experiment. In this group, the median accuracy is 20% and 15%, respectively. Within the Agnostic group, the accuracy was the same for both experiments, at 17.5%. The Buddhist group had only one data point, that being from the human-face experiment at 15%. The Unique beliefs group also only had one date point, also from the human-face experiment, at 15%.

Gender Based

Male and Female were the two genders represented in this study. The Male group had an accuracy of 18.75% in the human-face experiment, and 15.71% in the AI-face experiment. The median accuracy for Male participants was 20% and 15%, respectively. The Female group achieved an accuracy of 23.33% in the human-face experiment and 22.5% in the AI-face experiment. The Median accuracy for Female participants was 20% and 22.5%, respectively.

Limitations

For this study, there were twenty-three participants, however in Zwebner et al. (2017) and Miller et al.'s (2023) studies, their participant yield was significantly greater, generally over

one-hundred participants. Also, in many of the demographic sections, there is somewhat of a diversity, whereas in Zwebner et al.'s (2017) study —which discovered and claimed the face-naming effect— participants were generally of a single cultural demographic, which could have led to culture-specific results.

Suggestions for Future Research

Future researchers should work with a larger base of participants. This would increase the reliability of results. Secondly, the effect should be studied in more than one environment with different population demographics, such as environments with more conservative populations v. environments with more liberal populations. This would serve to show if political or social opinion impacts the effect through family names v. names new to a lineage. Thirdly, more female participants should be studied. My data might suggest that females have a much higher accuracy than that of male participants, and this should be studied for generalization and confirmation of the heightened accuracy. Next, the method for which human faces were collected should be improved, with more consistency in lighting, positioning, and emotion. Lastly, AI-generated faces where collected should be streamlined and improved, either with stricter descriptions or with more similarities in facial structure, body typing, or secondary ethnicity.

Conclusion

The joint aims of this study were to first, determine if the ‘face-naming’ effect is present in a mid-sized high school in a small town near Vancouver, BC (largely comprising people of European descent), and second, if facial features and characteristics have anything to with this face-naming effect. The results suggest that the face-naming effect did not appear in this context, possibly because of different naming conventions and local culture compared to the French and Israeli participants in Zwebner et al.'s (2017) study. The second hypothesis is that facial characteristics have an effect on the face-naming effect. I believe this is weakly supported. The mean accuracy of participants in the AI-faces experiment is approximately 17.8%, which is ~1.7% lower than the accuracy of participants in the human-faces experiment. This could show a disconnect between exact facial characteristics and the face naming effect. I believe the greatest factors that influence the accuracy of the face-naming effect are culture and gender of the participant and the culture of the person whose name is guessed.

Citations

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E.G., & Dawel, A. (2023, November 13). *AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones*. journals.sagepub.com. Retrieved October 7, 2024, from https://journals.sagepub.com/doi/10.1177/09567976231207095

Zwebner, Y., Sellier, A.-L., Rosenfeld, N., Goldenberg, J., & Mayo, R. (n.d.). *We Look Like Our Names: The Manifestation of Name Stereotypes in Facial Appearance*. APA.org. Retrieved Sept. 26, 2024, https://www.apa.org/pubs/journals/releases/psp-pspa0000076.pdf