The Minimal Group Paradigm Revisited: On the Design and Testing of AI-Generated Preference-Indifferent Avatars for Online Experiments

Ashley E. Oaks,1,2* Matthew I. Jones,3 Kai Xu,4 Feng Fu,5 Nicholas A. Christakis1,2,6

[1] *Human Nature Lab, Yale University*
[2] *Department of Statistics and Data Science, Yale University*[5]
[3]
[4]
[5]
[6] *Department of Sociology, Yale University*

*Corresponding author: ashley.oaks@yale.edu

Abstract

It is of broad interest to study human collective human behavior with laboratory experiments. Such experiments sometimes require subjects to be given avatars so that others in their group can identify and track them across time and distinguish one subject from another. Ideally, experimental subjects should have no a priori preference for any assigned group identifiers that originate outside the experimental setting. Here, in order to identify a set of preference-indifferent avatars that can be used in social experiments, we employ the CurmElo algorithm, a forced choice approach that uses the Elo ranking algorithm along with novel techniques for identifying preference polarization and intransitivity in preference rankings. We detail a process of generating differentiable circular images using AI tools (i.e., the "avatars") as well as the application of CurmElo to produce preference rankings from a population of human subjects. In addition, we detail preference consistency characteristics of both the subjects and the avatars. Our results lead to a handful of avatars that can reliably be used as group identifiers in experiments involving groupings while minimize confounding effects associated with their use. We also discuss the implication of our findings for future group behavior studies, especially those requiring or exploiting identity.

Introduction

A key feature of human society is group membership or group identity, which affects behavior in a wide range of settings, from cooperation to coalition building to social learning to ideology (Tajfel et al. 1971; see Christakis 2019 for a review). Having a common group identity makes people more likely to have positive regard for each other, be friends, and cooperate. Surprisingly, these identities can affect behavior even if group membership is completely arbitrary and very superficial, as shown by classic work in the "minimal group paradigm" (Diehl, 1990; Hartstone & Augoustinos, 1995; Jackson et al., 2019; Karp et al., 1993).

However, the minimal group paradigm, which assigns group identities arbitrarily, may not eliminate the possibility that the group *identifiers* themselves are also affecting behavior. People who are arbitrarily given red or blue shirts, or who make choices based on Klee or Kadinsky paintings, may have baseline preferences about these identifiers (Castano, 2002; Diehl, 1990; Hong and Ratner, 2021). That is, many types of experimentation within the social sciences (such as audit studies to identify racism or sexism, or evaluations of in-group preference among children) involve the use of ostensibly arbitrary identifiers or groups (Bloom, 2013; Brewer, 1979; Currarini & Mengel, 2016; Yamagishi, 1998). While these identifiers may indeed be randomly assigned, *a priori* one cannot say that a given population is indifferent between any two arbitrary identifiers. People may naturally prefer some sounds, colors, letters, or symbols over others, and so any behaviors in experiments (such as preferences for other groups, or attitudes towards one's own or others' behaviors) may relate to the assigned labels that may pre-exist the experiment.

As an exaggerated example, suppose we design an experiment with two groups that were experimentally manipulated in some way but who were represented by red and blue flags, and the data showed that all participants treated members of the blue group nicer than the red group. We could conclude that the experimental treatment caused the blue participants to be treated differently by other participants. However, it is also possible that people just like the color blue and so were unintentionally treating the blue group differently. If people really had no preferences over colors, then we could easily conclude that the experiment parameters is what led to their positive treatment; yet, people do have favorite colors and assign meanings to colors (Guilford & Smith, 1959; Park & Guerin, 2002; Schloss et al., 2013).

Hence, here, we set out to find a collection of group identifiers, or "avatars," over which real people have no significant preferences. People assigned to random colors, shapes, animals, or even symbols might have baselines preferences for their own or others' avatars or baseline expectations about how people with a particular avatar should act. Our work was indeed prompted by the recognition that people in online experiments who were assigned avatars (so that they might anonymously track the identity of others whom they were interacting with online) often had preferences about those avatars, and that these preferences might affect their behavior in the underlying experiment in ways that interfere with inference. In cooperation games, for instance, people randomly assigned animal avatars might act and treat others differently depending on their own or others' animals; a person assigned a lion avatar in a public goods game might offer smaller amounts to, or expect higher amounts from, their alters, and might do this especially if, say, their alter was assigned a bunny avatar (Emmerich et al., 2024).

Indeed, in our own pilot work, we found that switching to seemingly more abstract solutions – like colors or certain kinds of symbols – did not solve this problem of bias. For instance, people have preferences for colors and therefore might prefer to treat others with particular colors (say green) better than they treat people assigned to other colors (say gray). We even experimented with variations of Gabor patches (ordinarily used to study vision), but found that people had preferences, say, for whether the wavy lines were oriented vertically (like flames) or horizontally (like waves). See Appendix 1 for an description of earlier iterations of candidate avatars.

Indeed, in the search for monikers about which people might be indifferent, some of our previous work explored four- and five-letter nonsense words (for instance, people prefer the word PESAM to the word NUFIW, but they are indifferent between DOSOZ and BIHIR) (Sankaran et al., 2021). But these words are not visual and can be harder to remember and track across time. Thus, here, we focus on small images which, being more intuitive and immediately recognizable than strings of characters, are better suited for representing group identity (Callahan & Ledgerwood, 2016). We generated 110 small, circular images using artificial intelligence that were designed to be abstract and meaningless, and then recruited 360 human subjects to report their preferences between pairs of avatars. With this data, we identified sets of identifiable yet *preference-indifferent* avatars that are equally popular, providing us (and future researchers) with a set of group identifiers over which participants are indifferent and that would not unnecessarily distort any study of group behavior.

Methods

To identify a set of preference-indifferent avatars, we begin by creating a large set of abstract images (see Figure 1 for some examples). Using DALL-E 3, we created a prompt to generate a large set of patterns that appeared meaningless. After experimenting with various prompts and testing the resulting avatars, we used the following prompt that was able to generate various distinct patterns: "Black and white random pattern, no shading, minimalistic, very simple."

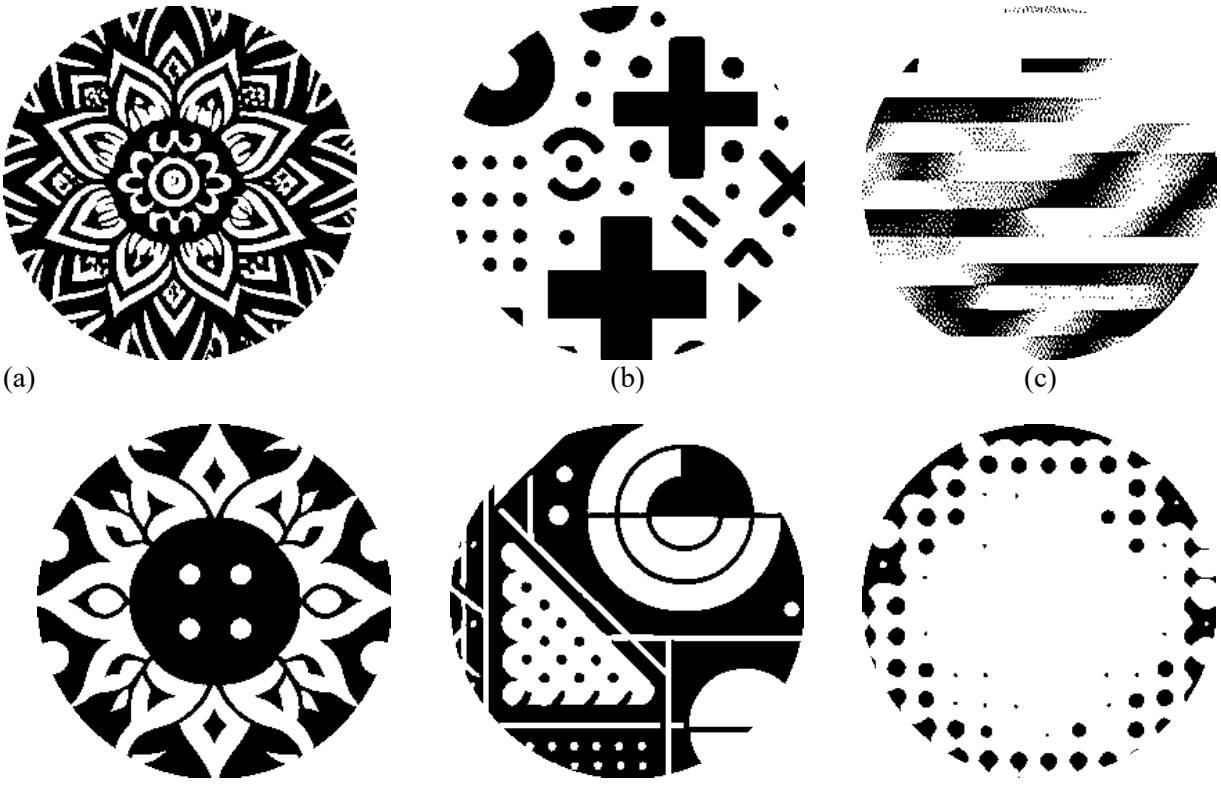


Figure 1: Six illustrative avatars from the set of 110 avatars created and evaluated. People preferred some of these avatars more than others; the order of preference we ascertained for these six was a/d/b/e/c/f, reflecting "CurmElo" scores of 1444/1294/1001/997/668/656.

After DALL-E created these patterns, we used a script to cut a single circle from each pattern. Some avatars were manually discarded for being indistinguishable, resulting in 110 candidate avatars. Some of the avatars' patterns are off-centered; and we hypothesized that this lack of symmetry could affect preferences. Furthermore, the avatars created have different styles – from flowery to geometric to abstract – and we hypothesized that those differences could possibly influence preferences as well.

After creating this corpus of avatars, we implemented a survey to identify preference-indifferent identifiers. We gave 360 online subjects forced-choice questions (in 2024) about which of two avatars they prefer. Every avatar in the set was compared.

As an attention check, we also included forced-choice comparisons between nonsense words from prior work (Sankaran et al., 2021). These questions were used to ensure that individuals were actually considering their preferences when faced with each forced-choice decision instead of simply answering randomly. The CurmElo scores of these nonsense words are known from the earlier work, and comparisons were always between top and bottom rated words from that study, where the top words should be chosen over the bottom words over 80% of the time. If too many low-rated words were selected during the quiz in the present study, then that individual's responses were removed from the data.

During the survey, each worker answered 60 questions in total: 30 unique avatar comparisons; 5 avatar comparisons that were repeated four times each; and 10 word comparisons for validation. The workers are paid $3.50 for roughly 5 minutes of work. Once the survey results were collected, the CurmElo algorithm was employed to find the preference-indifferent set of avatars. See Appendix 2 for a detailed description of the survey.

Measuring Preferences

The Elo algorithm

One way to assess preferences regarding identifiers is an algorithm called CurmElo (Sankaran et al., 2021). This algorithm extends the traditional Elo algorithm (Elo, 1978) to take into account polarization and the transitivity of preferences. Simply calculating the Elo scores and naively using the identifiers that score close together could include avatars that are polarizing, strongly preferred by some and strongly disfavored by others, instead of being truly preference-indifferent. CurmElo also identifies sets of objects that break transitivity, i.e., when object preferences create cycles instead of being well-ordered. Therefore, the CurmElo algorithm is the integration of three separate measures: the Elo ranking system, an estimate of polarization, and a transitivity break estimate.

In the Elo algorithm (Elo, 1978), objects have ratings which are updated as head-to-head matches are decided. If objects *a* and *b* are matched and object *a* wins, the ratings are updated as follows:

$$R'_a = R_a + k\frac{1}{1 + e^{\frac{R_a - R_b}{R_D}}}$$

$$R'_b = R_b - k\frac{1}{1 + e^{\frac{R_a - R_b}{R_D}}}$$

After a large number of matches, the objects (in our case, the avatars) that are the most preferred have the highest resulting Elo scores, and the avatars with the lowest scores are the least preferred. In this setting, k and R_D are free parameters used to tune how sensitive the ratings adjustment is to the results of new matches, and R_0 is the initial rating for all objects. We used the following arbitrary parameters: $k = 20$, $R_0 = 1000$, $R_D = 400$.

Quintile Polarization Estimator

After we have a ranking of avatars from the Elo scores, we divide them into five quintiles $Q_1, \ldots Q_5$. Quintiles with higher integer subscript values contain lower rated objects. The Quintile Polarization Estimator measures how often avatars are ranked "incorrectly," that is, counting the number of times an avatar is preferred over avatars in higher quintiles and rejected versus avatars in lower quintiles.

Formally, let W_{iQ_j} represent the rate at which avatar *i* is chosen compared against avatars in quintile Q_j, W_{iQ_jk}represent the kth (out of n) comparison between *i* and an avatar in Q_j being 1 if avatar *i* is chosen, and 0 otherwise. Using the following equations, we can compute the polarization score for avatar *i*, which ranges from 0 to 1, where higher polarization scores indicate less coherent preferences for the avatar. Avatars with larger polarization scores P_i should be discarded when choosing a preference-indifferent set.

$$\underline{W}_{iQ_j} = \frac{1}{n}\sum_{k=1}^{n} W_{iQ_jk}$$

$$N = \sum_{l=1}^{5}(l-1)$$

$$P_i = \frac{1}{N}\sum_{j\leq 5}\sum_{k:\, j<k\leq 5} 1_{(\underline{W}_{iQ_j} - \underline{W}_{iQ_k} > 0)}$$

where the indicator function $1(\cdot)$ equals one if the statement inside is true and zero otherwise.

Quintile transitivity breaks estimator

By design, our survey does not need to ensure well-orderedness, given the way that humans make preference decisions. It is possible that there are three avatars *a, b,* and *c* where *a* is preferred to *b, b* is preferred to *c*, and *c* is preferred to *a*, creating a cycle and making it difficult to order the three avatars. The transitivity break estimator counts how often each avatar creates a cycle with an avatar in a higher quintile and an avatar in a lower quintile.

Let $W_{Q_iQ_j}$ represent the rate at which avatars in Q_i are chosen when compared against avatars in Q_j. $W_{Q_iQ_jk}$ represents the kth comparison from the data and is 1 if the avatar in Q_i is chosen and 0 otherwise. Avatars with high transitivity breaks scores T_i should also be discarded for not being well-ordered.

$$\underline{W}_{Q_iQ_j} = \frac{1}{n}\sum_{k=1}^{n} W_{Q_iQ_jk}$$

$$N = \sum_{l=1}^{5}(l-1)$$

$$T_i = \frac{1}{N}\sum_{j\leq 5}\sum_{k:\, j<k\leq 5} 1_{(\underline{W}_{Q_iQ_j} < 0.5)\wedge(\underline{W}_{Q_iQ_k} \geq 0.5)}$$

Preference Consistency

We also assessed the consistency of individuals and the consistency induced by certain avatars. In order to learn about the reliability of individuals in their rankings, we included repeated comparisons. Each participant had five comparisons that were repeated four times each throughout the survey. Analyzing these responses, every individual was given a weighted consistency score based on the number of times they made the same choice when faced with the same decision. The weights are based on the difference in Elo scores of the two avatars being compared; it should be easier to remain consistent

between avatars with wildly different preference scores.

Let the weighted consistency score for each individual i be given by, WCS_i.

$$WCS_i = \frac{\sum_{c=1}^{5} f_c(R_a - R_b)}{20}$$

When two different avatars a and b are compared repeatedly, f_c is the number of times the preferred avatar is chosen. Therefore $f_c \in \{2,3,4\}$ since each of the repeated comparisons is repeated 4 times. Each avatar in the comparison, c, has Elo scores R_a and R_b respectively.

Each avatar was included in the repeated comparisons for at least 42 of the 360 participants. In order to learn more about which avatars induce consistency, we calculate win rate consistency and average flip rate. For each avatar, the Win Rate is the percentage of comparison where that avatar is selected.

$$Win\ Rate_i = \frac{Total\ number\ of\ wins}{Total\ number\ of\ comparisons}$$

The average flip rate for each avatar is the average number of times that an avatar is involved in a comparison where the individual changes their preferred avatar. P_i is the set of all comparisons in which avatar i appears. x_{ij} is the preferred avatar from the comparison between avatars i and j.

$$AFR_i = \frac{1}{|P_i|} \sum_{(i,j) \in P_i} 1({x_{ij}}^{(1)} \neq {x_{ij}}^{(2)})$$

The superscripts (1) and (2) indicate two successive comparisons in the sequence.

A perfectly consistent avatar would have a flip rate of zero while a completely random avatar would expect to have a flip rate of 0.5. Note that this definition is also dependent on the order in which comparisons are made. If two avatars, A and B, are being compared, and they were chosen in the order AAABB, the flip rate for A would be 0.2, but if the order was ABABA, the flip rate would be 0.8. This is due to the fact that the average flip rate is based off of the preceding forced choice decision made. Unlike with the Win Rate where order is ignored, the change in preference over time is the exact artifact we want to measure with the average flip rate. Therefore, order matters here.

Avatar Features and Preferences

Finally, we assessed whether any potentially discernable features of the AI-generated avatars could induce preferences. The features we examined included: vertical symmetry, horizontal symmetry, complexity, and darkness. For vertical and horizontal symmetry, a percentage of the number of pixels that mirror that of the other half of the image after the prescribed dissection was calculated. We define complexity to be the density of edges, which we measure using Canny (Canny, 2017). Darkness is simply the ratio of black versus white pixels.

We also considered the different styles of the avatars; for example, avatars *a* and *d* in Figure 1 are vaguely floral, while avatars *b* and *e* are made of classic geometric shapes. However, our attempts to automate/quantify these styles were fairly arbitrary and they did not yield any statistically significant results. See the appendix for a breakdown of our regression model which accounts for three styles: modern, minimalist, and flowery.

Results

Identifying Preference-Indifferent Avatars

Applying the Elo algorithm to the survey responses, we can obtain a score for each avatar. Figure 2 (a) shows a histogram of all of the avatars' Elo score rankings. As expected, the mean is equal to $R_0 = 1000$

(rounded to the nearest integer). With a standard deviation of 148, there is significant variation from the mean, which means that the avatars are not equally preferred.

Figures 2 (b) and (c) show histograms of the number of identifiers for each Polarization score and Transitivity Breaks Score. While moderate values are to be expected for both the polarization and transitivity break scores due to random noise, avatars with high values in one or both metrics should be discarded for being poorly behaved in pairwise comparisons. We used thresholds of 0.3 for the polarization score and 0.2 for the transitivity score

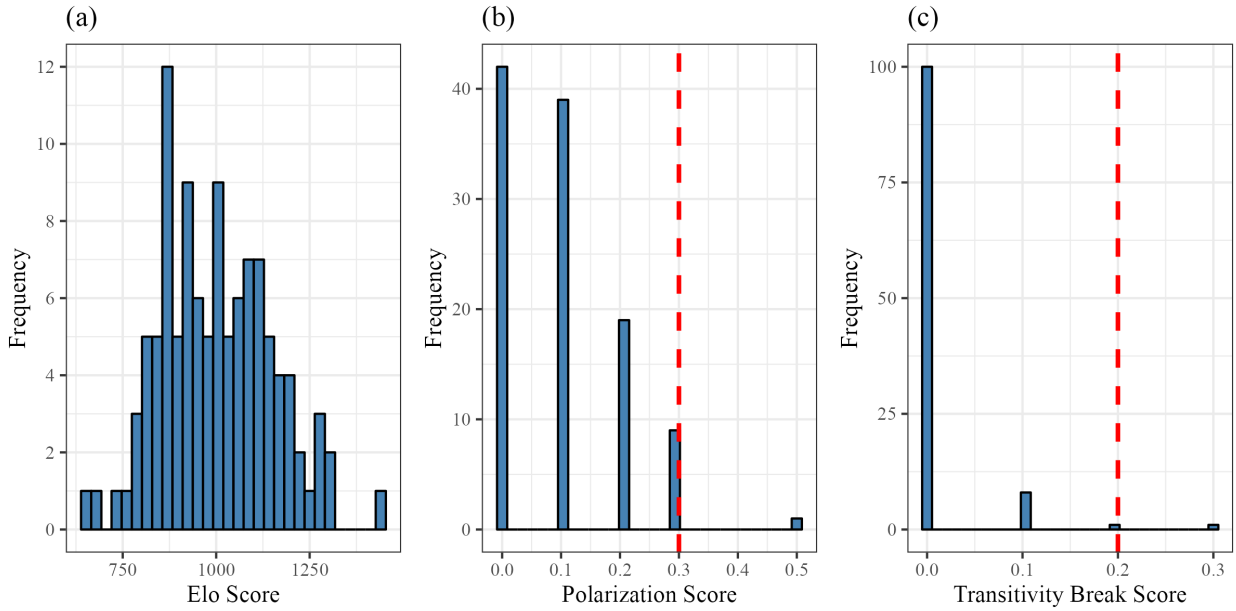
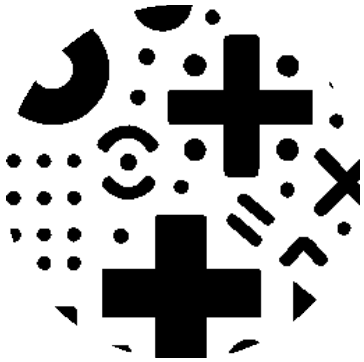


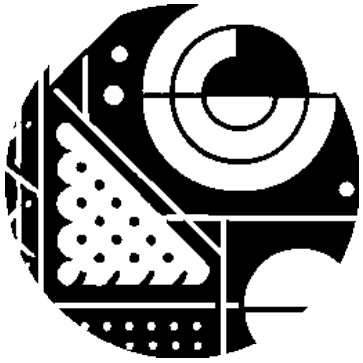
Figure 2: Distribution of Elo scores for 100 candidate avatars from subject rankings and estimates of their polarization and transitivity scores. This figure includes: (a) The distribution of Elo Scores, (b) The distribution of Polarization Scores, (c) The distribution of Transitivity Break Scores. Note the differing x and y axis scales in this figure.

The full corpus of avatars, together with Elo rankings, quintile polarization scores, and transitivity breaks scores can be found at Appendix 3-6. With this information, one can identify a set of avatars that are preference-indifferent. For example, an experiment needing four distinct group identities could use the avatars in Figure 3. All four avatars have nearly identical Elo rankings and have polarization/transitivity scores below 0.2. These four avatars are also near the mean Elo score, indicating that they are of moderate desirability, but, to be clear, a preference-indifferent set of avatars could also consist of highly (or lowly) rated avatars as well.

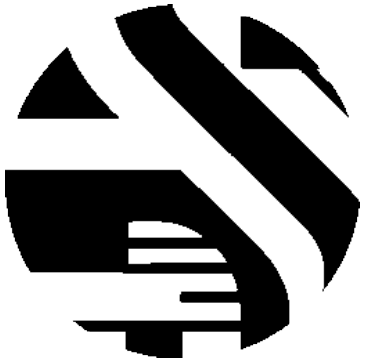
(a) Avatar 10; Elo Score: 1003



(b) Avatar 100; Elo Score: 1001



(c) Avatar 108; Elo Score: 997



(d) Avatar 38; Elo Score: 997

Figure 3: A subset of 4 neutral and preference-indifferent identifiers, selected out of the 100 avatars based on our implemented procedure.

Preference Consistency

A win rate consistency score was calculated for each avatar and can provide a richer understanding of the response from participants it invokes, especially when paired with an avatar's average flip rate. An avatar with a high win rate and low average flip rate is frequently preferred and has stable pairwise preferences. Such an avatar induces strong, consistent preferences. An avatar with a high win rate and high flip rate frequently wins, but there are uncertain or context-dependent preferences. An avatar with a low win rate and low flip rate induces durable but limited preferences. An avatar with a low win rate and high flip rate induces weak or unstable preferences.

The distribution of the average flip rate is shown in Figure 4. All of the average flip rates are below 0.3 which means that, on average, the preferred avatars continue to be preferred at least 70 percent of the time. The mean of the average flip rates is 0.179, so preferences regarding the avatars seem to be stable.

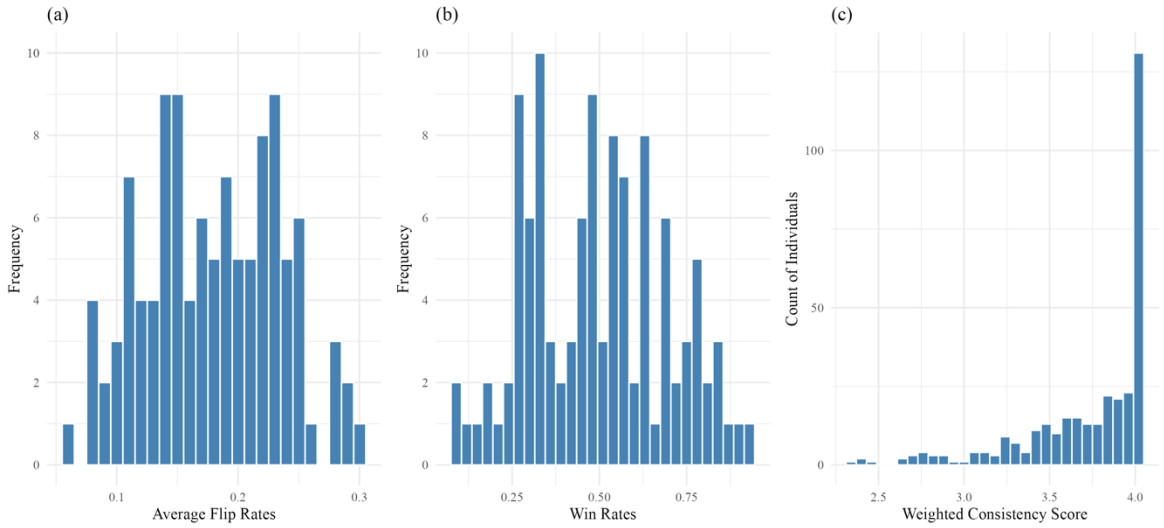


Figure 4: This figure includes: (a) Distribution of Average Flip Rates for the Avatars, (b) Distribution of Win Rates for the Avatars, (c) Distribution of weighted consistency scores for 360 subjects

The distribution of the Win Rate Consistency is shown in Figure 4 (b). This distribution is largely symmetric around 0.5, so our corpus of randomly generated avatars has an even divide of desirable and undesirable avatars. No avatars have win rates above 0.922 or below 0.1. Intriguingly, the distribution appears to be almost normal apart from spikes at the 0.26 and 0.28. These spikes could be due to the repeated comparisons. There are comparisons that are made 4 times; if the losing avatar consistently wins at least one of the repeated comparisons, the win rate would approach values around 0.25; the win rate for this losing avatar would be much lower if not for the repeated comparison. Finally, the distribution of the weighted consistency scores for each individual is shown in Figure 4 (c).

Each of the participants was asked five comparisons 4 different. Most individuals were either perfectly consistent, choosing the same winning avatar all four times, or moderately consistent, choosing the same winning avatar three out of the four times. Figure 5 shows that the smaller the difference in Elo score, the more inconsistent individuals are in repeated comparisons while the greater the Elo score difference the more consistent individuals tend to be.

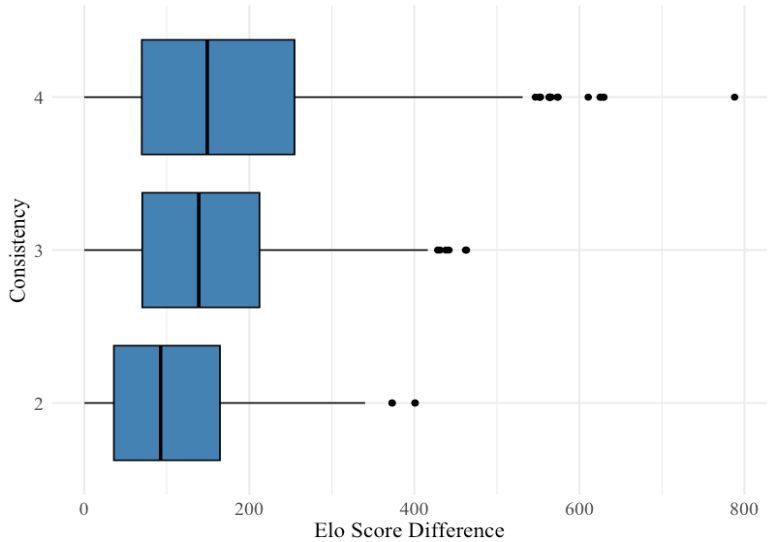


Figure 5: Difference in Elo Score and Repeated responses box and whisker plot. Each data point represents an individual comparison between two avatars, and the dots indicate outliers. The y-axis is the difference in Elo score of the two avatars that are being compared. The x-axis is how consistently the individual reported the same preferred avatar in repeated comparisons.

The Elo score rankings for each type of participant differs. Hence, perfectly consistent, moderately consistent, and inconsistent individuals each rank the avatars differently. If the inconsistent individuals, who are in the minority, are excluded, then the Elo rankings of the avatars are mostly unchanged. However, all individuals included in this distribution did successfully select the high-scoring words over the low-scoring words at least 65% of the time in our validation test.

Avatar Attributes

While we were primarily motivated by the objective of identifying preference-indifferent avatars, our data can also be used to study human preferences across avatar features. A simple linear regression on our five avatar properties (vertical symmetry, horizontal symmetry, complexity, darkness, and style) shows that complexity and the percentage of dark pixels are associated with statistically significant increases in the Elo Score (Figure 6). Avatars that are darker and have more edges seem to be more visually appealing to our participants.

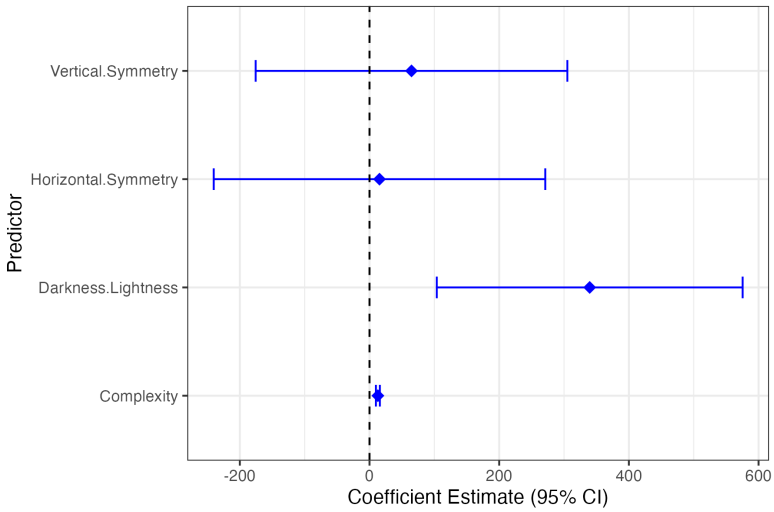


Figure 6: Forest plot of linear regression where the dependent variable is the avatar's Elo Score and the independent variables are the avatar features, including vertical symmetry, horizontal symmetry, complexity, and darkness.

Conclusions

People have intrinsic, albeit variable, preferences for particular numbers, letters, colors, animals, and so on. For instance, people prefer letters in their own name over others (Callahan & Ledgerwood, 2016; Heywood, 1972; Nuttin, 1985). The existence of such preferences is also demonstrated by the passwords people choose (Bonneau, 2012; Riddle, 1989). Human raters even impute category information to nonsense words in systematic ways (Lupyan and Casasanto, 2015). Such preferences extend not just to nonsense words, but to any identifier that might be used, including images, sounds, physical objects, colors, and so on.

Such preferences have material implications for the use of identifiers in experimental social science where people are assigned labels so others in their group can track them during anonymous interactions. For instance, group identity influences behavior, it is also true that behavior influences inter-personal connections and therefore group identity, for instance via homophily (Fu et al., 2012; McPherson et al., 2001). Identifier preference, if not accounted for, might significantly skew results by heterogeneously changing effect size on a per-identifier basis, as well as make replication difficult due to cross-population preference heterogeneity. Indeed, identifier preference may be unacknowledged confounder in a large number of experiments.

In sum, we rigorously evaluate and identify a set of AI-generated avatars toward which subjects show near preference indifference. Unlike other identifiers like colors or shapes, these avatars should minimize any noise from individuals' preferences regarding the identifiers. Though these avatars may not completely eliminate bias for all subjects, they are designed to minimize confounding effects associated with the use of avatars in experiments involving groupings. Future work can systematically examine cross-cultural and cross-population preferences in the design of group identifiers.

References

Bloom, Paul. 2013. *Just Babies: The Origins of Good and Evil*. New York: Crown.

Bonneau, Joseph. 2012. "The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords." In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (SP)*, 538–52. IEEE. https://doi.org/10.1109/SP.2012.49.

Brewer, Marilynn B. 1979. "In-group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis." *Psychological Bulletin* 86 (2): 307.

Callahan, Sean P., and Amber Ledgerwood. 2016. "On the Psychological Function of Flags and Logos: Group Identity Symbols Increase Perceived Entitativity." *Journal of Personality and Social Psychology* 110 (4): 528–50. https://doi.org/10.1037/pspi0000047.

Canny. 2017. *Canny*. Version 1.0. https://canny.io.

Castano, Emanuele, Vincent Yzerbyt, David Bourguignon, and Eléonore Seron. "Who may enter? The impact of in-group identification on in-group/out-group categorization." *Journal of experimental social psychology* 38, no. 3 (2002): 315-322.

Christakis, Nicholas A. 2019. *Blueprint: The Evolutionary Origins of a Good Society.* New York: Little, Brown Spark.

Currarini, Sergio, and Friederike Mengel. 2016. "Identity, Homophily and In-group Bias." *European Economic Review* 90: 40–55.

Diehl, Michael. 1990. "The Minimal Group Paradigm: Theoretical Explanations and Empirical Findings." *European Review of Social Psychology* 1 (1): 263–92.

Elo, Arpad E. 1978. *The Rating of Chessplayers, Past and Present*. New York: Arco Publishing.

Emmerich, Katharina, Alexander Krekhov, and Jürgen H. Krüger. 2024. "Playing for a Better Future: How Animal Avatars Influence the Attitude of Players." *Proceedings of the ACM on Human-Computer Interaction* 8 (CHI PLAY): 1–25. https://doi.org/10.1145/3677096.

Fu, Feng, Martin A. Nowak, Nicholas A. Christakis, and James H. Fowler. 2012. "The Evolution of Homophily." *Scientific Reports* 2 (1): 845.

Guilford, Joy Paul, and Patricia C. Smith. 1959. "A System of Color-Preferences." *The American Journal of Psychology* 72 (4): 487–502.

Hartstone, Margaret, and Martha Augoustinos. 1995. "The Minimal Group Paradigm: Categorization into Two versus Three Groups." *European Journal of Social Psychology* 25 (2): 179–93.

Heywood, Stephen. 1972. "The Popular Number Seven or Number Preference." *Perceptual and Motor Skills* 34 (2): 357–58. https://doi.org/10.2466/pms.1972.34.2.357.

Hong, Youngki, and Kyle G. Ratner. "Minimal but not meaningless: Seemingly arbitrary category labels can imply more than group membership." *Journal of Personality and Social Psychology* 120, no. 3 (2021): 576.

Jackson, Christopher Michael, Joshua Conrad Jackson, David Bilkey, Jonathan Jong, and Jamin Halberstadt. 2019. "The Dynamic Emergence of Minimal Groups." *Group Processes & Intergroup Relations* 22 (7): 921–29.

Karp, David, Nobuhito Jin, Toshio Yamagishi, and Hiromi Shinotsuka. 1993. "Raising the Minimum in the Minimal Group Paradigm." *The Japanese Journal of Experimental Social Psychology* 32 (3): 231–40.

Lupyan, Gary, and Daniel Casasanto. 2015. "Meaningless Words Promote Meaningful Categorization." *Language and Cognition* 7 (2): 167–93. https://doi.org/10.1017/langcog.2014.21.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27 (1): 415–44.

Nuttin, J. M., Jr. 1985. "Narcissism beyond Gestalt and Awareness: The Name-Letter Effect." *European Journal of Social Psychology* 15 (3): 353–61. https://doi.org/10.1002/ejsp.2420150309.

Park, Youngsoon, and Denise A. Guerin. 2002. "Meaning and Preference of Interior Color Palettes among Four Cultures." *Journal of Interior Design* 28 (1): 27–39.

Riddle, William. 1989. "Passwords in Use in a University Timesharing Environment." *Computers & Security* 8 (4): 497–505. https://doi.org/10.1016/0167-4048(89)90049-7.

Sankaran, Soham, Jacob Derechin, and Nicholas A. Christakis. 2021. "CurmElo: The Theory and Practice of a Forced-Choice Approach to Producing Preference Rankings." *PLOS ONE* 16 (5): e0252145. https://doi.org/10.1371/journal.pone.0252145.

Schloss, Karen B., Eli D. Strauss, and Stephen E. Palmer. 2013. “Object Color Preferences.” *Color Research & Application* 38 (6): 393–411.

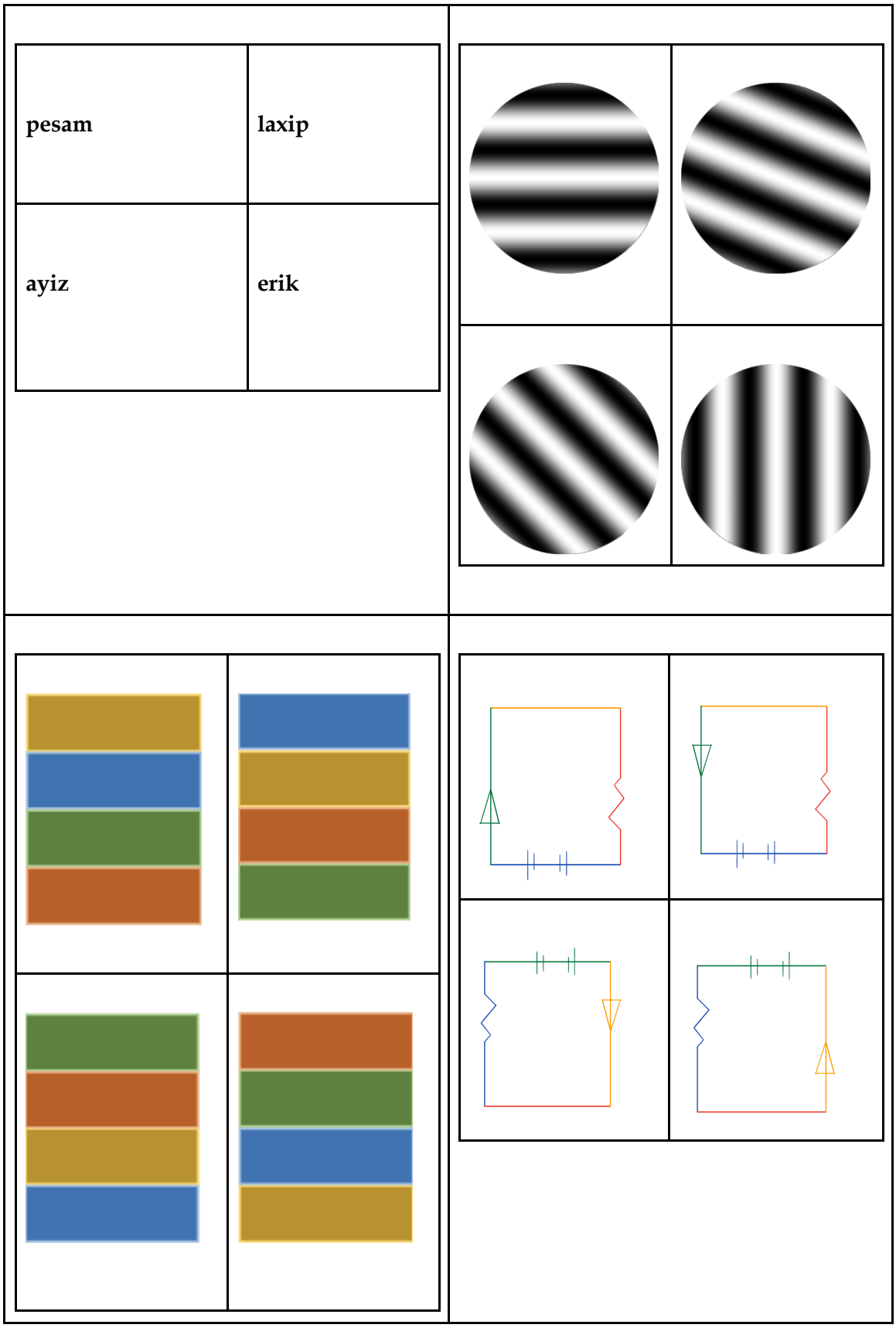
Tajfel, Henri, Michael G. Billig, Robert P. Bundy, and Claude Flament. 1971. “Social Categorization and Intergroup Behaviour.” *European Journal of Social Psychology* 1 (2): 149–78. https://doi.org/10.1002/ejsp.2420010202.

Yamagishi, Toshio, Nobuhito Jin, and Allan S. Miller. 1998. “In-Group Bias and Culture of Collectivism.” *Asian Journal of Social Psychology* 1 (3): 315–28.

Supporting Information

S1 Appendix. Previous avatar attempts.

This figure shows the progression of attempts to generate a set of preference indifferent avatars. The CurmElo paper examined four and five letter nonsense words as group identifiers. The first attempt at visual avatars included black and white lines of differing thickness and angle. The next attempt used primary colors in different orders. Another attempt used circuits with differing colors and arrow orientations. All of these avatars induced strong preferences and were thus unsuitable for our purposes.



S2 Appendix. Description of the AMT survey.

The survey can be viewed at https://comparisonsurvey.pages.dev/.

Participants were directed to the survey after accepting the task on AMT. Participants were collected in several waves and typically had a few days to complete the survey after accepting through AMT.
After consenting to the experiment, participants were given the 60 comparisons, one at a time. Avatars were placed side by side, and it was random which avatar was on the left and which on the right.
Shortly after completing the survey, participants would receive their payment through AMT.

S3 Appendix. Summary Statistics for Elo Ratings.

Statistic	Value
Mean	999.53
Standard Deviation	148.03
Minimum	656.00
Maximum	1444.00

S4 Appendix. Summary Statistics for Polarization.

Statistic	Value
Mean	0.10
Standard Deviation	0.10
Minimum	0.00
Maximum	0.50

S5 Appendix. Summary Statistics for Transitivity Breaks.

Statistic	Value
Mean	0.01
Standard Deviation	0.04
Minimum	0.00
Maximum	0.30

S6 Appendix. Full corpus of avatars.
This attachment includes all of the 110 avatars created.
(PNG)

S7 Appendix. Elo rankings for all avatars.
This attachment includes the Elo score for each of the 110 avatars.
(CSV)

S8 Appendix. Polarization scores for all avatars.
This attachment includes the Polarization score for each of the 110 avatars.
(CSV)

S9 Appendix. Transitivity break scores for all avatars.
This attachment includes the Transitivity break score for each of the 110 avatars.
(CSV)

S10 Appendix. Elo Score Linear Regression.
The linear regression shown in Figure 6.

lm(formula = Elo_Scores ~ Vertical.Symmetry + Horizontal.Symmetry +

Complexity + Darkness.Lightness, data = avatar_attributes1)

Residuals:

Min	1Q	Median	3Q	Max
-429.18	-73.34	-6.61	73.18	261.47

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	639.145	104.758	6.101	1.79e-08 ***
Vertical.Symmetry	64.768	121.219	0.534	0.59426
Horizontal.Symmetry	15.405	128.944	0.119	0.90513
Complexity	12.901	1.537	8.396	2.35e-13 ***
Darkness.Lightness	339.589	118.946	2.855	0.00519 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.9 on 105 degrees of freedom
Multiple R-squared: 0.4199, Adjusted R-squared: 0.3978
F-statistic: 19 on 4 and 105 DF, p-value: 8.831e-12

S10 Appendix. Win rate consistency score for each avatar.
This attachment includes the Win rate consistency score for each of the 110 avatars.
(CSV)

S11 Appendix. Weighted consistency score for each individual in the survey.
This attachment includes the weighted consistency score for each of the 360 participants in the survey.
(CSV)

S12 Appendix. Average flip rate for each avatar.
This attachment includes the Average flip rate for each of the 110 avatars.
(CSV)

S13 Appendix. Elo regression model. Regression including the style variable on Elo scores.
(PDF)

S14 Appendix. Data.
This attachment includes the forced choice responses from the survey.
(Zip)

This study was approved by the Yale University Committee on the Use of Human Subjects and all experiments were performed in accordance with all relevant guidelines and regulations. All subjects gave their informed consent in accordance with the Yale University IRB before participating in the study.