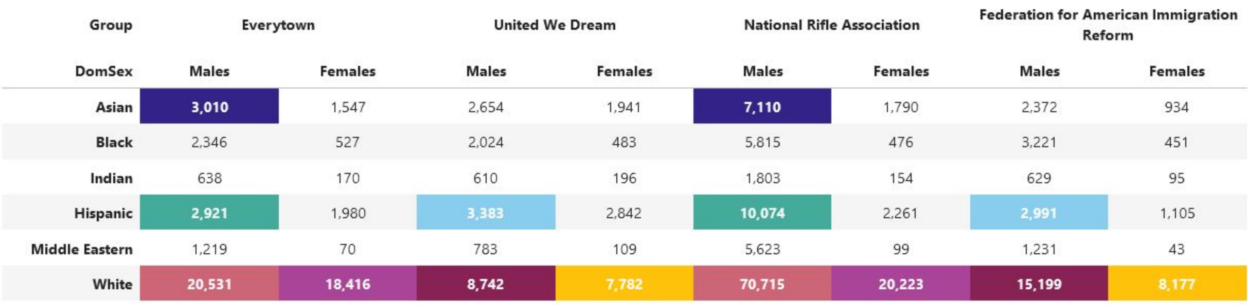
Chapter 9: Results

This chapter is divided into two sections, broadly. The first section examines the control variables that were gleaned from the images and attempts to elucidate differences in these variables across groups using traditional inferential statistics. Although these findings are not the primary focus of this research, identifying differences across subgroups should be illuminating from a theoretical perspective while giving the reader an overall impression of the data. These analyses were performed on the entire corpus of images.

Section two will utilize a machine learning approach to determine the role that facial morphology plays in group categorization. To eliminate the possibility of a classifier categorizing subjects based on their race or gender, all groups of analysis were constrained by these two variables for this section. In other words, the overall sample was broken into subsamples, each specific to a particular group membership, as well as the sex and race of the subject. With four different issue groups, two sex categories, and six ethnic-racial categories, this resulted in a total of 48 subgroups (e.g., white female followers of the NRA; Hispanic male followers of Everytown). Given that this approach was predicated on comparisons of different group members with different political leanings but identical demographic characteristics, this implies a maximum of 24 pairwise comparison.

Note that this approach relies upon an ample sample size in order to detect potentially small effects. Using a similar method, Wang and Kosinski (2018) previously relied on a minimum sample size of 3,441. In the present analyses, the intent was to include any subsamples that included at least 3,000 subjects. Because this resulted in only five viable pairwise compairsons, the lower bound was reduced to 2,900 to allow for the inclusion of two additional analyses. These are the sizes for the sample that is unrestricted in terms of pitch and yaw; however, companion analyses were included in the Appendices for the reduced sample set regardless of sample size. As illustrated in Table 1, this expanded the number of viable comparisons to seven.

Table 1
Groups of Analysis



Group	Everytown		United We Dream		National Rifle Association		Federation for American Immigration Reform	
DomSex	Males	Females	Males	Females	Males	Females	Males	Females
Asian	3,010	1,547	2,654	1,941	7,110	1,790	2,372	934
Black	2,346	527	2,024	483	5,815	476	3,221	451
Indian	638	170	610	196	1,803	154	629	95
Hispanic	2,921	1,980	3,383	2,842	10,074	2,261	2,991	1,105
Middle Eastern	1,219	70	783	109	5,623	99	1,231	43
White	20,531	18,416	8,742	7,782	70,715	20,223	15,199	8,177

Note: Fields with identical colors identify the viable comparison between a left-learning and a right-leaning group pertaining to the same issue. Everytown and United We Dream are left-leaning groups; National Rifle Association and the Federation for American Immigration Reform are right-leaning groups.

Because the sample skewed male (as does Twitter in general), the subgroup selection did as well, with five pairwise subgroup comparisons for men and only two such comparisons for women. Only women whom the classifier deemed white could be included in the present analyses.

For the sake of brevity, only one of the subgroups is examined in detail. For this portion of the analysis, the subgroup with the largest sample size was utilized: white males in the gun topic. It is important to understand that this subgroup is not of any particular interest in regards to these analyses. Rather, this subgroup is utilized merely as an illustration for the reader of how analyses were performed on all seven subgroups. However, more important than any specific subgroup comparison, we are looking for overall trends in regards to this methodology. Thus, after the detailed analysis of the white male gun subgroup, broader trends and realizations across all subgroup analyses will be presented. Metrics for all models are presented in the Appendices.

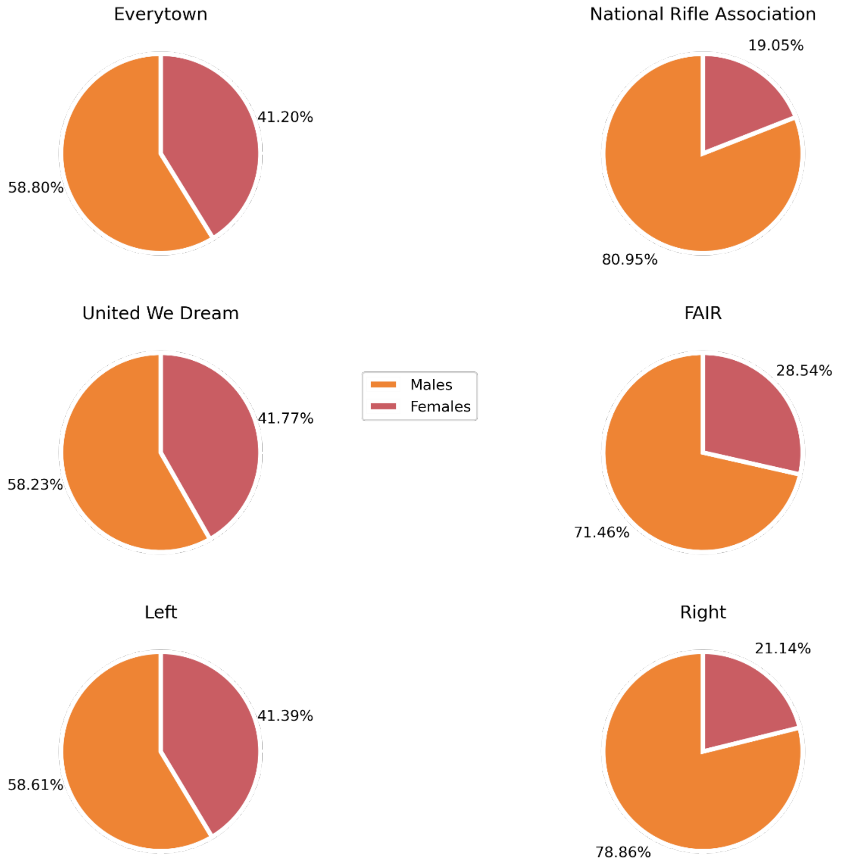
Section 1 – Control Variables

Each of the images in the sample was evaluated by DeepFace, a deep learning facial recognition network that was developed by a team at Facebook. DeepFace utilizes a nine-layer neural network that was trained on four million images from Facebook users (Serengil, n.d.). The network involves over 120 million parameters and describes itself as 'the most lightweight face recognition and facial attribute analysis library for python'(Serengil, n.d.). Using this architechture, attributes for sex, race, age, and emotional expression were extracted from each image. These factors were examined for Section 1, with all images being included in these analyses.

Sex

Research on Twitter demographics has previously shown that Twitter users are majority male (Dixon, 2022, 56.4%; Yildiz et al., 2017, 73%). This sample did not differ in that respect, with the sex identifier classifying nearly 71% of subjects as male. See Figure 3.

Figure 3
Sex by Organization Barplot



Note: FAIR = Federation of American Immigration Reform. Everytown and United We Dream are left-leaning groups; National Rifle Association and FAIR are right-leaning groups.

Several chi-square tests of independence were used to determine whether followers of organizations differed by sex. In comparing all images, right-leaning subjects were significantly more likely to be male than left-leaning subjects (χ^2(4, N = 247,515) = 11,311.91, $p < .001$, Φ = .21). The same held true for both the gun subgroup (χ^2(4, N = 179,518) = 9,925.86, $p < .001$, Φ = .24) and for the immigration subgroup (χ^2(4, N = 67,997) = 1,186.43, $p < .001$, Φ = .13), although the effect was weaker in the immigration subgroup. The phi value of .21 translates to a weak overall effect according to statistical conventions (Bhandari, 2021; Zaiontz, n.d.).

The sex classifier provided a propensity score to classify subjects in images as being male or female with some overall likelihood. Each image was provided a ‘score’ as to whether the individual was male or female, on a scale from zero to one. If the score is closer to zero, the algorithm predicts that the individual in the photo is a woman; if it is closer to one, the algorithm predicts that the individual is a man (or vice versa). In the present context, these categories are mutually exclusive: if the algorithm is predicting the image is a woman at a likelihood of 10%, this necessarily means that it is also predicting the image is a man with a likelihood of 90%. Thus, one can simply subtract the propensity score for women (in this example) from one, providing a proxy score for ‘sex typicality’ for both sexes with a scale from .50 to 1.

Analyzing the data in such a way, we identified if followers of organizations differed in their sex typicality, at least in terms of what the sex classifier deemed as sex typical qualities. Previous research has suggested that conservative leaning individuals might endorse more traditional gender roles, suggesting that males and females that are right-leaning might adopt appearances that are more ‘sex typical’ in nature (Duncan et al., 1997). At the same time, left-leaning individuals tend to score higher on the personality trait openness, to demonstrate a greater propensity to break with tradition, and to espouse for equity across social constructs, suggesting that the ‘borders’ between the sexes might be more traversable for left-leaning individuals (Carney et al., 2008).

A linear model was created using all images with the ‘sex typicality’ variable being independent and sex, orientation, topic, and the interactions of the three being dependent variables. All linear models were examined with a Type-3 analysis of variance test, and all categorical variables utilized 0/1 coding.. The results for this sex typicality ANOVA model were significant ($F(7, 247{,}507) = 1{,}335.00$, $p < .001$). Individuals on the right were significantly more likely to demonstrate 'sex typicality' than individuals on the left ($b = 3.45$, $CI_{95\%} = [3.33, 3.57]$, $t(247{,}507) = 57.11$, $p < .001$). There was also an effect by sex, with males on average across all images demonstrating greater sex typicality than females ($b = -0.72$, $CI_{95\%} = [-.88, -.56]$, $t(247{,}507) = -8.86$, $p < .001$). There was an interaction effect between orientation and sex, with males demonstrating a significant increase in sex typicality while moving from left to right in orientation, but with females being relatively consistent across the ideological gap ($b = 2.86$, $CI_{95\%} = [-3.06, -2.65]$, $t(247{,}507) = -27.39$, $p < .001$). Topic (gun vs. immigration) was non-significant ($b = .029$, $CI_{95\%} = [-.14, .20]$, $t(247{,}507) = .34$, $p = .74$), as was the interaction with topic and sex ($b = 0.050$, $CI_{95\%} = [-.31, .21]$, $t(247{,}507) = -.37$, $p = .71$). The interaction of topic and orientation was significant, however ($\beta = -1.72$, $CI_{95\%} = [-1.93, -1.51]$, $t(247{,}507) = -15.86$, $p < .001$). Left-leaning subgroups did not differ substantially in their sex typicality across the topics of gun control and immigration. Right-leaning subgroups, contrastingly, demonstrated greater sex typicality in the gun control conditions than the immigration conditions. Finally, the three-way interaction between sex, orientation, and topic was also significant ($b = 0.92$, $CI_{95\%} = [0.56, 1.28]$, $t(247{,}507) = 5.03$, $p < .001$), although the effect was rather modest. The R^2 for the model was .036, explaining just over 3.5 percent of the variance in the model. Figure 4a summarizes the pattern of findings for the collapsed data. Figures 4b and 4c display the data separately for gun and immigration subgroups, while 4d demonstrates the interaction between orientation and topic. Results for this analysis are presented in Appendix F.

Figure 4

Sex Typicality Line Plots

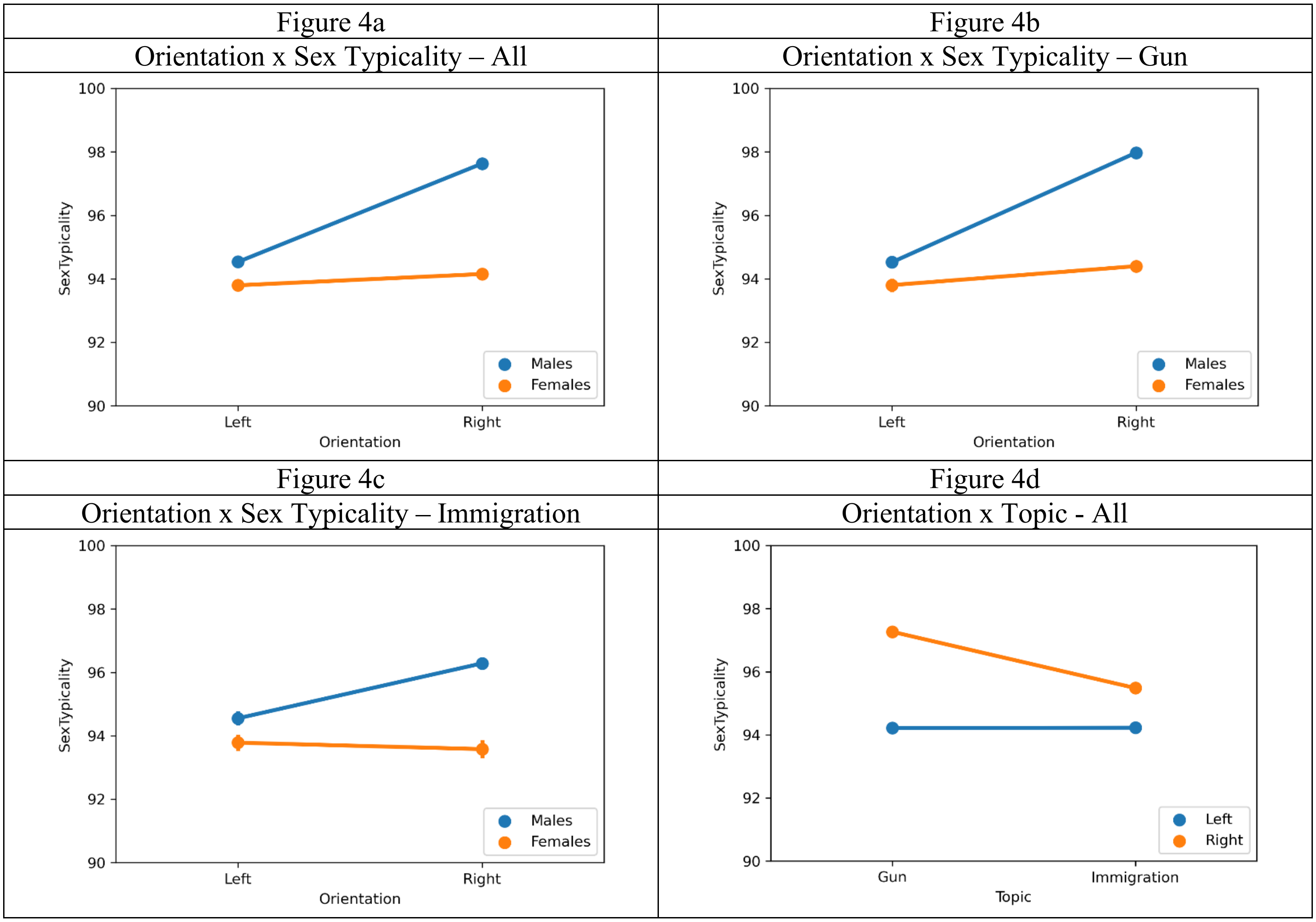


Figure 4a

Orientation x Sex Typicality – All

Figure 4b

Orientation x Sex Typicality – Gun

Figure 4c

Orientation x Sex Typicality – Immigration

Figure 4d

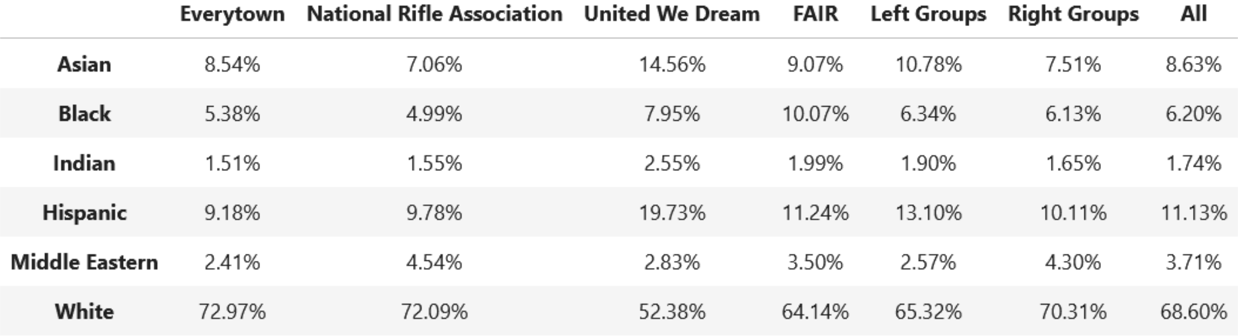
Orientation x Topic - All

Note: All plots contain confidence intervals for each point estimate. Due to the large sample sizes, they are very small and difficult to see.

Race

The python library DeepFace was used to retrieve the racial information for subjects in the images. The sample was majority white, with approximately 69% of the sample being Caucasian. The next largest ethnicity represented was Hispanics, with nearly 11% of the sample. Asians, African-Americans, Middle-Easterners, and Indians comprised the rest of the sample, each comprising 9%, 6%, 4%, and 2% of the sample, respectively (rounded to the nearest integer) (see Table 2).

Table 2
Percent Race by Group



	Everytown	National Rifle Association	United We Dream	FAIR	Left Groups	Right Groups	All
Asian	8.54%	7.06%	14.56%	9.07%	10.78%	7.51%	8.63%
Black	5.38%	4.99%	7.95%	10.07%	6.34%	6.13%	6.20%
Indian	1.51%	1.55%	2.55%	1.99%	1.90%	1.65%	1.74%
Hispanic	9.18%	9.78%	19.73%	11.24%	13.10%	10.11%	11.13%
Middle Eastern	2.41%	4.54%	2.83%	3.50%	2.57%	4.30%	3.71%
White	72.97%	72.09%	52.38%	64.14%	65.32%	70.31%	68.60%

To determine if left- and right-followers differed significantly in their racial composition, two sets of analyses were run. First, a chi-square test was performed on all of the images, with all of the ethnicities represented. Left-leaning subgroups were significantly more diverse than right-leaning subgroups, (χ^2(12, *N* = 247,515) = 1,820.07, *p* < .001, Φ_c = 0.09), although the effect was rather weak. The effect was the same albeit greatly reduced when looking at just the gun subgroups, with Everytown being more ethnically diverse than the NRA (χ^2(12, *N* =179,518) = 571.71, *p* < .001, Φ_c = 0.06). The effect was much stronger for the Immigration subgroups, with the images from United We Dream being significantly more diverse than FAIR and with the strongest effect overall in regards to ethnicity, although overall still relatively weak (χ^2(12, *N* = 67,997) = 1,773.81, *p* < .001, Φ_c = 0.16).

Second, a chi-square test was carried out on a dichotomized ethnicity variable with Caucasians representing one subgroup and members of all other ethnicities representing a non-Caucasian/'minority' subgroup. Results were similar although less pronounced.

When accounting for all images, left-leaning subgroups demonstrated greater racial diversity than right-leaning subgroups (χ^2(4, *N* = 247,515) = 644.41, *p* < .001, Φ = 0.05). For gun subgroups, the test result was significant but the effect was negligible (χ^2(4, *N* = 179,518) = 14.39, *p* < .001, Φ < 0.01). Similar to the previous analysis, the immigration subgroups demonstrated the largest effect (χ^2(4, *N* = 67,997) = 964.00, *p* < .001, Φ < 0.12).

Age

Each of the images was evaluated by DeepFace to provide an age estimation for each participant. Mean predicted age across all images was 31.54 years old (*SD* = 5.48). Research on Twitter users has previously shown that, on average, they are relatively young (Dixon, 2021, 64% ≤ 34). This sample was no different, with 75% of subjects being deemed 34 years of age or lower. The maximum age predicted by the classifier was 62, and the minimum age was 19.

A linear model was used to assess if there were significant differences in age across the images, with sex, orientation, and topic as the independent variables. The model proved significant overall (*F*(3, 247,507) = 841.10, *p* < .001). Females appeared significantly older than males in the sample (*b* = 1.20, $\mathrm{CI}_{95\%}$ = [1.11, 1.29], *t*(247,507) = 25.28, *p* < .001). At the same time, right-leaning individuals appeared significantly older than left-leaning individuals (*b* = 0.20, $\mathrm{CI}_{95\%}$ = [0.13, 0.27], *t*(247,507) = 5.70, *p* < .001). This is perhaps not surprising, as previous research has suggested people become more conservative with age (Truett, 1993). Topic was also significant in the model, with subjects in the immigration subgroups appearing older on average than those in the gun subgroups (*b* = -1.04, $\mathrm{CI}_{95\%}$ = [-1.14, -0.95], *t*(247,507) = -20.63, *p* < .001). There was a significant interaction effect between topic and orientation (*b* = 2.62, $\mathrm{CI}_{95\%}$ = [2.50, 2.74], *t*(247,507)

= 41.43, p < .001), with left-leaning and right-leaning followers being very similar in age for the gun topic but right-leaning subjects being around 2.5 years older on average than left-leaning subjects for the immigration topic. The interaction between sex and orientation approached significance (b = -0.12, $\mathrm{CI}_{95\%}$ = [-0.24, 0.00], t(247,507) = --1.93, p = .054), illustrating a slight decrease in age difference between the sexes when moving from left to right in orientation. The interaction between sex and topic was non-significant, (b = .02, $\mathrm{CI}_{95\%}$ = [-.13, .18], t(247,507) = .29, p = .77), although the interaction between the three independent variables was significant, (β = -.37, $\mathrm{CI}_{95\%}$ = [-.58, -.17], t(247,507) = -3.51, p < .001). The R^2 for the model was .023. See Figures 5a-5d. Results for this model are in Appendix G.

Figure 5

Age Line Plots

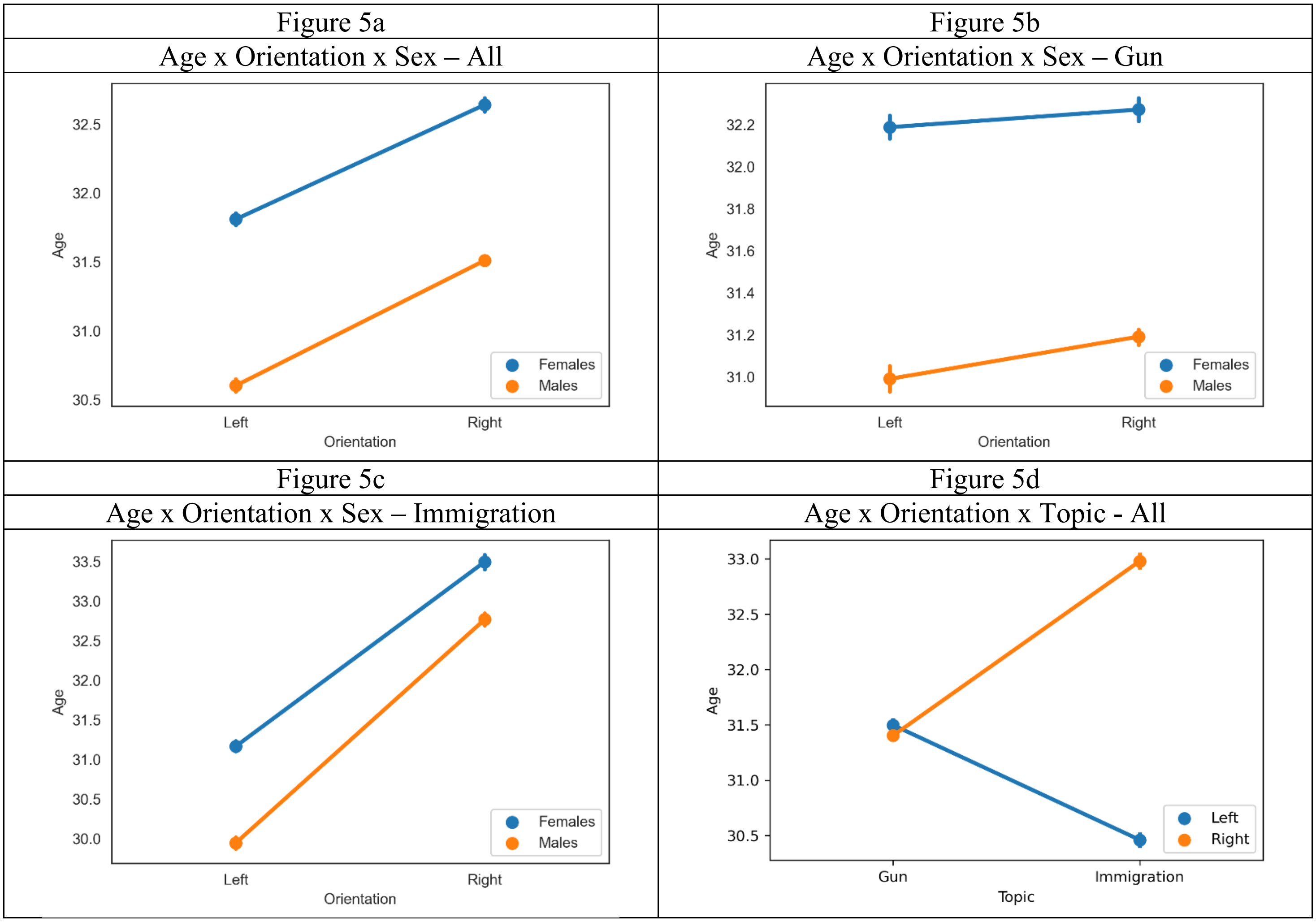


Figure 5a

Age x Orientation x Sex – All

Figure 5b

Age x Orientation x Sex – Gun

Figure 5c

Age x Orientation x Sex – Immigration

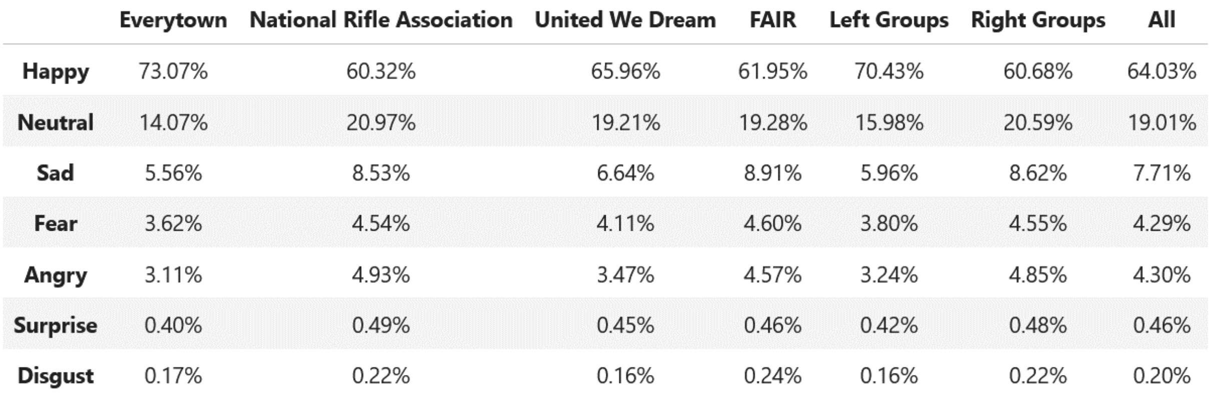
Figure 5d

Age x Orientation x Topic - All

Emotional Expression

Images in the sample were evaluated by DeepFace to provide a measure for subjects' facial expression. Every image was given a propensity score on each of seven emotions: happy, neutral, sad, fear, angry, surprise, and disgust. Over 90% of the sample was covered by just three of those emotions: happy, neutral, and sad. A chi-square goodness of fit test was performed on the sample, testing to see if emotional expression differed by political orientation. It did (χ^2(14, N = 247,515) = 2,384.25, p < .001, Φ = 0.10). See Table 3.

Table 3
Percent Emotional Expression by Organization, Orientation



	Everytown	National Rifle Association	United We Dream	FAIR	Left Groups	Right Groups	All
Happy	73.07%	60.32%	65.96%	61.95%	70.43%	60.68%	64.03%
Neutral	14.07%	20.97%	19.21%	19.28%	15.98%	20.59%	19.01%
Sad	5.56%	8.53%	6.64%	8.91%	5.96%	8.62%	7.71%
Fear	3.62%	4.54%	4.11%	4.60%	3.80%	4.55%	4.29%
Angry	3.11%	4.93%	3.47%	4.57%	3.24%	4.85%	4.30%
Surprise	0.40%	0.49%	0.45%	0.46%	0.42%	0.48%	0.46%
Disgust	0.17%	0.22%	0.16%	0.24%	0.16%	0.22%	0.20%

However, this did not reveal much about the individual emotions at play. Thus, binary variables were created for happy and sad, the two most common facial expressions aside from neutral, to see if followers differed by these emotions. Results were mixed.

Happy. A chi-square test for independence was run on the sample to see if followers differed significantly by proportion of happy subjects. Right-leaning followers had significantly fewer people per capita in their groups with happy expressions than left-leaning followers ($\chi^2(4, N = 247{,}515) = 2{,}300.13$, $p < .001$, $\Phi = 0.10$). Both of the subgroups were tested and both results were significant, although the results for the immigration subgroups were much weaker overall (gun: $\chi^2(4, N = 179{,}518) = 2{,}650.13$, $p < .001$, $\Phi = 0.12$, immigration: $\chi^2(4, N = 67{,}997) = 117.88$, $p < .001$, $\Phi = 0.04$).

Mean happiness propensity scores were also examined. The sample was reduced to just those subjects whom the classifier believed demonstrated a happy expression on their face. Among only these subjects, a linear model was fit to the data with happy propensity scores for the dependent variable and orientation, sex, topic, and the interactions as the independent variables. The model was significant overall ($F(7, 158{,}471) = 307.40$, $p < .001$, $R^2 = .01$). Among those who were happy, orientation was significant, with happiness scores being significantly higher for those on the left than for those on the right ($b = -1.63$, $CI_{95\%} = [-1.81, -1.45]$, $t(158{,}471) = -17.48$, $p < .001$) . At the same time, female subjects were scored as expressing more happiness in their photos than male targets on average ($b = 1.75$, $CI_{95\%} = [1.53, 1.98]$, $t(158{,}471) = 15.09$, $p < .001$). Happy subjects in the immigration topic demonstrated significantly lower happiness scores on average than those in the gun topic, ($b = -.83$, $CI_{95\%} = [-1.09, -0.57]$, $t(158{,}471) = -6.16$, $p < .001$). Both the interactions between orientation and sex as well as orientation and topic were significant, ($b = 1.04$, $CI_{95\%} = [0.75, 1.34]$, $t(158{,}471) = 6.89$, $p < .001$) and ($b = 0.60$, $CI_{95\%} = [.26, .93]$, $t(158{,}471) = 3.47$, $p = .001$), respectively. Females were happier on average than males, and the magnitude of this difference increased when moving from left to right.

Similarly, happy left subjects were happier than their right counterparts, and this difference was much larger across the gun topic than it was across the immigration topic. Neither the interaction between topic and sex (*b* = -0.20, CI$_{95\%}$ = [-0.59, 0.18], *t*(158,471) = -1.03, *p* = .30) nor the three way interaction (*b* = 0.30, CI$_{95\%}$ = [-0.23, 0.83], *t*(158,471) = 1.11, *p* = .27) were significant. See Figures 6a-6d. Model results are presented in Appendix H.

Figure 6
Happy Line Plots

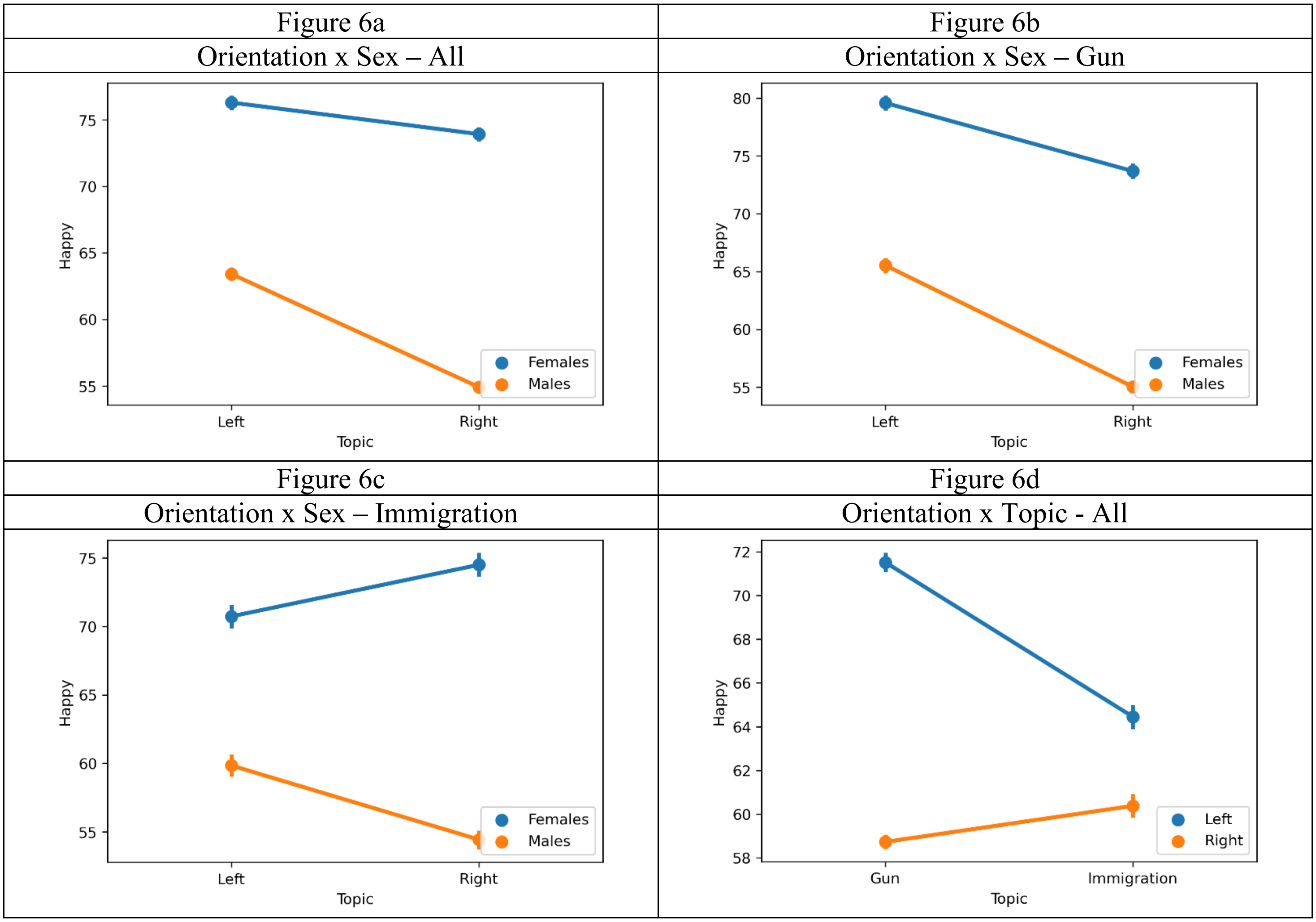


Figure 6a
Orientation x Sex – All
Figure 6b
Orientation x Sex – Gun
Figure 6c
Orientation x Sex – Immigration
Figure 6d
Orientation x Topic - All

Sad. Results were more opaque in regards to the 'sad' emotional expression. A chi-square test for independence found that right leaning followers had proportionally more photos with sad expressions than left leaning followers, $\chi^2(4, N = 313{,}302) = 745.44$, $p < .001$, $\Phi = 0.05$. This effect remained when testing only the gun subgroups ($\chi^2(4, N = 232{,}017) = 648.99$, $p < .001$, $\Phi = 0.05$) and the immigration subgroups ($\chi^2(4, N = 81{,}285) = 133.43$, $p < .001$, $\Phi = 0.04$).

Similar to the 'happy' expression, sad subjects were isolated, and a linear model for sadness was fit with orientation, sex, topic, and their interactions as predictors. The model was significant, although it explained little of the variance ($F(7, 19{,}065) = 2.29$, $p = .03$, $R^2 < .01$). Only orientation was significant in this model, with subjects on the right demonstrating significantly more sadness in their sad pictures than subjects on the left ($b = -1.29$, $\text{CI}_{95\%} = [-2.21, -0.36]$, $t(19{,}065) = -2.72$, $p < .01$). Results for the linear model are presented in Appendix I.

Pitch and Yaw

One criticism of the original work by Wang and Kosinski (2018) was that the researchers did not account for head positioning. Groups of analysis might differ by head positioning, critics argued, in which case the classifier might be predicting based on head position rather than facial morphology itself (Agüera y Arcas et al., 2018). To test for this, all images were subjected to an algorithm that estimated the roll, pitch, and yaw of the head. Because all facial images were cropped and rotated, the roll variable demonstrated little variation across images. Thus, the factors of interest are primarily pitch and yaw.

Pitch. Critics of Wang and Kosinski (2018) describe that heterosexual males and females might orient their heads in different positions when taking their photographs, specifically in the pitch orientation, the axis used when nodding your head up and down in agreement. It is posited that this is due to sexual dimorphism regarding height in human beings and subconscious sexual signaling to potential mates (Sedgewick et al., 2017). If this were true, we would expect female faces to demonstrate a larger pitch number, indicating that women had taken pictures from above with their faces tilted up, while men would have a lower pitch number indicating that their pictures on average were taken from a lower angle.

To test for such a hypothesis, a linear model was created with pitch as the dependent variable and orientation, sex, topic, and their interactions as predictors. The model was significant overall ($F(7, 247{,}507) = 1{,}820.00$, $p < .01$, $R^2 = .05$). Women demonstrated a significantly higher pitch value in comparison to men ($b = 3.90$, $CI_{95\%} = [3.69, 4.10]$, $t(247{,}507) = 37.48$, $p < .001$), as did people on the left in comparison to people on the right ($b = -2.18$, $CI_{95\%} = [-2.33, -2.03]$, $t(247{,}507) = -28.17$, $p < .001$). Topic was also significant ($b = -0.63$, $CI_{95\%} = [-0.85, -0.42]$, $t(247{,}507) = -5.70$, $p < .001$), with people in the gun groups demonstrating a slightly higher pitch than those in the immigration groups. The interaction between orientation and sex was significant ($b = 2.96$, $CI_{95\%} = [2.70, 3.22]$, $t(247{,}507) = 22.18$, $p < .001$). Females demonstrated an increase in pitch angle when moving from left to right in orientation, while males experienced a decrease in pitch angle. Put another way, right-leaning subjects demonstrated greater variation in pitch between the sexes, while left-leaning subjects exhibited less variation. These findings parallel the findings concerning sex typicality, and provide some additional support for increased gender role adoption among right-leaning subjects.

The interaction between orientation and topic was significant ($b = 0.66$, $CI_{95\%} =$

[0.38, 0.93], *t*(247,507) = 4.73, $p < .001$). Whereas participants on the left demonstrated greater mean pitch angles than those on the right, these differences were reduced in the immigration subset in comparison to the gun subset. The final two-way interaction between topic and sex was non-significant, (*b* = -0.12, $CI_{95\%}$ = [-0.46, 0.21], *t*(247,507) = -.71, *p* = .48). However, the three-way interaction between all of the variables did reach significance, (β = -0.68, $CI_{95\%}$ = [-1.13, -0.22], *t*(247,507) = -2.89, $p < .01$). See Figures 7a-7d. Model results are presented in Appendix J.

Figure 7
Pitch Line Plots

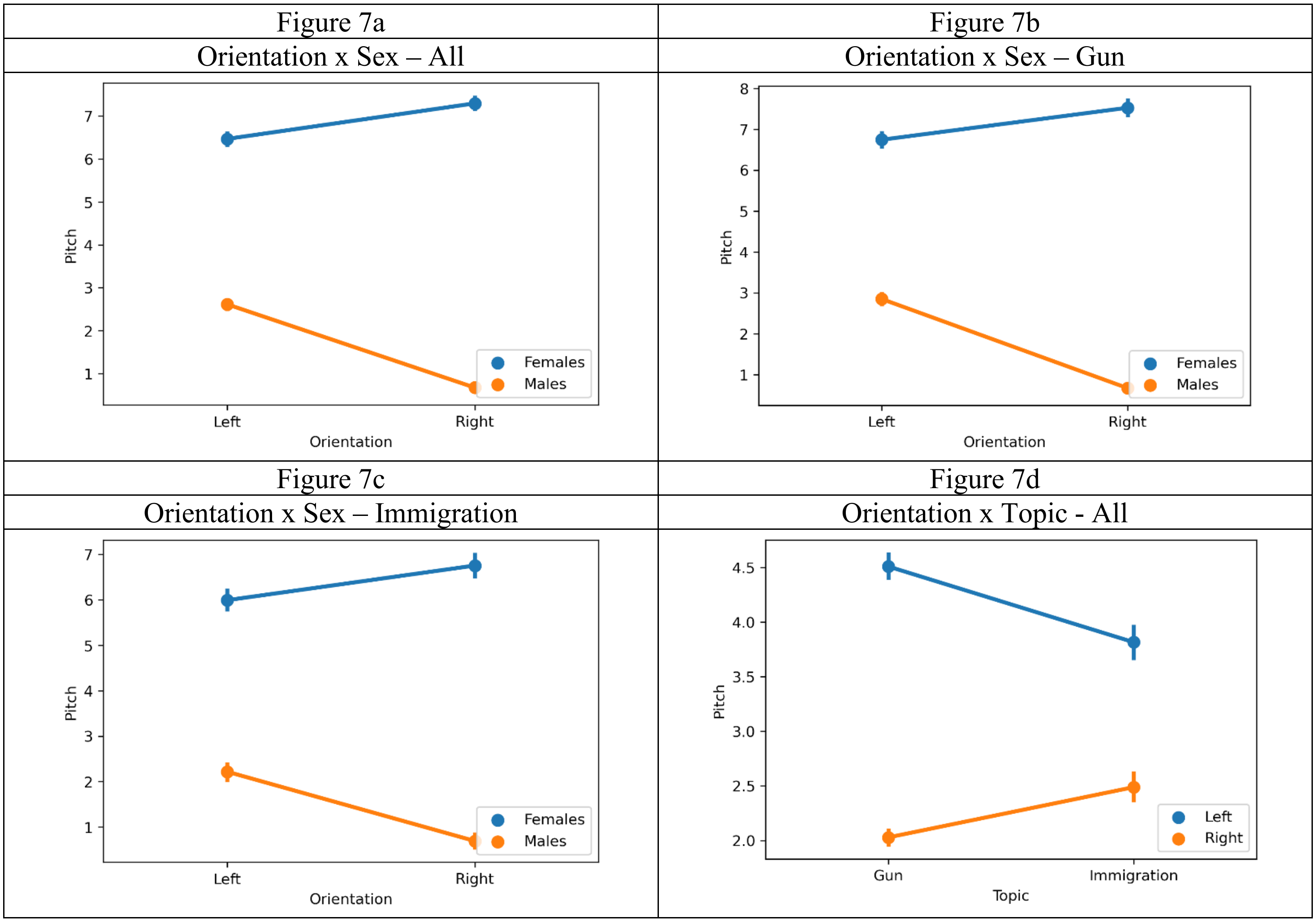


Figure 7a
Orientation x Sex – All

Figure 7b
Orientation x Sex – Gun

Figure 7c
Orientation x Sex – Immigration

Figure 7d
Orientation x Topic - All

Yaw. Although there was no reason to believe that groups differed by the yaw of their head position (shaking your head no in disagreement), a linear model was created to determine if groups differed in this dimension, with sex, orientation, topic, and their interactions as predictors. Recall that positive numbers in the yaw dimension translate to a head position facing towards the subject's right, the viewer's left. Results for the model were significant ($F(7, 247{,}507) = 45.84$, $p < .001$, $R^2 = .001$), but the small amount of variance explaine suggests that yaw does not aid much in classification of images.

There was a main effect for sex ($b = -0.89$, $\mathrm{CI}_{95\%} = [-1.08, -0.71]$, $t(247{,}507) = -9.67$, $p < .001$), with women on average, facing slightly more to their left than men. There was also a main effect for orientation ($b = -0.40$, $\mathrm{CI}_{95\%} = [-0.53, -0.26]$, $t(247{,}507) = -5.79$, $p < .001$), with subjects on the political right facing slightly more to their left than those on the political left. The main effect for topic approached significance ($b = 0.19$, $\mathrm{CI}_{95\%} = [-0.01, 0.38]$, $t(247{,}507) = 1.89$, $p = .06$), with those subjects in the gun groups demonstrating a slightly lower yaw than those in the immigration groups, translating to subjects in the immigration groups facing slightly more to their left than those in the gun groups.

Two interactions were significant. First, there was a significant interaction between sex and orientation ($b = 0.28$, $\mathrm{CI}_{95\%} = [0.04, 0.51]$, $t(247{,}507) = 2.33$, $p = .02$). Females on average had a lower yaw than males, but this difference was reduced among right leaning subjects in comparison to left leaning subjects. Second, there was a significant interaction between sex and topic, ($b = -0.38$, $\mathrm{CI}_{95\%} = [-0.68, -0.08]$, $t(247{,}507) = -2.50$, $p = .01$). Males and females in the immigration groups differed more than strongly than males and females in the gun groups.

Neither the two-way interaction of topic and orientation ($b = 0.01$, $\mathrm{CI}_{95\%} = [-0.23,$

0.25], *t*(247,507) = .07, *p* = .94) nor the three-way interaction of topic, orientation, and gender (*b* = 0.27, $CI_{95\%}$ = [-0.14, 0.68], *t*(247,507) = 1.30, *p* = .19) were significant. See Figures 8a-8d. Results for this model are presented in Appendix K.

Figure 8
Yaw Line Plots

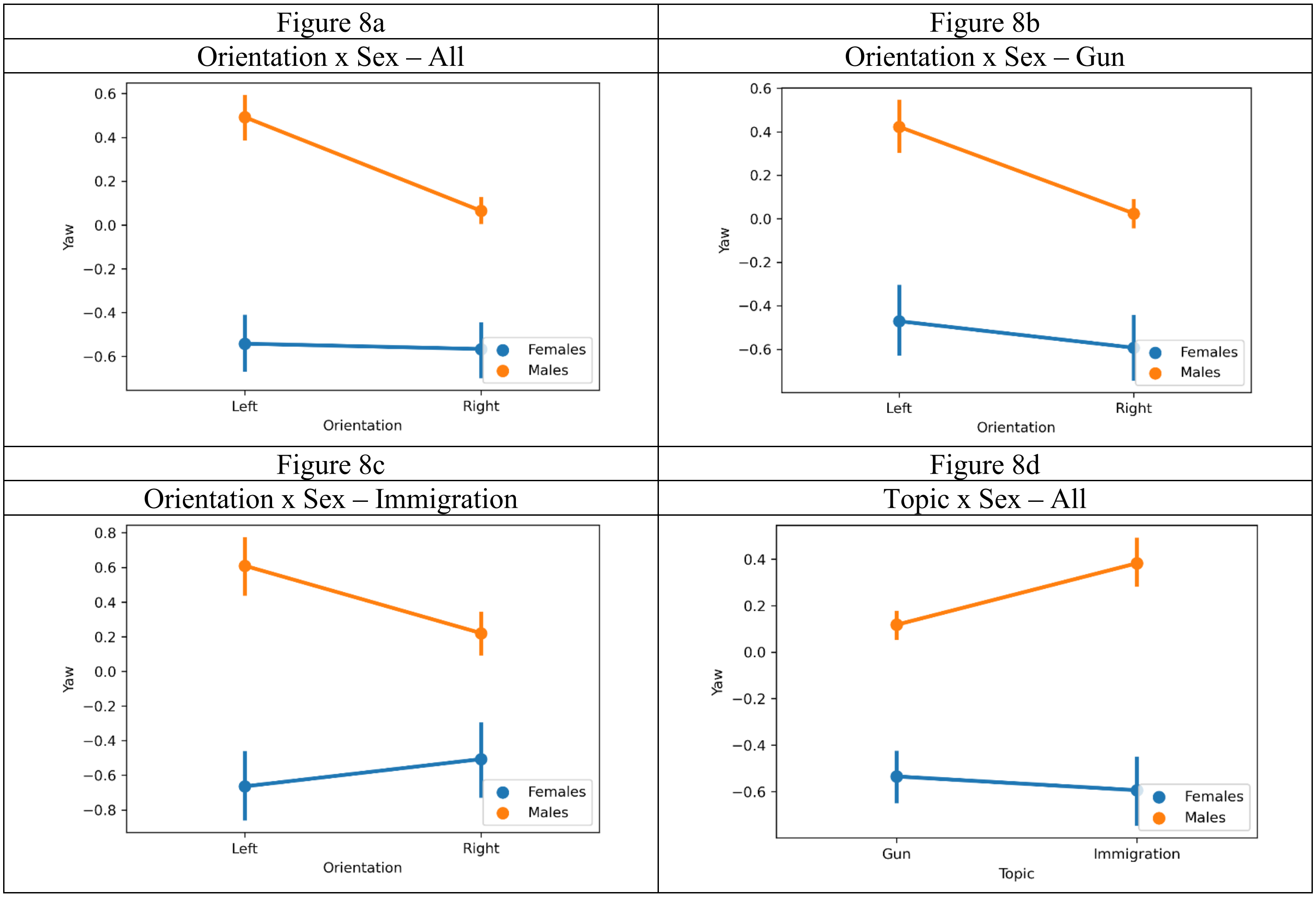


Figure 8a
Orientation x Sex – All

Figure 8b
Orientation x Sex – Gun

Figure 8c
Orientation x Sex – Immigration

Figure 8d
Topic x Sex – All

Section 2 – Hypothesis Testing

To test for the hypotheses presented in Chapter 8, the sample of images was grouped by race, sex, and topic, so that only members of the same race, sex, and topic were compared to one another in (e.g., white female pro-immigration v. white female anti-immigration). To illustrate the process of image analysis, we begin with the group demonstrating the largest sample size as an example. Our largest group of comparison was white males confined to the gun topic, with a sample size of 20,531 for the smaller sample.

White Males – Gun

White males in the gun domain was the largest subsample. Two sets of parallel analyses were run, one on the entire corpus of images for white males in the gun domain, and a second set of analyses with the pitch and yaw data constrained to the limits proposed by Wang and Kosinski (2018). The results for two sets of analyses do not differ by much, and as such only the larger sample is presented here.

In order to compare white males following Everytown to white males following the NRA, the two subgroups needed to be equivalent in number of observations. To make these two groups of analysis even in sample size, the larger group was randomly sampled to reach the same number of subjects that were available in the smaller group. For every logistic regression analysis, 10-fold cross-validation was performed, a sampling procedure that ensures that every element of the data is part of both the train and test sets. For each model the data were standardized, and each logistic regression used a Least Absolute Shrinkage and Selection Operator (LASSO) penalty for regularization. LASSO regression is best utilized when attempting to reduce overfitting in a model, when features or columns might outnumber sample size, or when many of the components of the model can be reduced to zero without losing much information (Maina, 2021; McNeish, 2015). Additionally, models employed a ‘SAGA’ solver, an optimization method related to stochastic average gradient (SAG) but with better convergence (Defazio et al., 2014). SAGA solvers are optimal for sparse regression matrices as well as large data, and are often the best choice for solvers according to sklearn documentation (Defazio et al., 2014).

For each model, metrics related to both accuracy and area under the curve (AUC) are reported as measures of classification power. Accuracy is defined as the ratio of correct predictions to total predictions, while area under the curve is the ratio of true positives to false positives. AUC is typically seen as a superior metric for model fit in comparison to accuracy, because models with high accuracy can sometimes be poor classifiers. Despite this, accuracy is perhaps more intuitive, so both are presented here. Error is presented in standard error of the mean (σ_M) of cross validated scores, and is a measure of how closely the model mean approximates the probable population mean.

Age, Pitch, Yaw, Emotions, Sex Typicality – 'Apyest'. Of the variables for these analyses, age was the least predictive, with the model demonstrating an average AUC = .505 ($CI_{95\%}$ = [.494, .517], σ_M = .0051). The accuracy of the model was 51%, not much better than chance. The model for pitch and yaw was a better classifier, AUC = .569 ($CI_{95\%}$ = [.562, .577], σ_M = .0033), accuracy = 55%. Using the emotional expression variables as predictors was comparable to the success of the pitch and yaw model, with AUC = .586 ($CI_{95\%}$ = [.578, .594], σ_M = .0036), achieving an overall accuracy in predictions of 55%. Sex typicality was the most effective predictor of these variables, AUC = .613 ($CI_{95\%}$ = [.598, .629], σ_M = .0067). Sex typicality proved an accurate classifier for 59% of the images.

These predictors were combined into one logistic regression model (age, pitch, yaw, emotions, and sex typicality, or together 'Apyest') with orientation being the criterion variable. This 'Apyest' model achieved a better AUC score than any of the previous models, AUC = .646 ($CI_{95\%}$ = [.632, .659], σ_M = .0060). Perhaps unsurprisingly, this model also had the best accuracy thus far, correctly categorizing 60% of the data.

ROC curve plots are presented in Figure 9a – 9e. For each of these plots, the dashed red line represents a model that performs no better than chance. Each of the semi-transparent lines represents one fold of the 10-fold cross validated model, while the solid blue line represents the mean across all folds. One can interpret the classification power of the model by assessing the curve of the blue line. Models that are better classifiers will have blue lines that arc towards the upper left corner of the plot, while models that are poor classifiers will have mean lines that hug the random chance line. Models with less error will have the fold lines tightly surrounding the mean line, while models with more error will demonstrate a greater spread around the mean. ROC curve plots for all analyses are presented in Appendix L.

Figure 9

ROC Plot – Apyest – White Male Gun

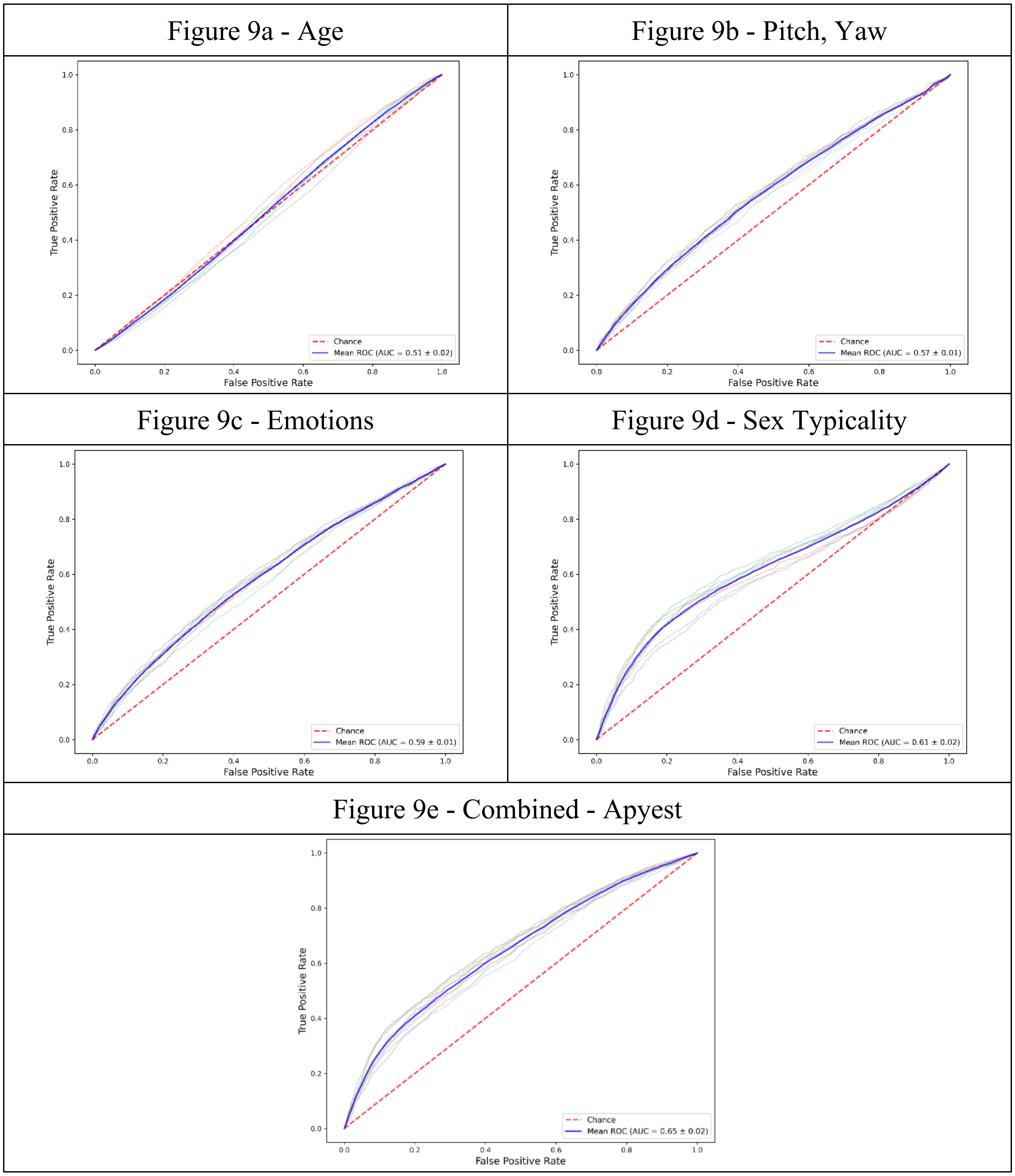


Figure 9a - Age

Figure 9b - Pitch, Yaw

Figure 9c - Emotions

Figure 9d - Sex Typicality

Figure 9e - Combined - Apyest

Feature Analysis. Recall that for each image, 4,096 features were extracted. Singular Value Decomposition (SVD) was performed on these features for just the group of interest (white males in the gun topic, in this case), leaving the 500 feature columns that were most important to classification for the group of interest.

Results from the cross-validated feature model were more impressive than any of the previous models, AUC = .746 ($CI_{95\%}$= [.736, .756], σ_M = .0044). Accuracy on the model using only features averaged 68% across all folds. Adding age, emotions, pitch, yaw, and sex typicality to the feature model did not improve classification substantially, AUC = .750 ($CI_{95\%}$ = [.740, .761], σ_M = .0045), accuracy = 69%. This confirms Hypothesis 1 for the white male gun analysis, replicating the primary finding presented in Wang and Kosinski (2018) and Kosinski (2021). Features alone were a superior classifier than the 'Apyest' model, and the inclusion of the 'Apyest' data to the feature model did not offer much new information in regards to classification.

The same analysis was performed on the masked image set. A total of 4,096 features were extracted from each masked image, and SVD was performed, reducing the feature set to the 500 most influential features. Comparing the classification power in the masked model to the classification power of the whole image model should reveal the importance of the background in image classification.

The removal of the background data did not appear to be very important to the model, causing only a minor reduction in predictive power, AUC = .735 ($CI_{95\%}$ = [.721, .748], σ_M = .0059). The accuracy of this model using only masked features was 67%, a slight reduction from the whole image feature model, but still far superior to the 'Apyest' model constrained on a sample by sex and race. Adding the 'Apyest' data to the masked feature model had a negligible effect, AUC = .741 ($CI_{95\%}$ = [.727, .755], σ_M = .0063), accuracy = 68%. These findings confirm Hypothesis 2 for this comparison, and suggest that the background of the image is not particularly effective in regards to being a classifier, at least not in comparison to features related to the face in the image. See Figure 10a – 10d.

Figure 10
ROC Plot – Features – White Male Gun

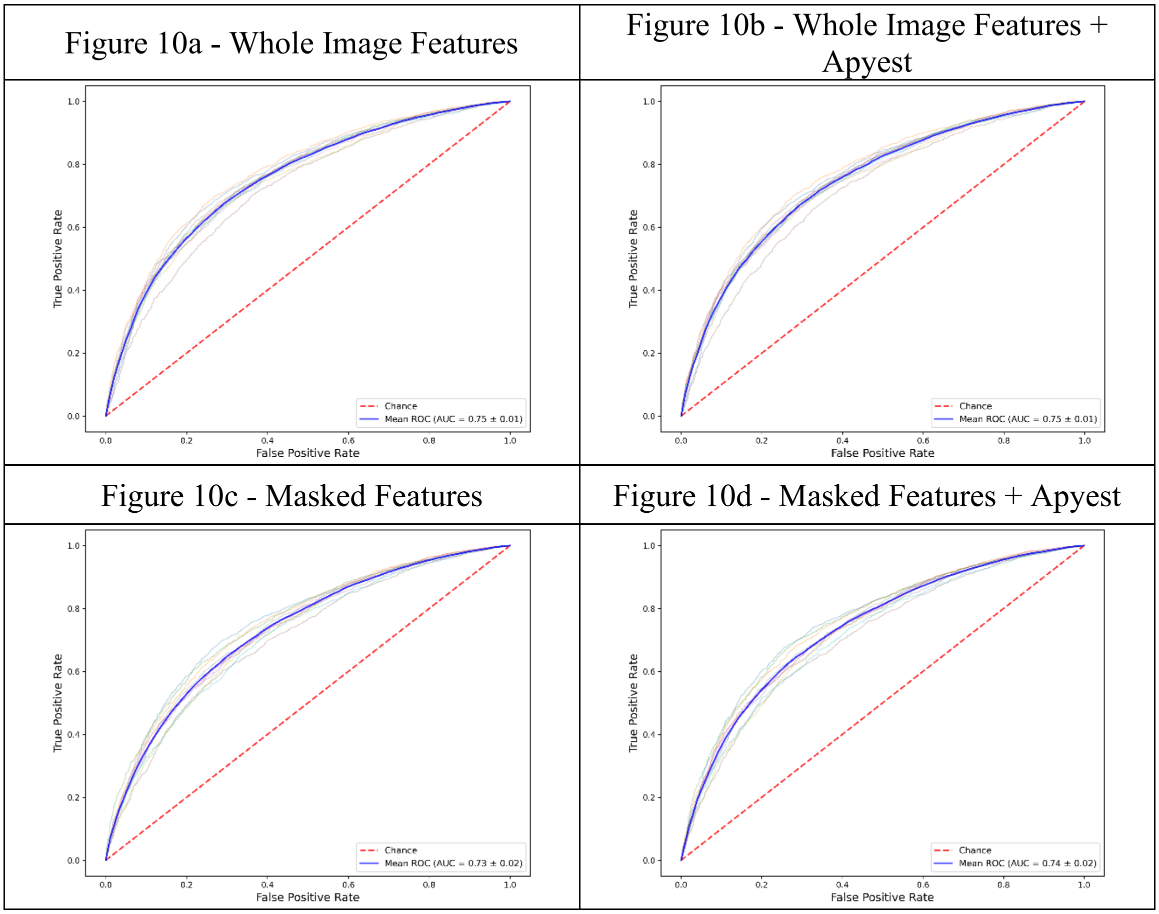


Figure 10a - Whole Image Features

Figure 10b - Whole Image Features + Apyest

Figure 10c - Masked Features

Figure 10d - Masked Features + Apyest

Point Coordinates. Using the dlib library, 68 facial point coordinates were taken for each image. These point coordinates were then fit to a logistic regression model, utilizing only the point coordinates for prediction. By utilizing these point coordinates in such a manner, we can reduce or eliminate the influence of anything unrelated to feature morphology.

This model was successful in classification, AUC = .701 ($CI_{95\%}$ = [.689, .712], σ_M = .0049), although not as successful as the feature models. The accuracy of the model was 65%, slightly reduced from the feature models but far enough from chance to demonstrate that facial morphology is almost certainly influential in terms of model success, at least for white males in the gun topic, confirming Hypothesis 3 for this comparison. Adding 'Apyest' to the model improved metrics slightly but not dramatically, AUC = .710 (CI = [.698, .721], σ_M = .0051, accuracy = 66%). See Figure 11a – 11b.

Figure 11
ROC Plot – Facial Points – White Male Gun

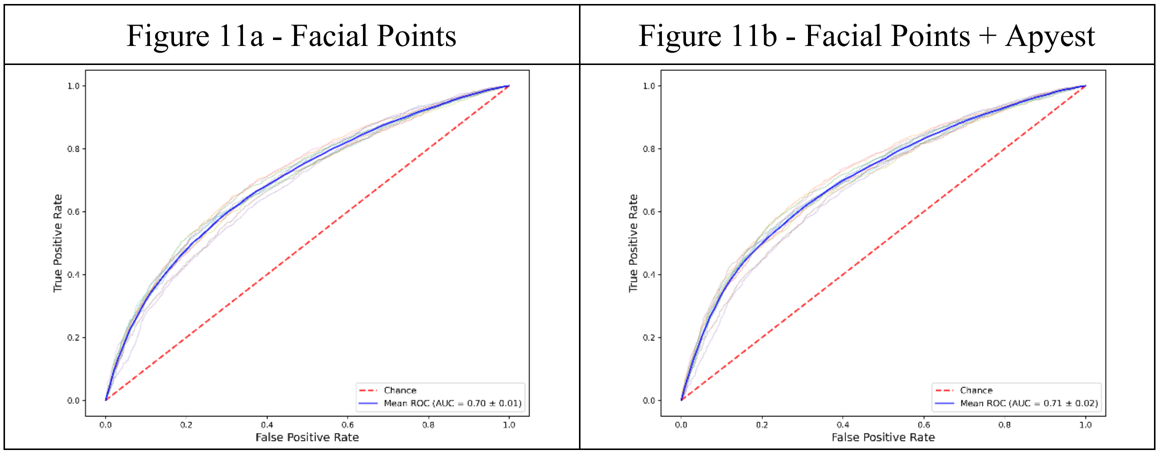


Figure 11a - Facial Points

Figure 11b - Facial Points + Apyest

Previously, neural networks have been described as 'black box' systems, because researchers do not necessarily have easy access to their internal workings. In other words, researchers often know their neural networks are working from the output of the model, rather than understanding each transformation the data is going through at each layer of the network itself. This provides an opaque understanding of how the classifier is coming to make its decisions. However, we might reverse engineer some data in order to see what the classifier is basing its decisions on.

To do this, the probability likelihood ratio of belonging to either the left or right orientations was captured for each image. These probabilities were then divided into quartiles. By comparing those images most likely to be classified into both left and right groups, that is, the first quartile compared to the fourth, one might get an idea of what differentiates left and right subjects, according to the classifier. Mean point coordinates are demonstrated for the left, right, and combined white male gun groups in Figure 12. Blue points and lines represent the mean point coordinates for subjects in the left-most quartile, while red points and lines represent those of the right-most quartile. Facial area in the left groups appears to be quite a bit smaller than in the right groups, as well as left groups demonstrating a more open mouth and right groups demonstrating a slightly shorter nose in length.

Figure 12

Facial Quartile Points Plot – White Male Gun

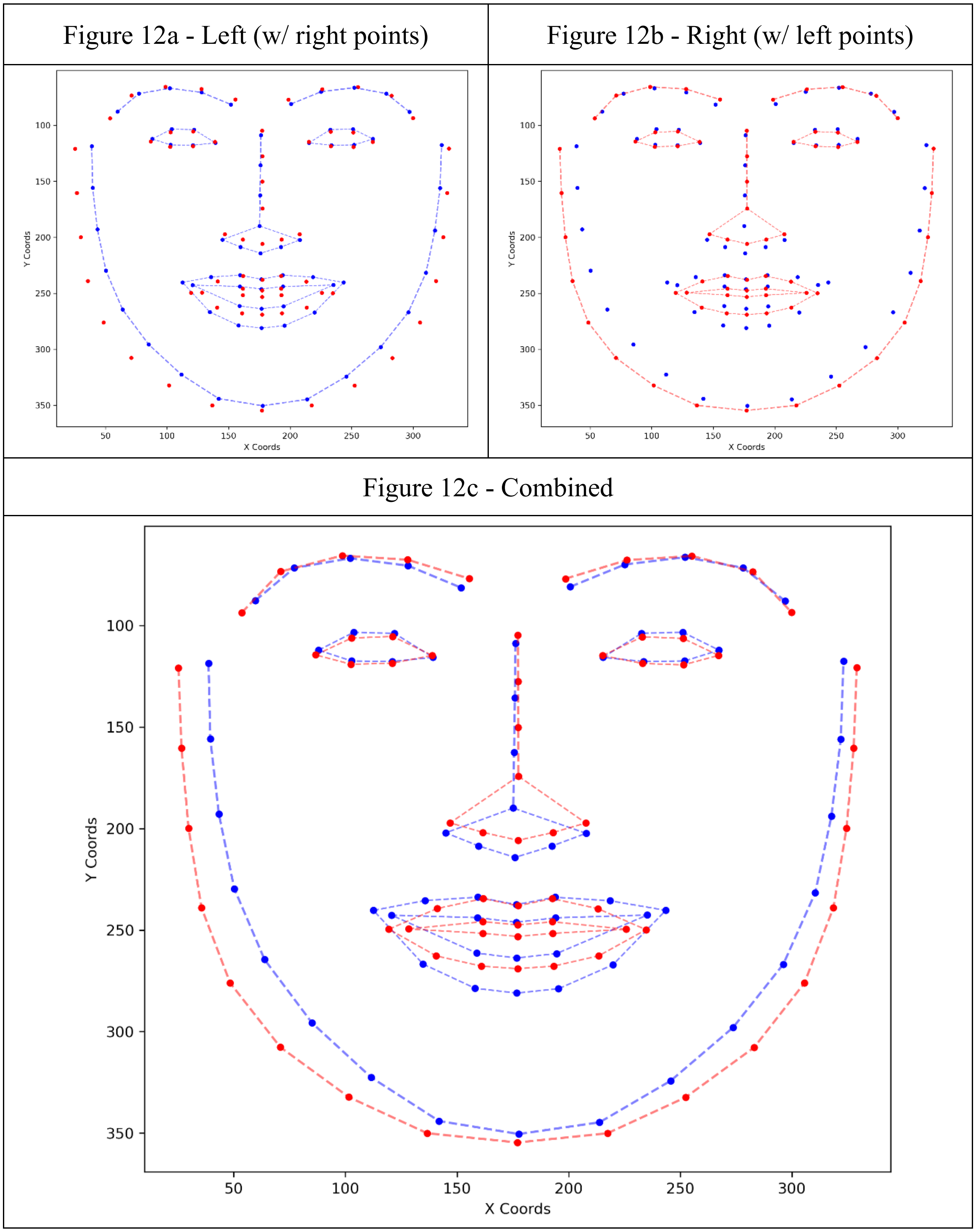


Figure 12a - Left (w/ right points)

Figure 12b - Right (w/ left points)

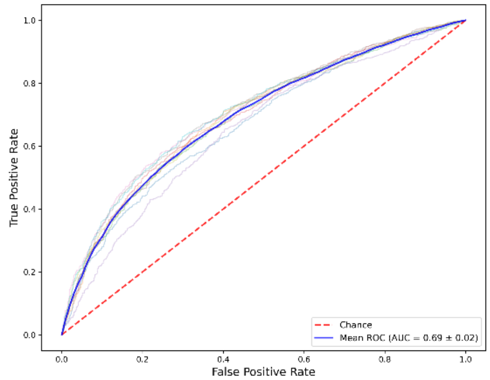
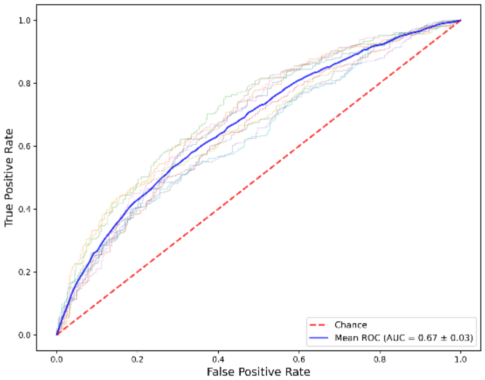
Figure 12c - Combined

Because the mouth can vary so much as a function of emotional expression, points related to the mouth were eliminated to determine if a model relying upon facial morphology but not facial points would be successful in image categorization. Utilizing only the remaining points resulted in a rather negligible decrease in model success, AUC = .692 ($CI_{95\%}$ = [.681, .701], σ_M = .0045). The accuracy of the facial points without mouth coordinates was 64%. This confirms Hypothesis 4. Even when eliminating the mouth points, this model was still able to accurately categorize images, adding further support to the idea that images can be categorized by facial morphology alone. Adding the 'Apyest' variables to the no-mouth model resulted in a slight increase in correct classification, AUC = .704 ($CI_{95\%}$ = [.693, .715], σ_M = .0050), accuracy = 65%.

The two most prominent emotions demonstrated in the dataset were 'happy' and 'neutral', with over 83% of the sample being in either one of these groups. The point coordinate data were reduced down to those subjects the emotion classifier deemed as being 'happy', thus attempting to isolate on images demonstrating similar facial expressions across the white male gun domain. When constrained to just 'happy' subjects, model classification still proved to be successful, AUC = .696 ($CI_{95\%}$ = [.683, .709], σ_M = .0059). Model accuracy for only facial points on just happy subjects was 65%. Adding the 'Apyest' data generated a modest increase in classification efficacy, AUC = .707 ($CI_{95\%}$ = [.698, .717], σ_M = .0042), accuracy = 65%. See Figure 13a – 13b.

Figure 13

ROC Plot – Happy – White Male Gun

Figure 13a - Face Points - Happy	Figure 13b - Face Points - Neutral
	

Images with neutral faces were also isolated on in such a manner. Both the neutral point only model and the point model with 'Apyest' data proved to be accurate classifiers, AUC = .691 (CI$_{95\%}$ = [.675, .706], σ_M = .0068) and AUC = .699 (CI$_{95\%}$ = [.684, .714], σ_M = .0066), respectively. Accuracy of the neutral point model was 64%, while the neutral point model with 'Apyest' data had an accuracy of 65%. First and fourth quartile facial meshes were taken for both 'happy' and 'neutral' models and are displayed in Figure 14.

Figure 14

Facial Quartile Points Plot - Happy and Neutral - White Male Gun

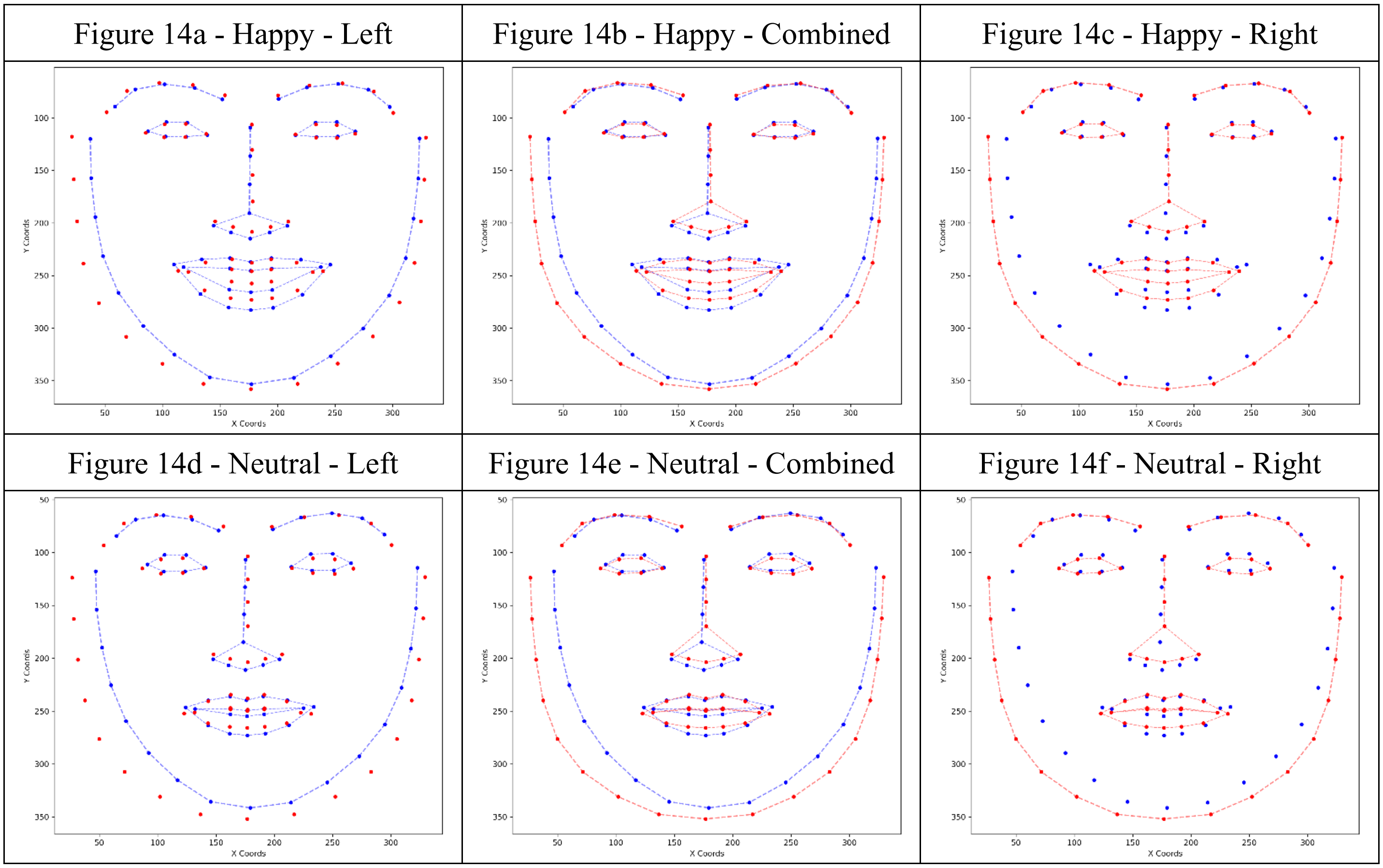


Figure 14a - Happy - Left

Figure 14b - Happy - Combined

Figure 14c - Happy - Right

Figure 14d - Neutral - Left

Figure 14e - Neutral - Combined

Figure 14f - Neutral - Right

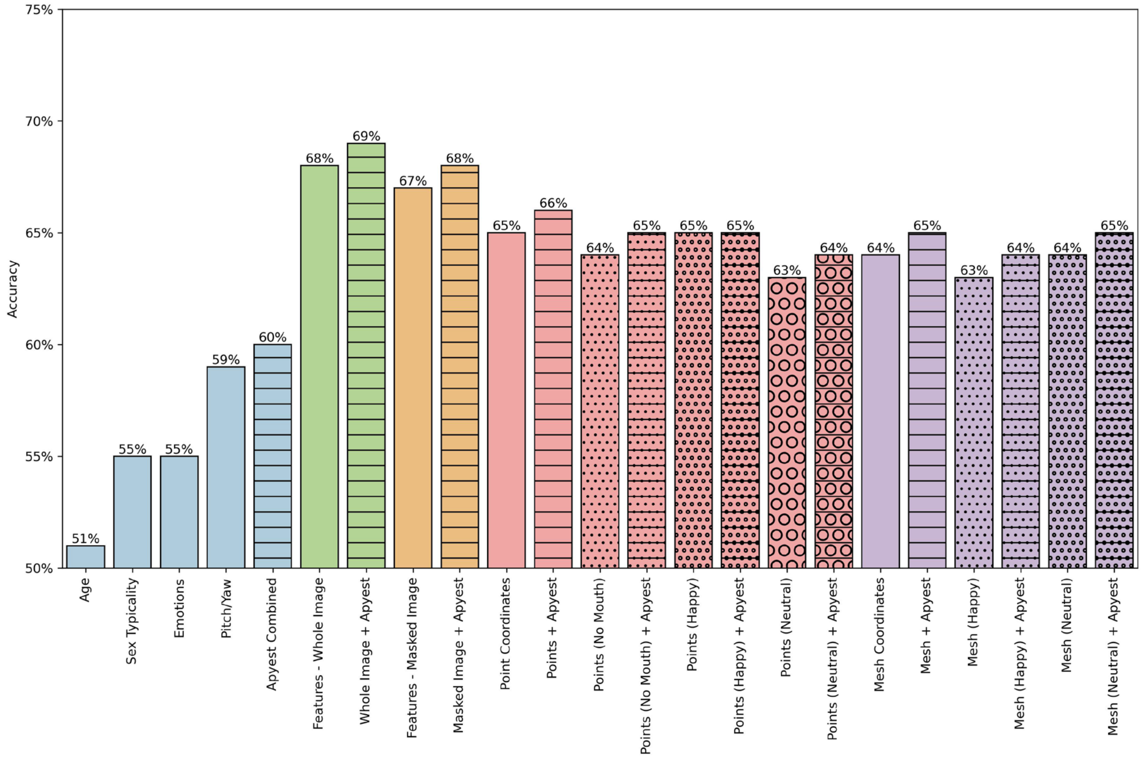
These findings confirm Hypothesis 5. Point models constrained by facial expression performed far better than chance, proving that images can be categorized by facial morphology alone and that facial expression cannot be the primary determinant of these models' success.

Mesh Coordinates. In addition to the point coordinate data, mesh coordinate data were also collected using the mediapipe library. In contrast to the dlib library which stores 68 facial points, media pipe stores 468 facial points to make its lattice. The X and Y coordinates for these 468 points were used as predictors in a logistic regression model. This model comprised only of mesh face coordinates was enough to accurately classify images 64% of the time, AUC = .691 ($CI_{95\%}$ = [.679, .703], σ_M = .0052), confirming Hypothesis 6. Adding the 'Apyest' variables improved classification slightly but not substantially, AUC = .704 ($CI_{95\%}$ = [.691, .716], σ_M = .0055), accuracy = 65%.

The mesh coordinates were also narrowed by facial expression, similar to the point coordinate data. With a sample of only happy subjects, the mesh coordinates were suitable predictors, accurately classifying 63% of subjects, AUC = .685 ($CI_{95\%}$ = [.677, .692], σ_M = .0034). Results for the same model with 'Apyest' variables were also positive, AUC = .698 ($CI_{95\%}$ = [.690, .707], σ_M = .0039), accuracy = 64%. Neutral subjects were also accurately classified at a rate of 64%, AUC = .692 ($CI_{95\%}$ = [.675, .710], σ_M = .0078), and the model with the 'Apyest' variables improved upon classification slightly, AUC = .705 ($CI_{95\%}$ = [.687, .723], σ_M = .0078), accuracy = 65%. These findings confirm Hypothesis 7, and prove that images can be classified when only utilizing mesh coordinates, even when controlling facial expression. Accuracies and AUCs are plotted for the all image, white male gun models in Figures 15 and 16. Accuracy and AUC plots are available for all other models in Appendix M.

Figure 15

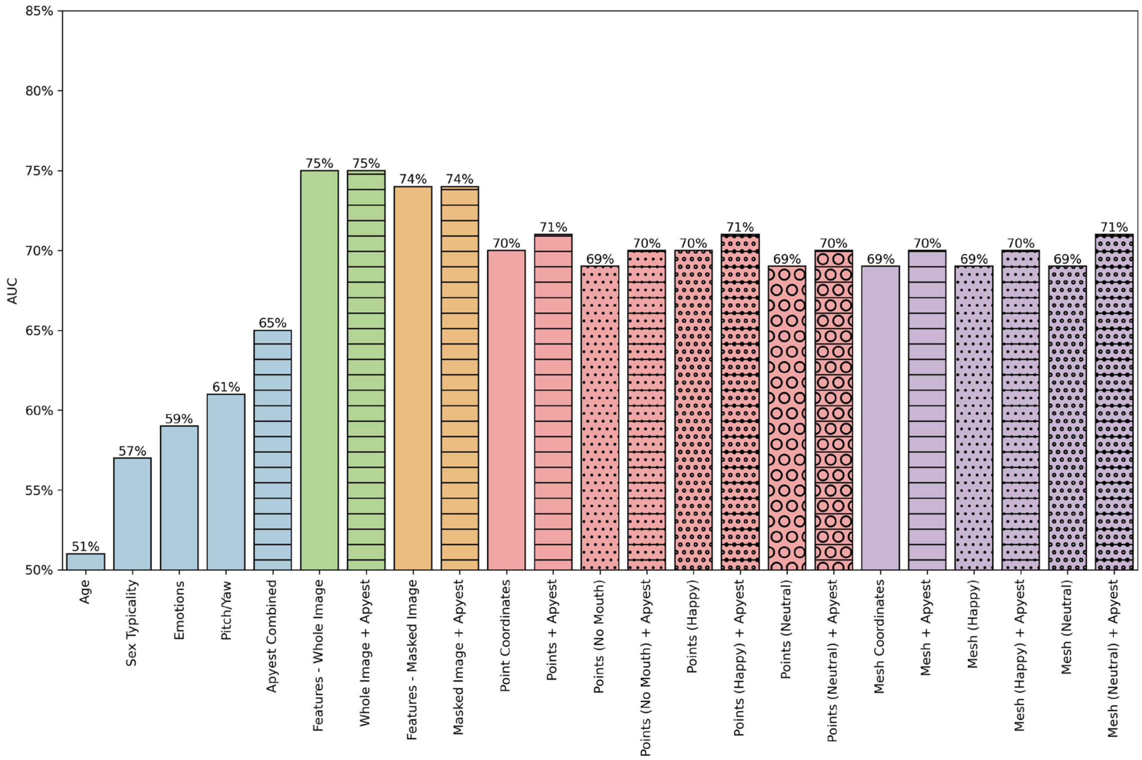
Accuracy – All Images – White Male Gun



Note: Blue bars represent the control variables and the combined ‘Apyest’ model. Green bars represent features of the whole image, while orange bars represent features of the masked image. Red bars represent point coordinates and purple bars represent mesh coordinates.

Figure 16

AUC – All Images – White Male Gun



Note: Blue bars represent the control variables and the combined ‘Apyest’ model. Green bars represent features of the whole image, while orange bars represent features of the masked image. Red bars represent point coordinates and purple bars represent mesh coordinates.

Summary – White Males – Gun. Results from this set of analyses confirm all of the hypotheses. First, the original conceptual effect from Wang and Kosinski (2018) and Kosinski (2021) was replicated, namely, being able to correctly classify images from their features alone. Features were a more powerful classifier than any of the other variables recorded. Further, removing the background from the images appeared to have little effect on classification power.

There were also several strong indicators that the classifier was utilizing facial morphology. Utilizing only facial point or facial mesh coordinates resulted in a decrease in classification power; however, the classifier was still able to categorize images well above chance levels. This was also the case across all of the analyzed subgroups (the no-mouth subgroups, the happy subgroups, and the neutral subgroups). These analyses provide the strongest support yet that feature only models are almost certainly utilizing facial morphology when classifying images.

All Groups

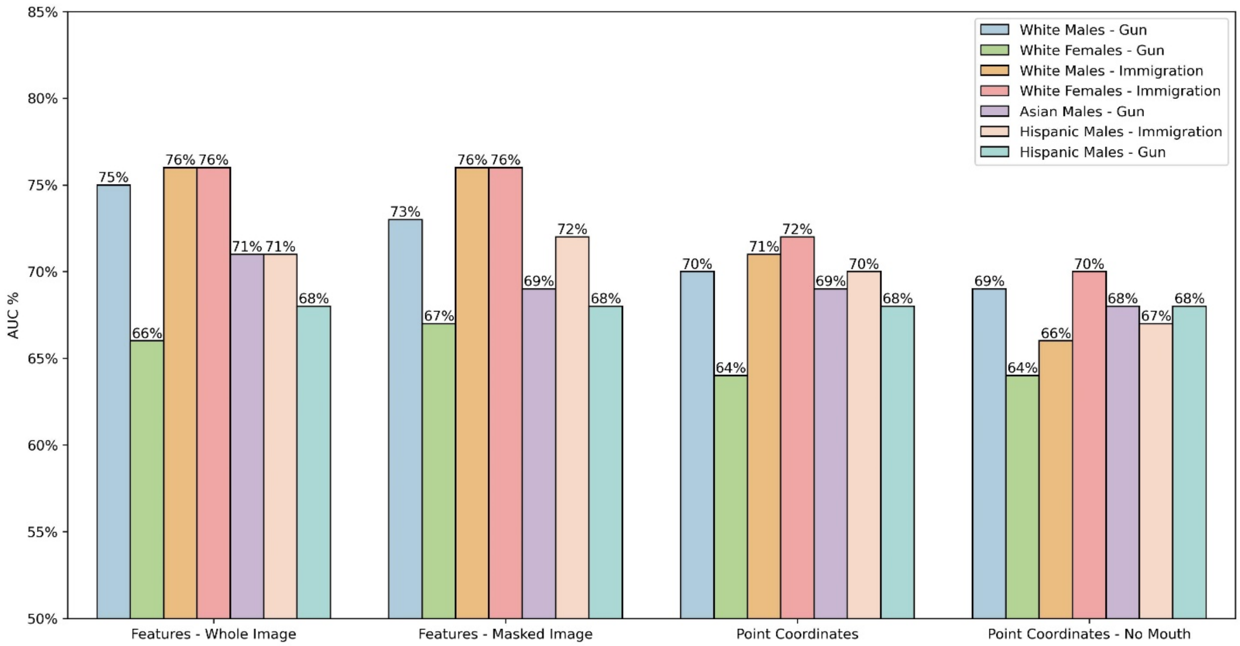
It is possible that the positive findings for the white male gun group are an aberration, and that the effects would not translate to the other groups of analysis.

In an attempt to determine if this methodology is consistent across all comparisons, the same analyses were performed on the remaining six groups of analysis. If the methodology is sound, we should expect a similar ability to classify subjects' political orientation across the rest of the comparisons using only facial morphology. If, however, the methodology is dependent upon the specific comparison in question, the negative results from these analyses should also be revealing.

Results for the models across all comparisons was confirmatory. Of particular interest is the comparison of the feature models to the masked models as well as to the point and mesh coordinates. AUCs are presented for all comparisons in Figure 17. Metrics for all models are presented in Appendix N.

Figure 17

AUC – Features and Point Coordinates – All Groups

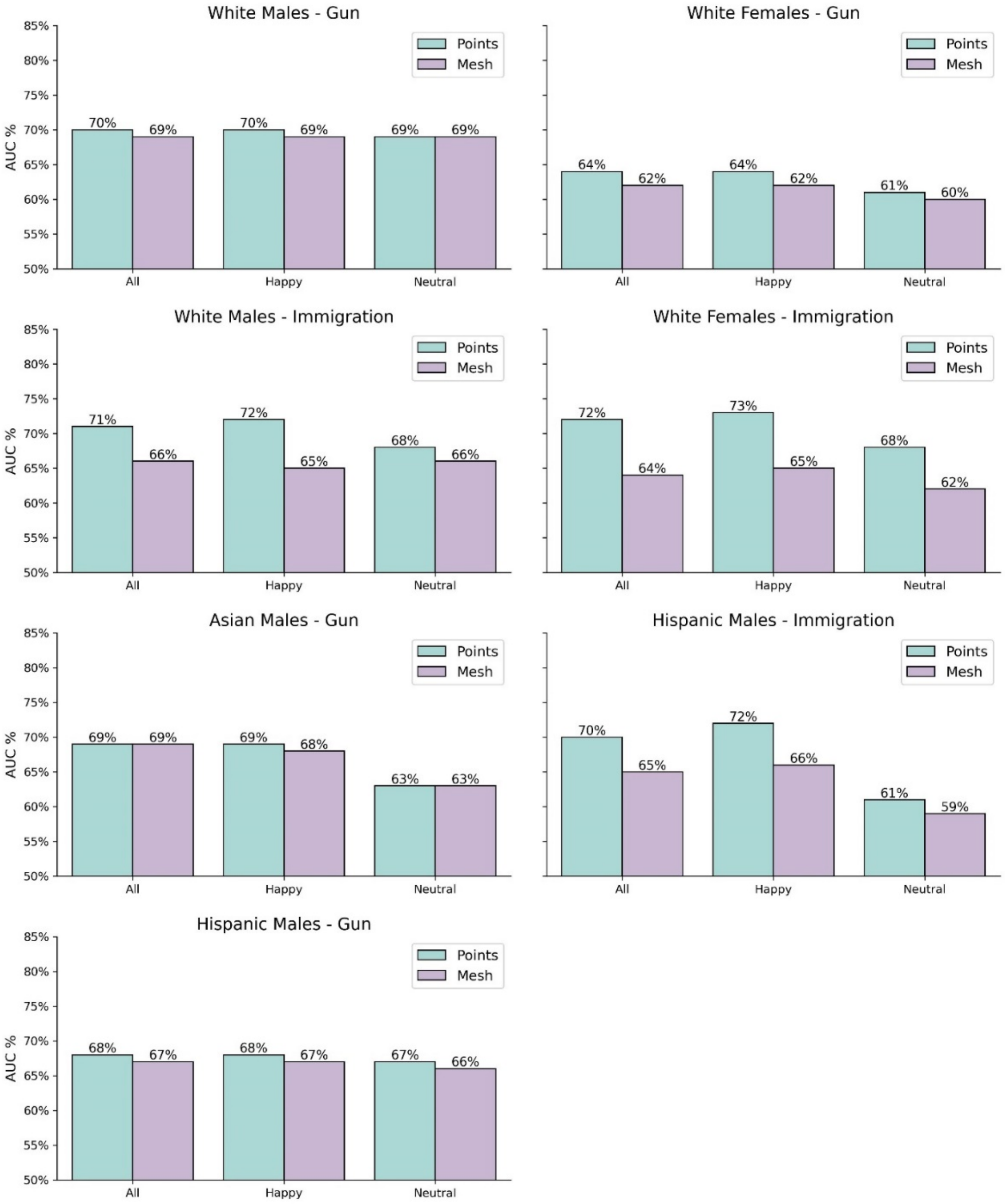


Several things are clear from the bar plot. First, every group of interest was able to be classified successfully using only features, replicating the principal finding across all conditions and confirming Hypothesis 1 for all groups of analysis. Second, removing the background from the image had little effect on classification power in any of the groups of analysis, proving that the background is not particularly informative in regards to classification and confirming Hypothesis 2 for each comparison. Third, the point coordinate models performed well above chance across every comparison, proving that images can be classified using only facial morphology, and that the initial confirmatory results were not exclusive to the white male gun comparison. This confirms Hypotheses 3 for all groups of analysis. Finally, even when removing the points related to the mouth, the classifiers were still able to accurately categorize images at rates well above chance, confirming Hypothesis 4 for all groups of analysis.

Regarding classifier quality, utilizing computer vision to attain point coordinates allowed for better classification than using the mesh coordinates, although both performed far better than chance. Across every analysis, the point coordinates performed similarly or better than the mesh coordinates. See Figure 18.

Figure 18

AUC – Point vs. Mesh Comparison – All Groups of Analysis



Examining Figure 18, we can draw several conclusions. First, point models constrained by happy and neutral facial expressions were solidly predictive, confirming Hypothesis 5 for all groups of analysis. Constraining the images by facial expression sometimes reduced model accuracy slightly, in particular for neutral facial expressions and only in some comparisons. However, each model demonstrated an AUC well above chance, and many of the comparisons demonstrated relatively consistent AUCs across conditions.

Figure 18 also illustrates that using only the mesh models was sufficient to categorize images, confirming Hypothesis 6 for all groups of analysis. Mesh models constrained by facial expression were also able to be classified, confirming Hypothesis 7 for all groups of analysis.

From these results, it is evident that the classifier was effective at classifying followers across all groups of analysis. However, it is still unclear if the effect is similar across groups. For example, it was previously discovered that, for the white male gun comparison, the facial area for conservative subjects was larger than the facial area for liberal subjects.

This effect was duplicated across all male groups of analysis. Males across all five of the comparisons demonstrated the same effect in regards to facial area. Conservative male subjects have a larger facial area, in general, than do liberal male subjects, independent of either the topic or of the racial background of the subject.

However, this effect was inverted across the female samples, with conservative women demonstrating, on average, less facial area than their liberal counterparts. The effect is more difficult to observe in the female immigration sample; nevertheless, liberal female faces remain longer and wider on average than conservative female faces.

Neutral facial point images for all groups are presented in Figures 19a – 19d and 20. All facial point images are presented in Appendix O.

Figure 19
Facial Quartile Points Plot - Neutral – White Males and Females

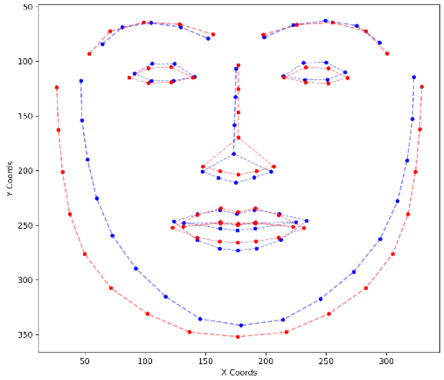
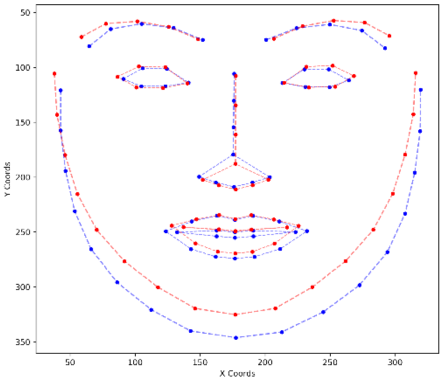
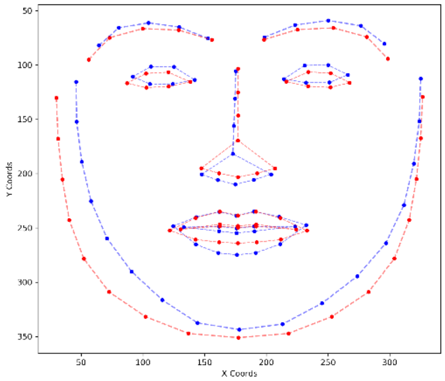
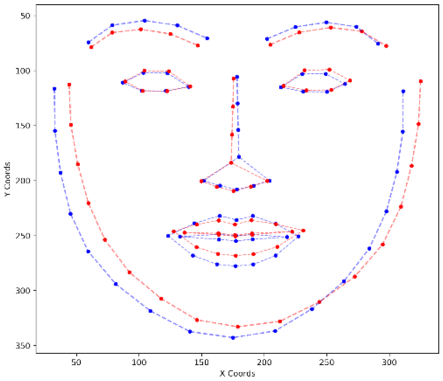
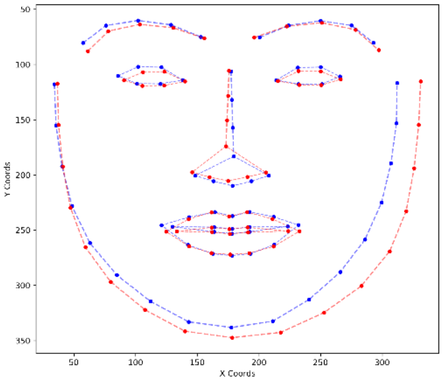
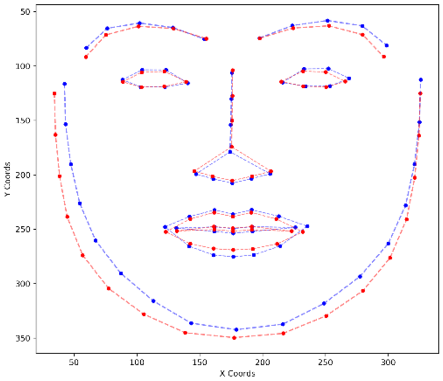
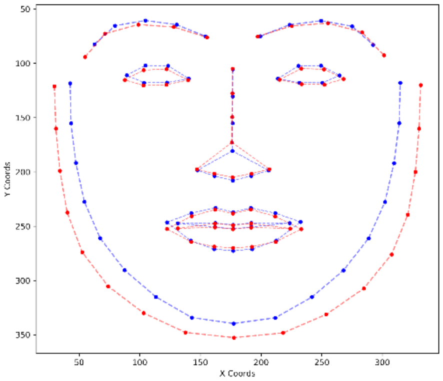
White – Gun – Neutral	
Figure 19a - Males	Figure 19b - Females
	
White – Immigration – Neutral	
Figure 19c - Males	Figure 19d - Females
	

Figure 20

Facial Quartile Points Plot - Neutral – Asian and Hispanic Males

	Males – Neutral	
Figure 20a - Asian – Gun	Figure 20b - Hispanic – Immmigration	Figure 20c - Hispanic – Gun
		

Finally, classification ability was markedly lower in the white female – gun comparison in comparison to the other comparisons. While it is unclear as to why this is, it is possible the sample for this particular subgroup was less ‘clean’ than the other groups, resulting in more noise in the model. Contrastingly, it is possible that the subjects in the white female gun groups simply demonstrate less variation between left and right groups. Further research is required to make the determination.