PNAS

Supplementary Information for

Deep models of superficial face judgments

Joshua C. Peterson, Stefan Uddenberg, Thomas L. Griffiths, Alexander Todorov, Jordan W. Suchow

Joshua C. Peterson.
E-mail: joshuacp@princeton.edu

This PDF file includes:

Supporting Information Text

1. Supplemental Results

Data Quality. Intra-rater (*i.e.,* test-retest) reliability was reasonably high on average across all of the tested attributes, as shown in Table S1 (keeping in mind that each observer re-rated 20% of all seen stimuli). Most individual participant sessions showed high levels of reliability, as can be seen in Figure S3's left skew. Sessions were not included in the models if their intra-rater reliability was below 0. Due to this conservative exclusion criterion, only 3.1% of all sessions tested were excluded from our attribute models. The attributes eliciting the lowest reliability (although still reasonably high) were familiarity and typicality, while those eliciting the highest were less subjective attributes such as age and gender. All other attributes had a median reliability above 0.6.

Qualitative Examination of Face Ratings. Figures S5 through S9 show the faces with the ten highest and ten lowest mean ratings for each perceived attribute. Selections for less subjective attributes such as *age*, *skinny/fat*, and *feminine/masculine* are straightforward, although there are some interesting observations. For example, the most *masculine*-looking men are not necessarily the most *dominant*-looking ones, who tend to look younger and have stronger jawlines. More subjective attributes also show clear patterns. Consistent with prior findings, children's faces look more *trustworthy* (3–5), while straight-faced *masculine*-looking faces with sunglasses appear least trustworthy. *Feminine*-looking faces were rated as more *attractive* (6), while older *masculine*-looking faces wearing glasses were rated as the least *attractive*. Faces rated as especially *smart* also often wore glasses (7), but appeared young to middle-aged, while the most *outgoing*-looking faces were often smiling. Finally, even perceived attributes with the lowest intra-rater reliability in the full set are reasonably interpretable. For example, young to middle-aged white *masculine*-looking faces were rated as more *typical*, while less *typical* faces were more diverse in terms of their race and gender. This is to be expected, given the fact that our MTurk sample reflected the demographics of the platform at large, and was therefore predominantly white. *Feminine*-looking faces were judged as looking more *familiar*, while less *familiar* faces were also more diverse. Our goal is to model the full extent of these effects, and not just what can be inferred qualitatively from inspecting such examples.

References

1. Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018.
2. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
3. Diane S Berry and Leslie Z McArthur. Some components and consequences of a babyface. *Journal of Personality and Social Psychology*, 48(2):312, 1985.
4. Diane S Berry and Leslie Z McArthur. Perceiving character in faces: the impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, 100(1):3, 1986.
5. Joann M Montepare and Leslie A Zebrowitz. Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology*, 30:93–161, 1998.
6. Christopher P Said and Alexander Todorov. A statistical model of facial attractiveness. *Psychological Science*, 22(9): 1183–1190, 2011.
7. Clare AM Sutherland, Julian A Oldmeadow, Isabel M Santos, John Towler, D Michael Burt, and Andrew W Young. Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, 127(1):105–118, 2013.

Perceived Attribute	Median Reliability	% Sessions Excluded
trustworthy	0.713	10.393
attractive	0.799	3.207
dominant	0.786	3.106
smart	0.740	4.615
age	0.955	0.303
fem./masc.	0.937	4.050
skinny/fat	0.778	0.000
typical	0.656	4.969
happy	0.867	1.286
familiar	0.520	16.393
outgoing	0.782	1.905
memorable	0.691	3.115
well groomed	0.794	1.274
long hair	0.933	0.322
smug	0.746	2.160
dorky	0.740	4.334
skin color	0.874	0.629
hair color	0.918	0.625
alert	0.697	2.532
cute	0.873	0.000
privileged	0.763	3.145
liberal	0.724	2.769
Asian	0.904	0.637
Middle Eastern	0.811	0.943
Hispanic	0.807	0.000
Pacific Islander	0.847	1.558
Native American	0.812	2.950
Black	0.894	7.599
white	0.919	0.312
looks like you	0.826	7.207
gay	0.721	4.545
electable	0.869	1.572
believes in God	0.674	2.839
outdoors	0.870	0.645

Table S1. Intra-rater test-retest reliability for all participant sessions and session exclusion statistics for each of the collected attributes.

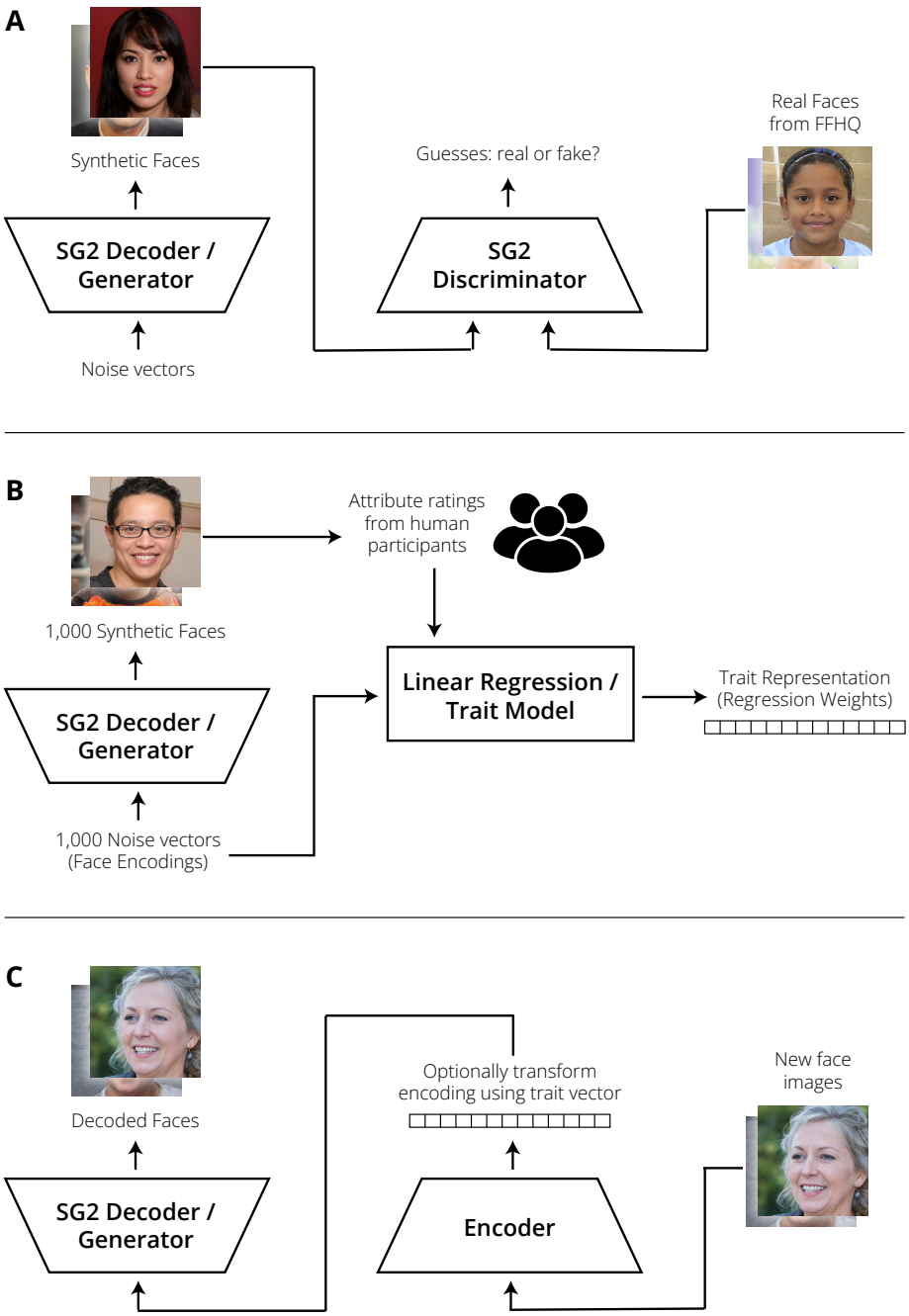

Fig. S1. A summary of our face attribute encoding and manipulation pipeline. **A.** StyleGAN2 (SG2) was trained independently using 70,000 images of real faces. It consists of a generator and discriminator network that are optimized jointly to produce realistic synthetic faces and tell synthetic ones from real ones respectively. **B.** We used a pre-optimized (fixed) SG2 generator to generate our 1,000 synthetic face stimuli that were used to obtain attribute ratings from human participants. We then regressed the encodings (features) used to produce the synthetic faces to the mean attribute ratings for each attribute, producing attribute representations (regression weights). **C.** An encoder algorithm is employed to output encodings for a new input face that would have produced the same face image when input to the generator network. Once an encoding is obtained, it can be modified using a attribute representation, which has the same dimensionality of the encoding, using vector arithmetic.

Fig. S2. Random example stimuli from our dataset of 1,000 curated synthetic face images generated using StyleGAN2 (1, 2) for use in all of our experiments.

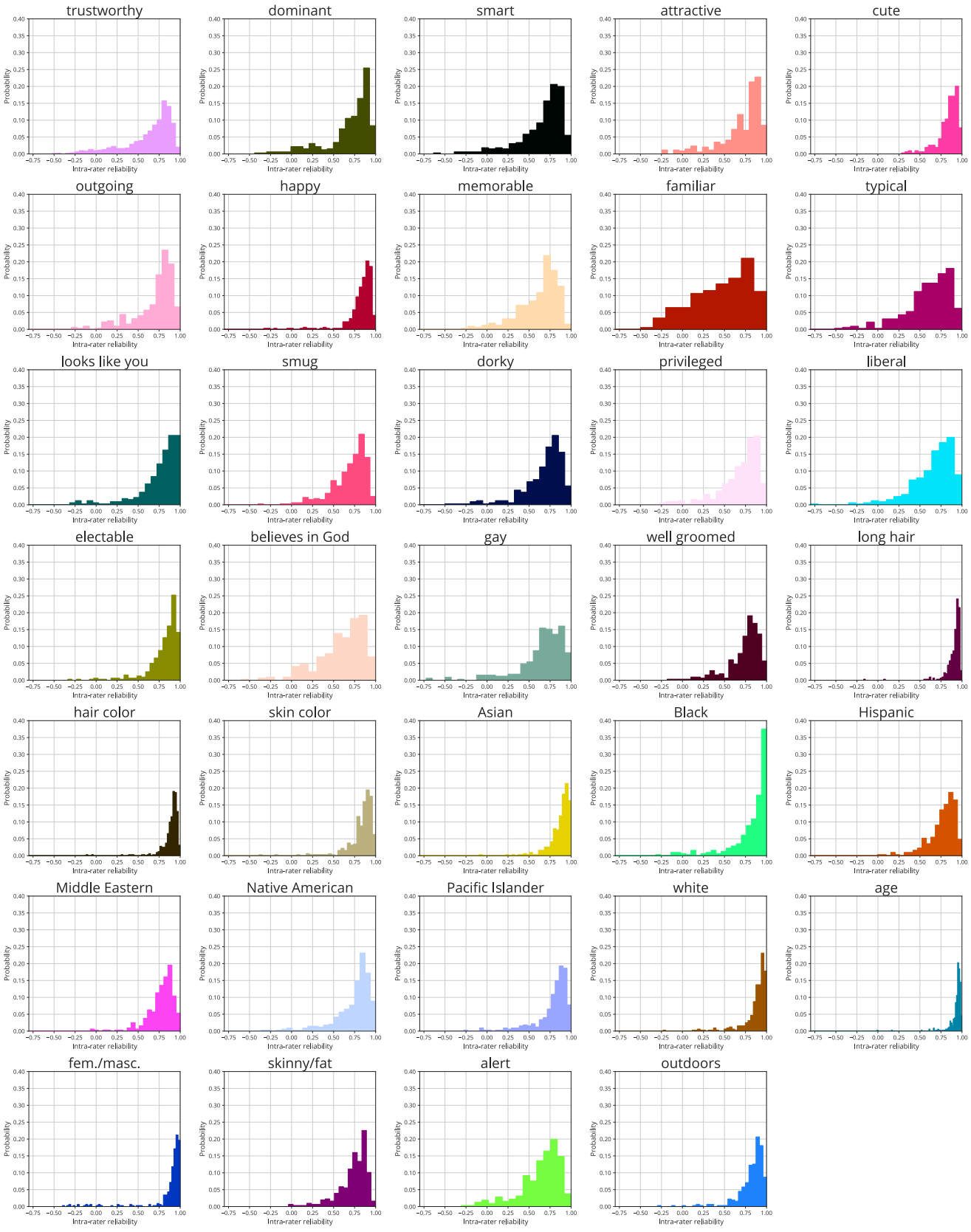

Fig. S3. Intra-rater reliability distributions for each measured attribute.

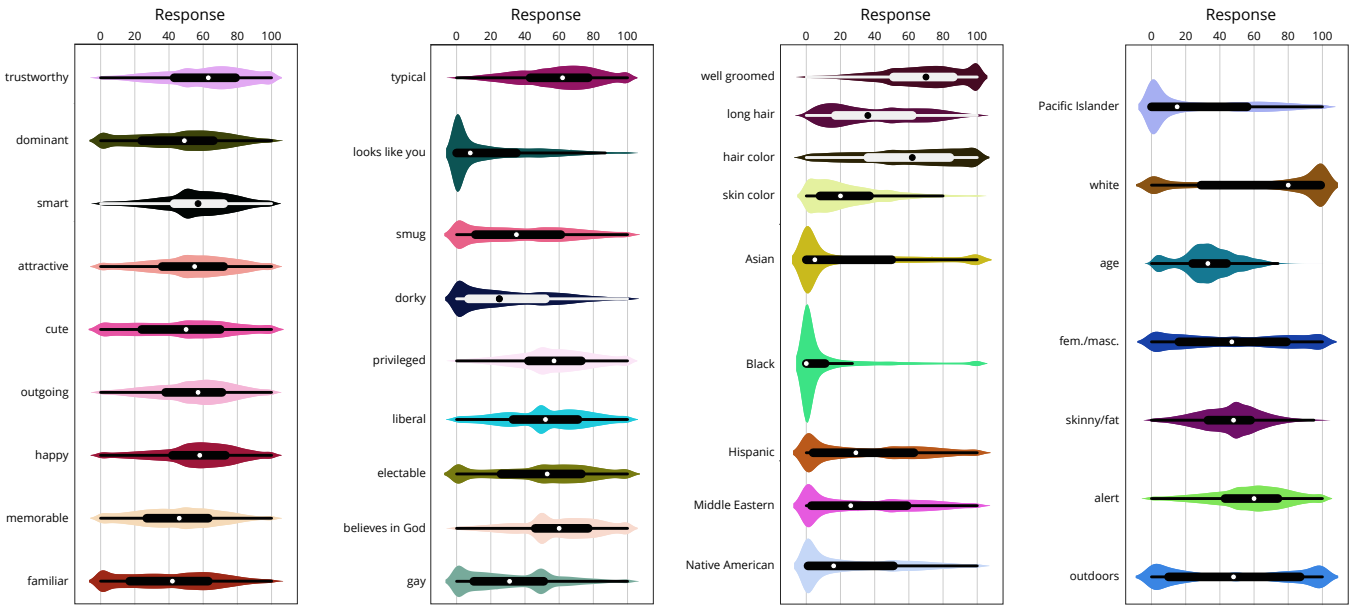

Fig. S4. Distribution of raw responses given by participants for images rated along each attribute. Boxplots at the center of each distribution represent the median as a white/black dot (depending on the contrast), the interquartile range as the thick opposite-colored line, and the remainder of the distribution (sans outliers) as the thinner lines (i.e., the "whiskers").

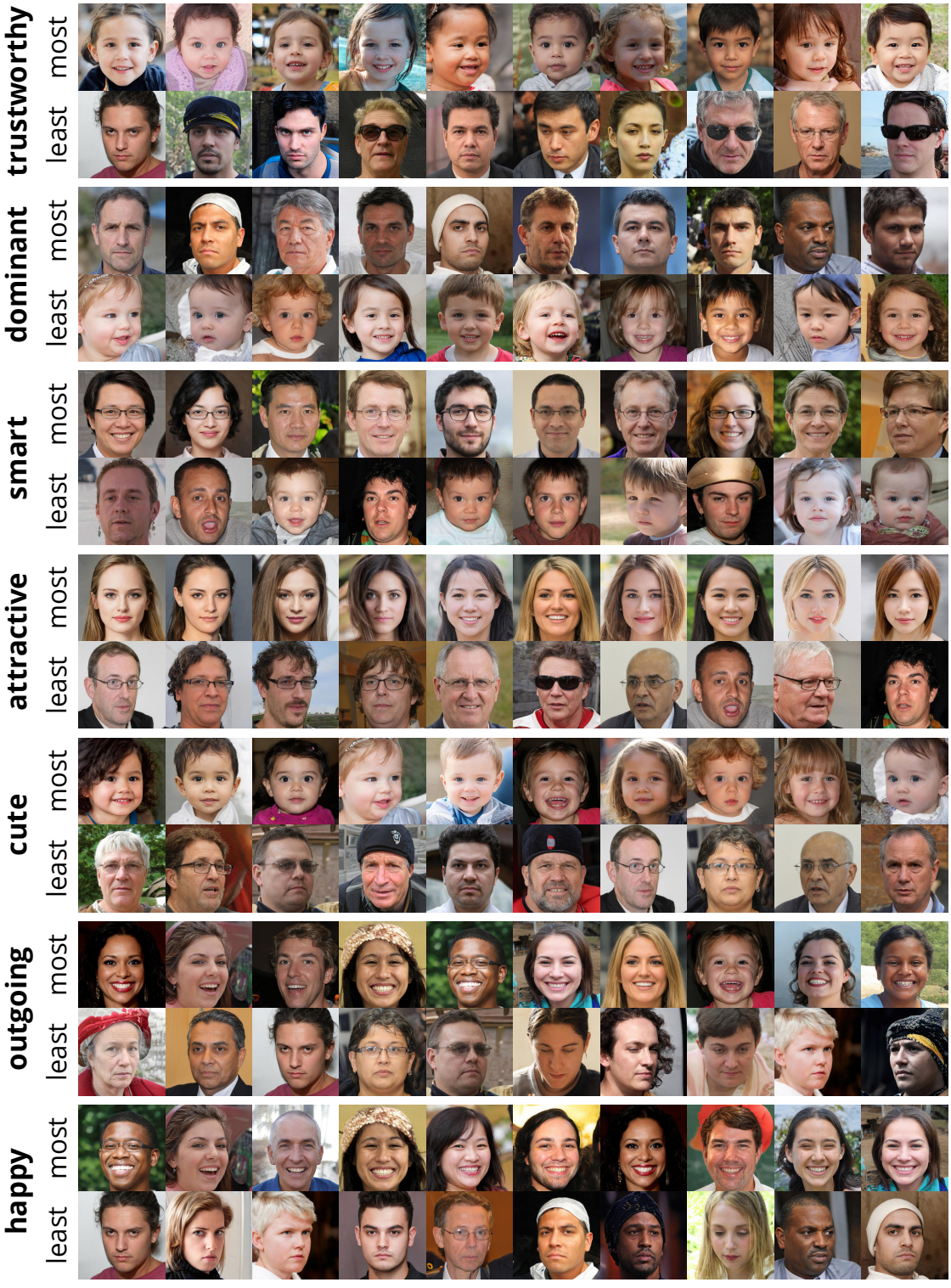

Fig. S5. Ten most and ten least congruent faces (highest/lowest mean ratings) for attributes *trustworthy*, *dominant*, *smart*, *attractive*, *cute*, *outgoing* and *happy*.

Fig. S6. Ten most and ten least congruent faces (highest/lowest mean ratings) for attributes *memorable*, *familiar*, *typical*, *looks like you*, *smug*, *dorky*, and *privileged*.

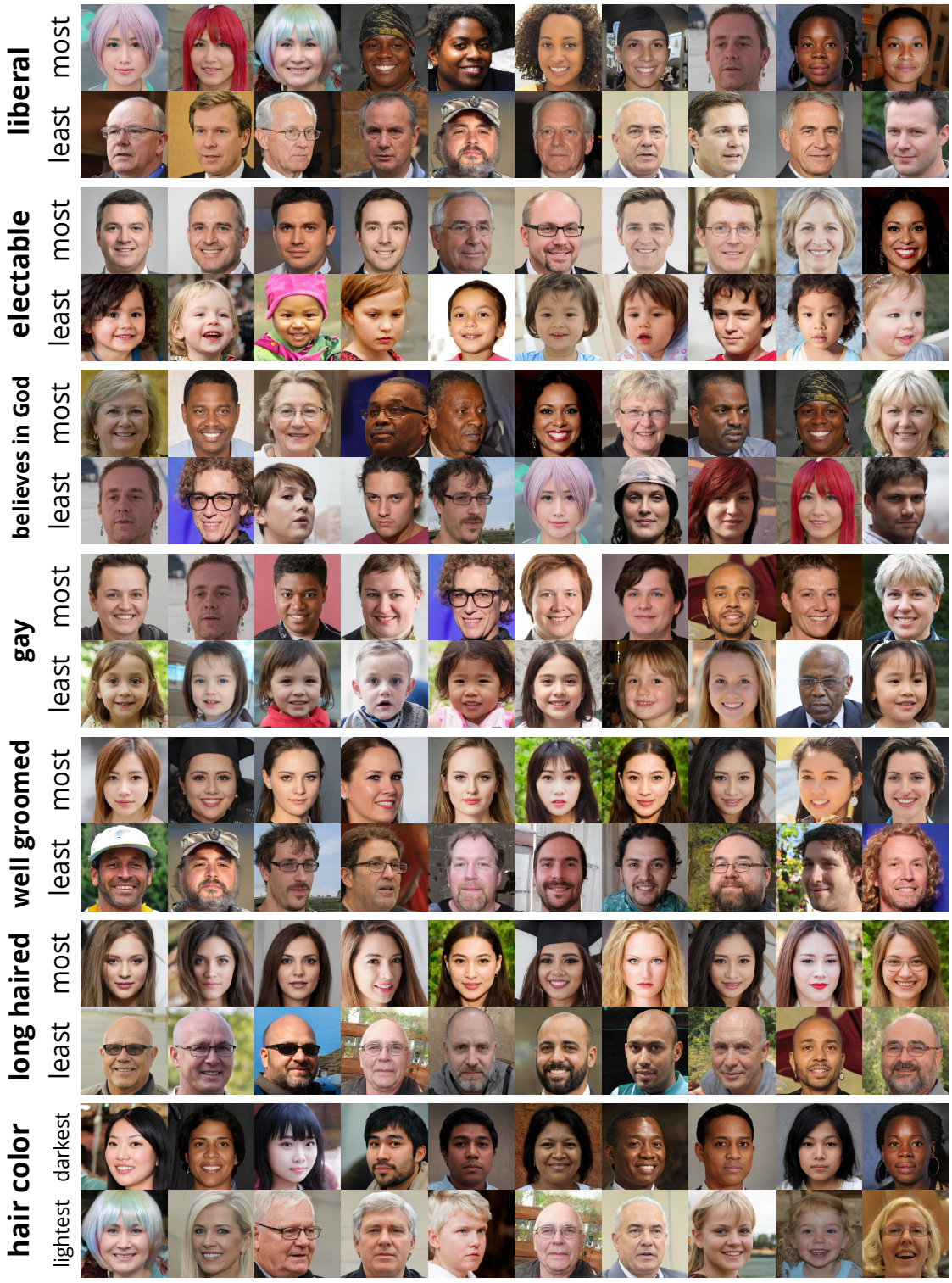

Fig. S7. Ten most and ten least congruent faces (highest/lowest mean ratings) for attributes *liberal*, *electable*, *believes in God*, *gay*, *well groomed*, *long haired*, and *hair color*.

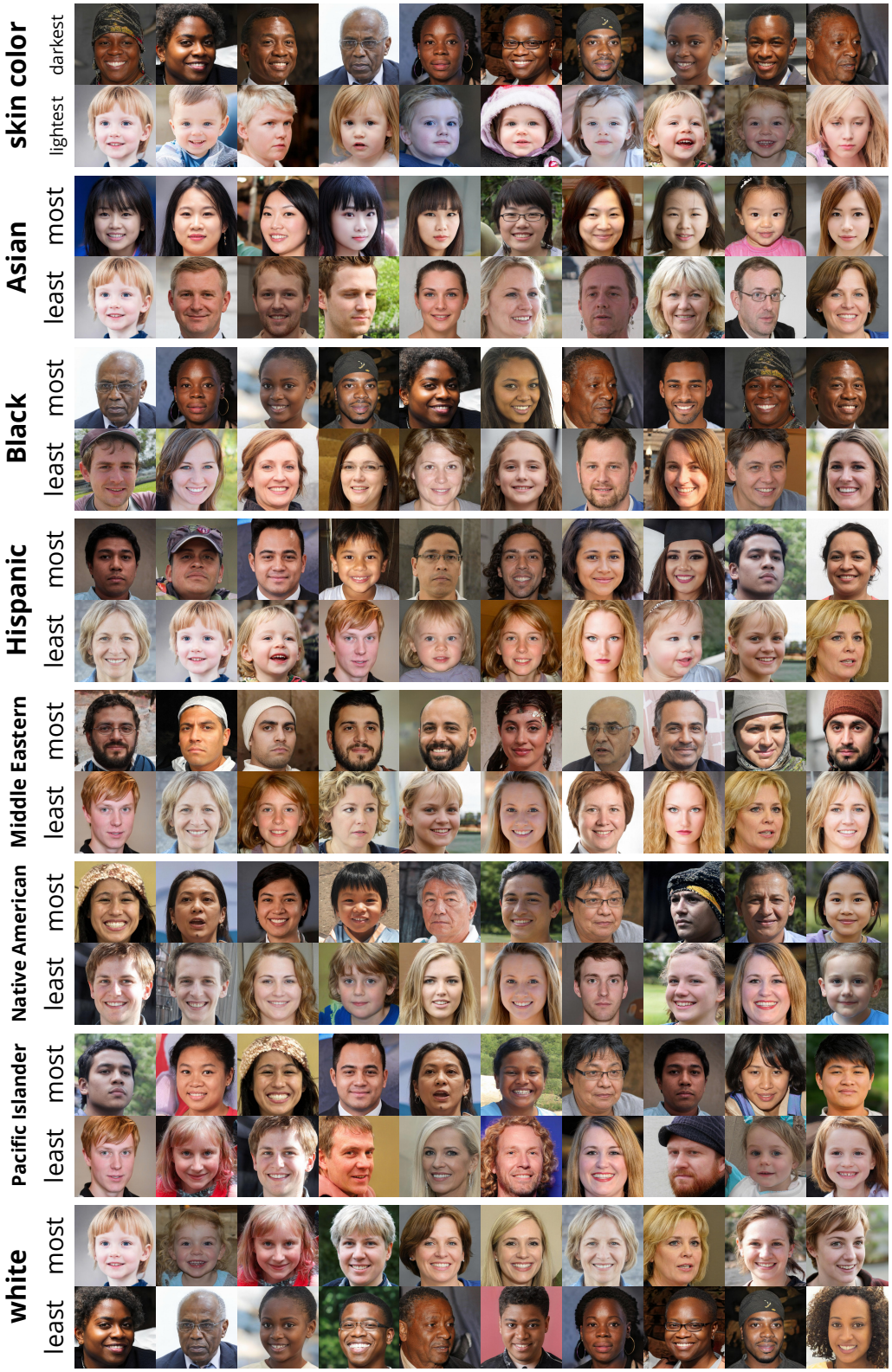

Fig. S8. Ten most and ten least congruent faces (highest/lowest mean ratings) for attributes *skin color*, *Asian*, *Black*, *Hispanic*, *Middle Eastern*, *Native American*, *Pacific Islander*, and *white*.

Fig. S9. Ten most and ten least congruent faces (highest/lowest mean ratings) for attributes *age*, *feminine/masculine*, *skinny/fat*, *alert*, and *outdoors*.

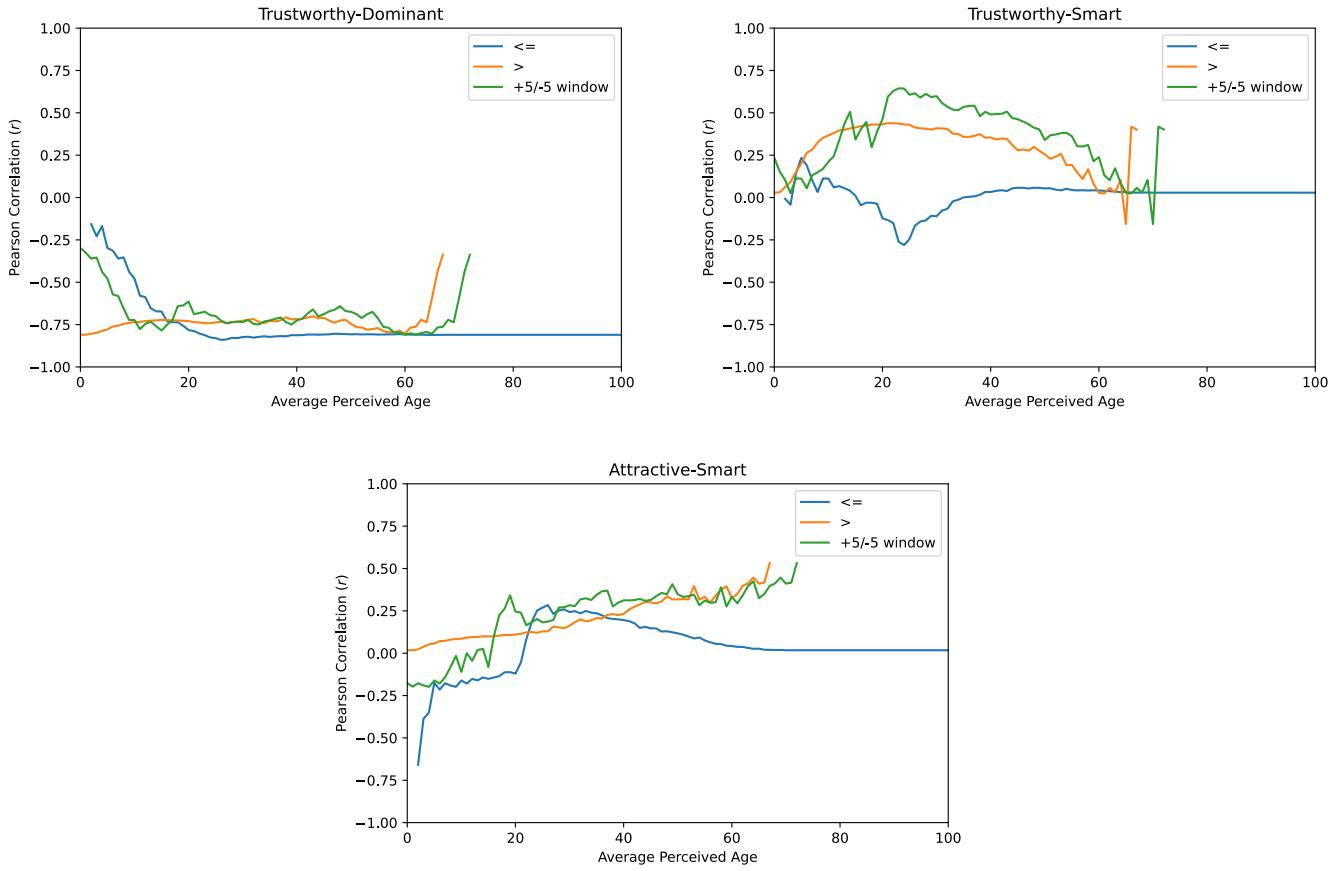

Fig. S10. Correlations between average attribute ratings as a function of age for three pairs of attributes. The blue curves plot correlations for faces less than or equal the age threshold on the x-axis. The orange curves plot correlations for faces greater than the threshold on the x-axis. The green curves plot correlations for faces within a sliding 10-year window around the values on the x-axis. For the attribute pair trustworthy-dominant (top left), the correlation is only affected at very young or very old ages. For trustworthy-smart (top right), the correlation increases up to ages around approximately 25, and then decreases again. For attractive-smart (bottom), the correlation becomes larger and more positive for older ages.

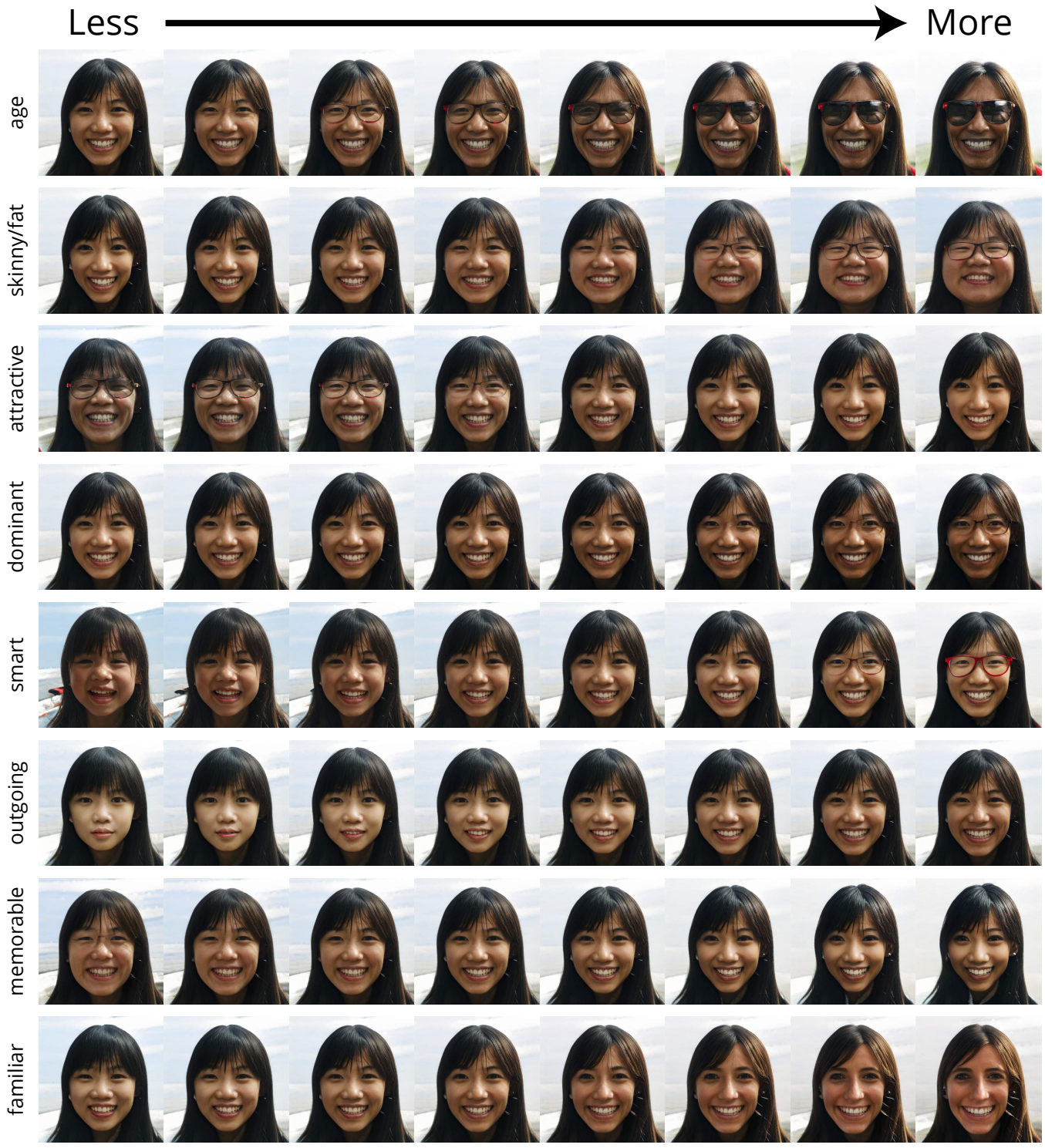

Fig. S11. Transformations controlling for perceived *trustworthiness* along perceived *age*, *weight*, *attractiveness*, *dominance*, *smartness*, *outgoingness*, *memorability*, and *familiarity*.

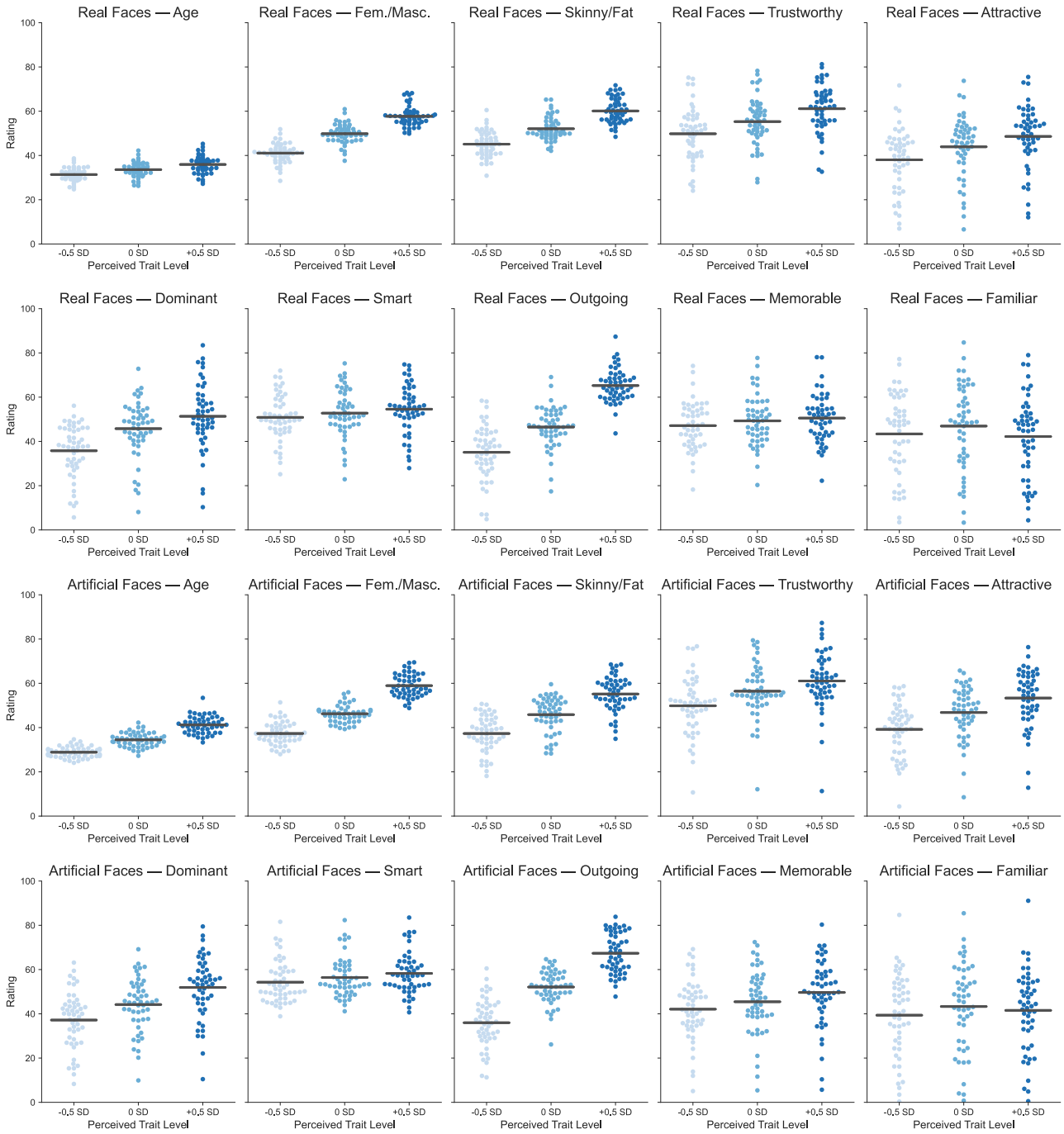

Fig. S12. Mean responses given by participants for each model validation experiment. Each dot represents the mean response of a participant at a given level of a perceived attribute manipulation for a given experiment. The mean of each distribution is given by the horizontal bars overlaid on the distributions.