Supporting Information for Individualized Models of Social Judgments and Context-Dependent Representations

Daniel N. Albohn, Stefan Uddenberg, and Alexander Todorov
Booth School of Business, The University of Chicago, Chicago, IL, United States

Detailed Method for Constructing Idiosyncratic Visual Models 1
Training A New Generative Model 4
Model Generation Average Test-Retest Correlations 7
Cosine Similarity Additional Results 8
Study 1 Additional Results 10
Study 2 Additional Results 12
Study 3 Additional Results 19
Study 4 Additional Results 21
Exploring Methods for Computing Idiosyncratic Visual Models 24
Group-Level and Individual Model Visualizations 26

Detailed Method for Constructing Idiosyncratic Visual Models

Here we describe the methodology for constructing idiosyncratic visual models. The method uses a modified version of StyleGAN-2 architecture (https://github.com/NVlabs/stylegan2-ada-pytorch). All code was run using Python 3.6.13 and PyTorch 1.10.0.

The procedure requires three steps: 1) stimulus creation; 2) stimulus selection (by participants); and 3) stimulus analysis (i.e., idiosyncratic visual model creation).

Step 1: Stimulus Creation

Stimuli for participants to categorize are generated from the latent space of a pretrained model. This is accomplished in one of two ways. The first way, which is used in Studies 1 and 2, is by projecting real faces into the latent space to act as a starting point. Stimulus projection/inversion, rather than random sampling, is necessary for certain models that have an overrepresentation of specific stimuli or attributes, like the StyleGAN-2 FFHQ face model, which largely consists of smiling faces due to the nature of the original training data (online portrait images) (*53*, *54*). An overrepresentation of smiling faces (or any other type of stimulus/attribute) is undesirable for obtaining an accurate and robust idiosyncratic visual model as it can bias representations in one direction. We decided on inverting real neutral face images into the StyleGAN-2 latent space in an effort to remove the oversaturation of smiling faces.

We projected 2,484 neutral faces from various available databases into the StyleGAN-2 latent space using a modified VGG encoder adapted by Peterson et al. (*3*) in Study 1 and a modified version of the *FeatureStyle* encoder (*55*) in Study 2. The neutral faces were taken from several face databases: the Chicago Face Database (*56*), FACES (*57*), NIMSTIM (*58*), RAFD (*59*), Face

Database (*60*), Face Research Set London (*61*), FERET (*62*), and RADIATE (*63*) image sets, as well as a number of internal face resources.

Next, we created the actual stimuli that participant would categorize by first averaging together the latents of a subset of randomly selected faces from the 2,484 faces inverted into the model latent space. Finally, we added a small amount of random Gaussian noise to the averaged latent to further differentiate it from the pool of inverted faces. This two-step process was repeated for each stimulus generated in Studies 1 and 2.

The second method is to sample stimuli directly from the latent space without first projecting specific faces into the space. As already noted, sampling faces directly from the StyleGAN-2 FFHQ pretrained model results in an over-representation of smiling faces. If over-representation of smiling faces is not an issue, one could theoretically sample directly from the model's latent space to obtain stimuli for generative reverse correlation. However, we also trained a new StyleGAN-2 model that outputs high quality face images that are largely neutral in appearance, which we used to sample images directly from the latent space in Studies 3 and 4. We discuss details about training this model in the next section.

Regardless of which stimulus sampling method is used, each image obtained from the latent space of the StyleGAN-2 model has a corresponding 18×512 matrix[1] of numeric values that represents that face in the model's latent space.

Step 2: Stimulus Selection (Participant Procedure)

After stimuli are generated (typically 300 – 1000), they are categorized by participants. On each trial, a single image is displayed to the participant along with three response options: 1) the judgment of interest (e.g., perceived "attractiveness"), 2) the conceptual opposite of this judgment (e.g., perceived "unattractiveness"), and 3) a "neutral" (or "neither/unsure") category. The rationale for including a "neutral" category was to obtain an unbiased, individualized starting point within the latent space for each participant to aid in creating high quality images from the idiosyncratic visual models. What one individual categorizes as "neutral" is likely to differ from one participant to the next (*64*, *65*).

Step 3: Idiosyncratic Visual Model Construction and Visualization

Constructing idiosyncratic visual models is done by binning participant selections of each of the three categories and averaging the latent vectors of the corresponding images. First, for each participant the averaged vector of the conceptually opposite target judgment is subtracted from the averaged vector of the target judgment. This is what we refer to as the idiosyncratic visual model. This vector represents the latent features that are unique to that single participant and judgment. Next, this idiosyncratic visual model vector is added to the average of all latent vectors that participant selected as the "neutral" or "neither" category. This is done to generate a high quality visualization for the individual participant, analogous to what typical noise-based reverse correlation refers to as a *classification image*. The images generated from the

[1]We used a 1×512 vector repeated 18 times for generating face stimuli in Study 3 and the Supplemental Study, as we observed higher quality final results from this procedure over the fully randomized 18×512 latent matrix.

idiosyncratic visual models reflect the mental prototypes that individuals hold of the particular target judgment. Finally, one can visualize both the target judgment and the conceptually opposite target judgment at varying levels of intensity by multiplying the idiosyncratic visual model by a constant. With a restricted latent space, such as the one used in Studies 1 and 2, we observed stable visualizations between +/- 6. However, with a larger latent space, such as the one used in the Supplemental Study and Study 3, visualizations typically go out of sample much faster. We discuss the implications of this in the next section and report an alternative interpolation method in the section, "Exploring Methods for Computing Idiosyncratic Visual Models".

More formally, constructing and visualizing idiosyncratic visual models is expressed as,

$$M_{iC} = \left(\hat{A}_i * C\right) + \bar{N}_i$$

Where M_{ic} represents the image latent vector for an individual participant, i, at a specific model interpolation value constant, C. $\hat{A}_i$ represents the idiosyncratic visual model for participant, i, and is computed by,

$$\hat{A}_i = \bar{A}_i - \bar{B}_i$$

where,

$$\bar{A}_i = \frac{1}{n}\sum_{j=1}^{n} A_j$$

is the average of all latent vectors that the individual participant selected as the target category, and,

$$\bar{B}_i = \frac{1}{m}\sum_{k=1}^{m} B_k$$

is the average of all latent vectors that the individual participant selected as the conceptually opposite of the target category. Finally,

$$\bar{N}_i = \frac{1}{p}\sum_{l=1}^{p} N_l$$

represents the average of all latent vectors for the images that the individual participant selected as "neutral" or "neither". A_i, B_i, and N_i are all an $M \times N$ matrix of real numbers representing latent values from the trained generative model.

Training A New Generative Model

In Studies 1 and 2, we projected real neutral faces into the StyleGAN-2 latent space to act as a starting point for generated stimuli. This method also ensured that each stimulus generated was neutral in expressivity. While this approach is effective and produces psychologically aligned idiosyncratic models, it is not without its limitations. Primarily, it artificially restricts the latent space to a subset of the entire space and, as a result, reduces the diversity of the images generated from the model's latent space. If generative reverse correlation is invariant to the underlying model the stimuli are generated from, more diverse stimuli should result in better and more accurate idiosyncratic visual models. In order to test this, we retrained a StyleGAN-2 model to have a more diverse and less biased latent space, particularly in terms of the over-representation of smiling faces present in the latent space of the original StyleGAN-2 FFHQ model.

Retraining the model results in two additional benefits over projecting neutral faces into the latent space. First, it eliminates the first step of our procedure whereby we need to project images into the latent space. While it may not be immediately apparent, eliminating this step has potential benefits for future work using generative reverse correlation (e.g., generating real time visual approximation of idiosyncratic models).

Second, if the stimuli being evaluated by participants were randomly selected from across the entire latent space rather than a subset of it, traversing the individual's own mental models should theoretically require less linear interpolation. More simply, and all else being equal, the interpolation between two random points in a restricted latent space will be closer together compared to two random points in an unrestricted latent space, resulting in smaller steps per constant for the restricted latent space and bigger steps per constant in the unrestricted latent space (see also the section titled) "Exploring Methods for Computing Idiosyncratic Visual Models").

Before retraining the model, we first had to acquire an adequate sample of high quality neutral-appearing face stimuli. Neutral face stimuli were obtained from a variety of sources, including the FFHQ dataset (*53*), the CelebA-HQ dataset (*66*), the neutral faces derived from datasets introduced in Study 1 (main text), and online image scraping. Every face used in the final training set was evaluated as "neutral in appearance or minimally expressive" by the first author and subsequently categorized as "neutral" by an emotion detection algorithm, *Deepface* (*17*). Aside from being neutral in appearance, images needed to be larger than 1024 × 1024 pixels, contain the majority of the face (e.g., not having the forehead out of frame), not too blurry or granulated, and not extremely rotated or oblique. In total, we acquired 47,724 high quality neutral faces that met the criteria for use in training the new model (over 75,000 training images with augmentation).

We fine-tuned our new model using the StyleGAN-2 FFHQ model as a base. All hyperparameters were the same as those used for StyleGAN-2-ADA FFHQ model training (*54*). Our model was trained for an additional 4,000 epochs and reached a final Fréchet inception distance score of 4.19, which is comparable to the original StyleGAN-2 FFHQ model trained on 70,000 face images. Following the naming convention of StyleGAN-2, we refer to this new model as the "Neutral and Minimally Expressive Faces-High Quality" model (NAMFHQ).

After training the NAMFHQ model, we compared the face images it produced to those produced by the original model to ensure that 1) the NAMFHQ model would generate high quality, realistic neutral face images at a higher rate than the original FFHQ model and 2) it would maintain the high diversity of the FFHQ model on other variables of interest, such as sex/gender, race, and age. We randomly generated 30,000 images from each model and had the same emotion detection classifier auto-classify the facial emotion of each image. The results from this classification are shown in Fig. S1.

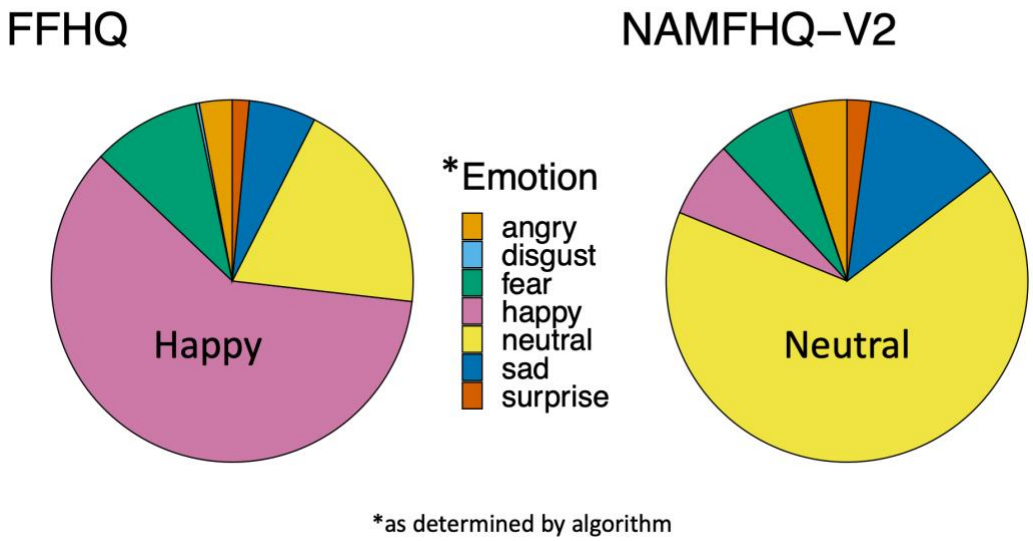

Fig. S1. Comparison of face images generated from pretrained StyleGAN-2 FFHQ model (left) and our new Neutral and Minimally Expressive (NAMFHQ; left) model. Whereas most of the images generated by the original FFHQ model are smiling or faces with happy expressions, the majority of the images generated from our model are neutral in appearance. Importantly, our model still maintains diversity in every other regard (e.g., sex/gender, race, and age; see Fig. S2).

The results from this comparison show that the new model is generating neutral face images as intended. The majority of the images generated from the original model are smiling or faces with happy expressions, whereas the majority of the images generated from our model are neutral in appearance. Importantly, our model still maintains diversity in every other regard. Examples of randomly generated images from our new NAMFHQ model are shown in Fig. S2.

Fig. S2. Randomly generated example images from our Neutral and Minimally Expressive Faces-High Quality (NAMFHQ) Model.

Observations and Comparison to Projection Methods

The one major analytical difference between our previous model and method (Studies 1 and 2) and this new NAMFHQ model is the number of linear interpolation steps that can be used before going out of sample. Previously, we were observing that it was possible to interpolate each visual model +/- 6 steps on average. With our new model, the interpolation steps are reduced to +/- 2 for judgments with consistent and clear visual properties associated with them (e.g., feminine/masculine) and +/- 3 or +/- 4 for more complex judgments (e.g., attractiveness). We predicted that this might be the case, as the previous models we were using had a tightly bound and restricted latent space. That is, by projecting real neutral faces into the model's latent space, we were artificially constraining the usable latent space to only a portion of what was available. Thus, each linear interpolation step was relatively small when we computed images from each individualized visual model. In contrast, we were able to use the full latent space in our new, neutral-only GAN model. In other words, any two selected stimuli from the constrained latent space model are more likely to be closer together in that latent space compared to any two randomly selected stimuli drawn from the unconstrained latent space model. These properties are likely driving why idiosyncratic models are more quickly going out of sample in the new model, as interpolation steps between the images represent larger changes in the latent space.

Model Generation Average Test-Retest Correlations

The average post-exclusion test-retest correlations for each study and condition was as follows:

Study 1: Masculinity-Femininity (r = .69, SD = .22); Trustworthiness (r = .42, SD = .25)

Study 2: Masculinity-Femininity (r = .61; SD = .17); Age (r = .46; SD = .22); Familiarity (r = .29, SD = .20); Attractiveness (r = .50, SD = .15)

Study 3: Masculinity-Femininity (r = .87; SD = .15); Attractiveness (r = .64, SD = .24)

Study 4: Car (r = .56; SD = .26); Child (r = .60; SD = .23); Money (r = .52, SD = .24)

Cosine Similarity Additional Results

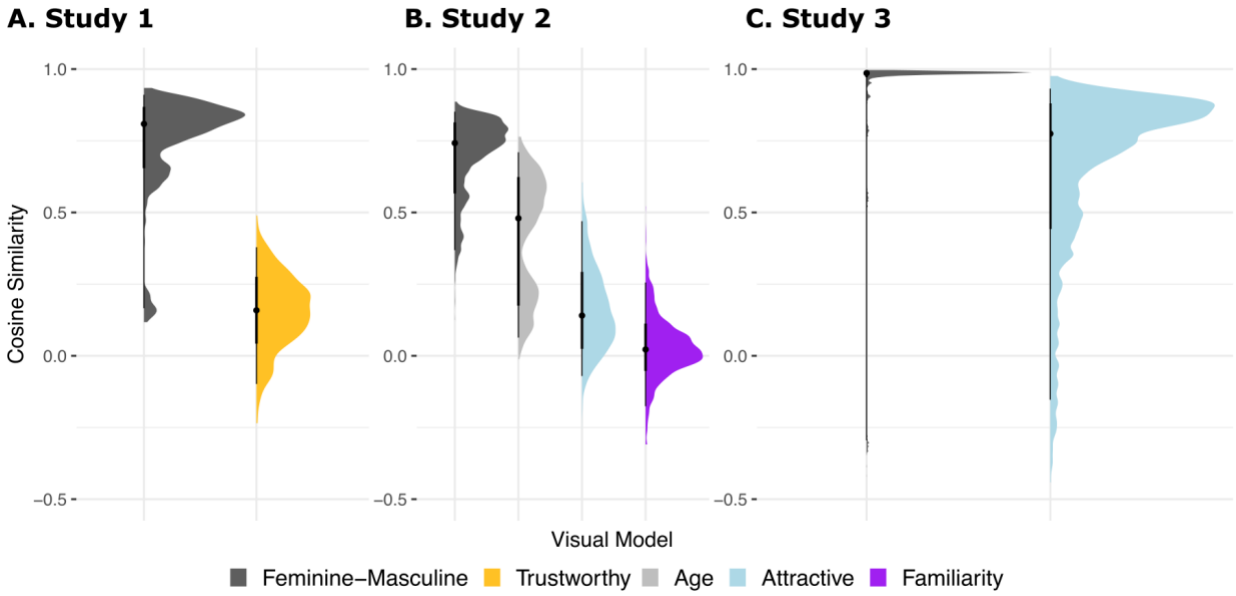

Figure S3. Distribution of the cosine similarities for Studies 1, 2, and 3 (Panels A - C, respectively). The cosine similarity (y axis) was calculated by taking the average similarity of each participant's idiosyncratic visual model and every other participant's visual model within a particular judgment category (x-axis; colored distributions). Across all studies, we predicted that the similarity for highly shared judgments (e.g., feminine-masculine, age) would be larger than highly idiosyncratic judgments (e.g., trustworthy, attractive). Error bars represent 95% confidence intervals.

Correlation with Test-Retest Reliabilities

Study 1. Cosine similarity was correlated with participants' test-retest reliability for feminine-masculine visual models, $r(33) = .76$, $p < .001$, but not for trustworthy-untrustworthy visual models, $r(28) = -.25$, $p = .18$, suggesting that whereas for highly shared visual models, differences from the average can be partially explained by noise (i.e., noisy, less reliable participants), for idiosyncratic models these differences reflect genuine idiosyncratic differences.

Study 2. Unlike Study 1, cosine similarity was correlated with participants' test-retest reliability for both highly shared visual models, $r(113) = .75$, $p < .001$ and highly idiosyncratic visual models, $r(95) = .54$, $p < .001$. While the correlation was significant for both types of judgments, the correlation for highly shared visual models was significantly larger than that of highly idiosyncratic visual models, $z = 2.76$, $p = .006$. Thus, this result theoretically replicates Study 1 and suggests that for highly shared visual models, differences from the average visualization can be partially explained by noise.

Study 3. Cosine similarity scores were correlated with participants' test-retest reliability for both feminine-masculine visual models ($r(56) = .50$, $p < .001$), as well as attractiveness visual models ($r(54) = .59$, $p < .001$). These correlations were not significantly different from one another, $z = -0.64$, $p = .518$. While the highly shared visual models do not show a more significant correlation

with test-retest reliability (i.e., theoretically replicating Studies 1 and 2), this may be due to the greater range of cosine values across participants in this sample (see distribution of scores in Fig. S4). One potential explanation for the increase in cosine similarity scores in both feminine-masculine and attractiveness visual models may be the new latent space used in this study. However, future work is needed to thoroughly examine this explanation.

Study 4. Like the previous studies, the average cosine similarity of each participant's idiosyncratic visual model was significantly correlated with their test-retest correlation (car: $r(43) = .31$, $p = .045$; child: $r(41) = .36$, $p = .018$; money: $r(29) = .39$, $p = .031$).

Study 4: Similarity Between Idiosyncratic Visual Models

While we did not have any specific predictions about the average similarities between context-dependent visual models of trustworthiness, we still computed similarity scores to assess whether the visual models were significantly different from one another (Fig. S5). The average cosine similarity of "trust to fix your car" visual models was significantly lower than both "trust to watch your child" visual models ($t(116) = 13.35$, $p < .001$, d = 2.85) and "trust to invest your money" visual models ($t(116) = 3.67$, $p < .001$, $d = 0.85$). Similarly, the "trust to watch your child" visual models were significantly higher than the "trust to invest your money" visual models, $t(116) = 8.45$, $p < .001$, $d = 1.99$.

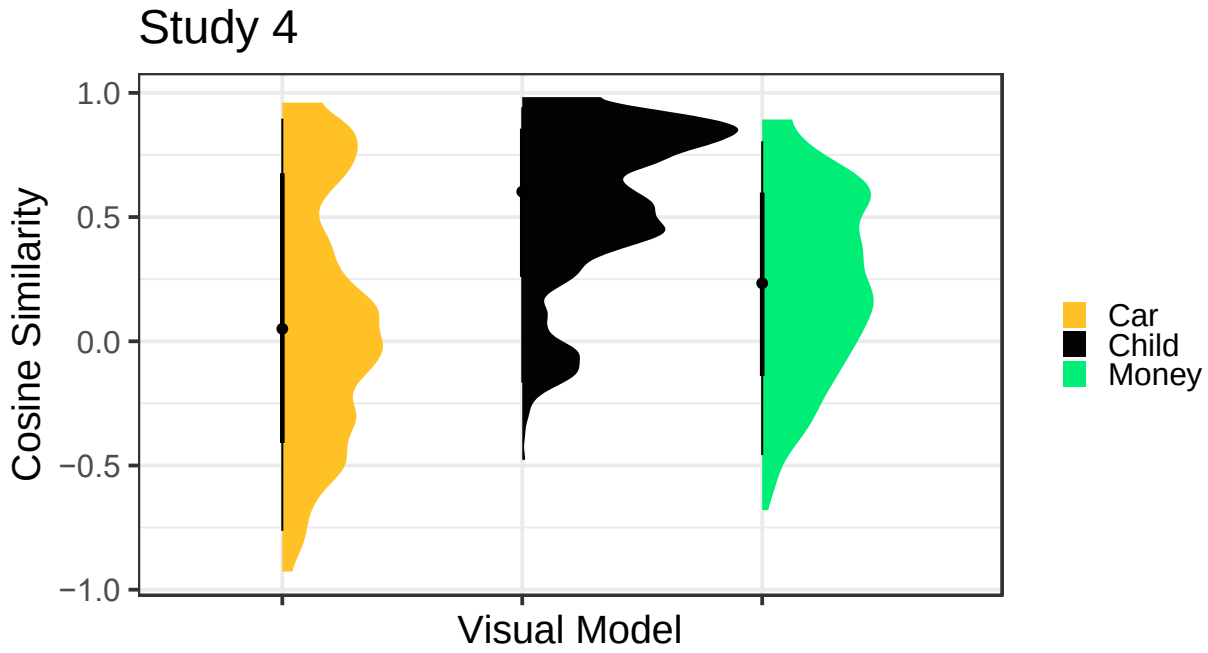

Fig. S4. Cosine similarity between all three context-dependent trustworthiness conditions in Study 4. The cosine similarity (y axis) was calculated by taking the average similarity of each participant's idiosyncratic visual model and every other participant's visual model within a particular judgment category (x-axis; colored distributions).

Study 1 Additional Results

Full Linear Mixed-Effects Models for each Judgment in Phase II (Stimulus Ratings)

	Ratings of Masculinity				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	3.93	3.73 – 4.13	38.96	**<0.001**	181.69
Visual Model	0.83	0.61 – 1.05	7.39	**<0.001**	255.89
Model Value	0.63	0.58 – 0.67	25.95	**<0.001**	258.20
Visual Model × Model Value	-0.96	-1.03 – -0.89	-27.14	**<0.001**	256.14
Random Effects					
σ^2	1.16				
$\tau_{00\ face}$	0.75				
$\tau_{00\ participant}$	0.22				
ICC	0.45				
$N_{participant}$	50				
N_{face}	260				
Observations	5140				
Marginal R^2 / Conditional R^2	0.562 / 0.761				

Table S1. Linear-mixed effects regression table for Study 1 "masculinity" ratings. The "Visual Model" variable compared images generated from the "feminine-masculine" visual model to images generated from the "trustworthiness" visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -4 to +4 in increments of two and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

	Ratings of Trustworthiness				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	4.65	4.42 – 4.88	40.09	<**0.001**	49.57
Visual Model	-0.46	-0.56 – -0.36	-9.08	<**0.001**	260.56
Model Value	-0.11	-0.13 – -0.09	-9.98	<**0.001**	268.64
Visual Model × Model Value	0.32	0.29 – 0.35	19.94	<**0.001**	260.01
Random Effects					
σ^2	1.05				
$\tau_{00\ face}$	0.10				
$\tau_{00\ participant}$	0.54				
ICC	0.38				
$N_{participant}$	44				
N_{face}	260				
Observations	4546				
Marginal R^2 / Conditional R^2	0.160 / 0.478				

Table S2. Linear-mixed effects regression table for Study 1 trustworthiness ratings. The "Visual Model" variable compared images generated from the "feminine-masculine" visual model to images generated from the "trustworthiness" visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -4 to +4 in increments of two and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

Study 2 Additional Results

Full Linear Mixed-Effects Models for each Judgment in Phase II (Stimulus Ratings)

	Ratings of "Old" (Age)				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	3.97	3.67 – 4.26	27.03	**<0.001**	75.05
Visual Model	0.01	-0.22 – 0.24	0.07	0.948	50.08
Model Value	0.48	0.42 – 0.54	16.53	**<0.001**	50.18
Visual Model × Model Value	-0.05	-0.12 – 0.01	-1.59	0.117	50.14
Random Effects					
σ^2	0.99				
$\tau_{00\ image}$	0.07				
$\tau_{00\ participant}$	0.36				
ICC	0.30				
$N_{\ participant}$	34				
$N_{\ image}$	54				
Observations	1870				
Marginal R^2 / Conditional R^2	0.633 / 0.745				

Table S3. Linear-mixed effects regression table for Study 2 "agedness" ratings. The "Visual Model" predictor variable compares images generated from other participants' visual models to images generated from participants' own visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -6 to +6 in increments of two and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

	Ratings of Masculinity				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	3.98	3.52 – 4.45	16.70	**<0.001**	55.02
Visual Model	0.02	-0.47 – 0.52	0.10	0.923	50.00
Model Value	0.55	0.43 – 0.68	8.47	**<0.001**	50.00
Idiosyncratic × Model Value	0.00	-0.14 – 0.14	0.05	0.958	50.00
Random Effects					
σ^2	0.66				
$\tau_{00\ image}$	0.47				
$\tau_{00\ participant}$	0.09				
ICC	0.46				
$N_{\ participant}$	33				
$N_{\ image}$	54				
Observations	1836				
Marginal R^2 / Conditional R^2	0.763 / 0.872				

Table S4. Linear-mixed effects regression table for Study 2 "masculinity" ratings. The "Visual Model" predictor variable compares images generated from other participants' visual models to images generated from participants' own visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -6 to +6 in increments of two and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

	Ratings of Attractiveness				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	3.95	3.53 – 4.37	18.52	<**0.001**	35.38
Visual Model	-0.19	-0.41 – 0.03	-1.71	0.087	49.83
Model Value	0.45	0.40 – 0.51	15.91	<**0.001**	49.31
Visual Model × Model Value	-0.18	-0.24 – -0.12	-5.76	<**0.001**	49.38
Random Effects					
σ^2	1.20				
$\tau_{00\ image}$	0.04				
$\tau_{00\ participant}$	0.85				
ICC	0.43				
$N_{\ participant}$	24				
$N_{\ image}$	54				
Observations	1322				
Marginal R^2 / Conditional R^2	0.372 / 0.639				

Table S5. Linear-mixed effects regression table for Study 2 "attractiveness" ratings. The "Visual Model" predictor variable compares images generated from other participants' visual models to images generated from participants' own visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -6 to +6 in increments of two and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

	Ratings of Familiarity				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	4.39	3.68 – 5.11	12.06	<**0.001**	23.27
Visual Model	-0.18	-0.47 – 0.11	-1.20	0.230	50.00
Model Value	0.21	0.14 – 0.29	5.57	<**0.001**	50.00
Visual Model × Model Value	-0.20	-0.28 – -0.12	-4.86	<**0.001**	50.00
Random Effects					
σ^2	1.84				
$\tau_{00\ image}$	0.08				
$\tau_{00\ participant}$	2.17				
ICC	0.55				
$N_{participant}$	19				
N_{image}	54				
Observations	1134				
Marginal R^2 / Conditional R^2	0.024 / 0.561				

Table S6. Linear-mixed effects regression table for Study 2 "familiarity" ratings. The "Visual Model" predictor variable compares images generated from other participants' visual models to images generated from participants' own visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -6 to +6 in increments of two and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

Optimal Number of Trials

The number of experimental trials used in reverse correlation studies varies widely between different methodologies. For example, early visual psychophysical reverse correlation studies used up to 20,000 trials per participant (e.g., *68*). However, many recent studies using noise-based reverse correlation have used a minimum of 300 trials (e.g., *37*, *38*). The dramatic reduction in experimental trials was possible due to the addition of a "base image" added underneath the randomly generated noise, which acted as a template or guide for participants. We used this prior work as a starting point in our past proof-of-concept work (*9*), but were interested in whether more or less trials were needed to obtain quality results.

To test this, we had participants in Study 2 complete 1000 trials in chunks of 100 randomized faces. Chunking trials in groups of 100 faces allowed for us to measure the quality of the images produced from an individual's model across each group of 100 stimuli (e.g., after 100 trials, 200 trials, etc.). To compare these models and images, we computed the cosine similarity on both the face *image* latents and the *idiosyncratic visual model* latents after each block of 100 trials. The comparison for each cosine similarity score was the participant's final latent vector produced by using data from all 1000 trials.

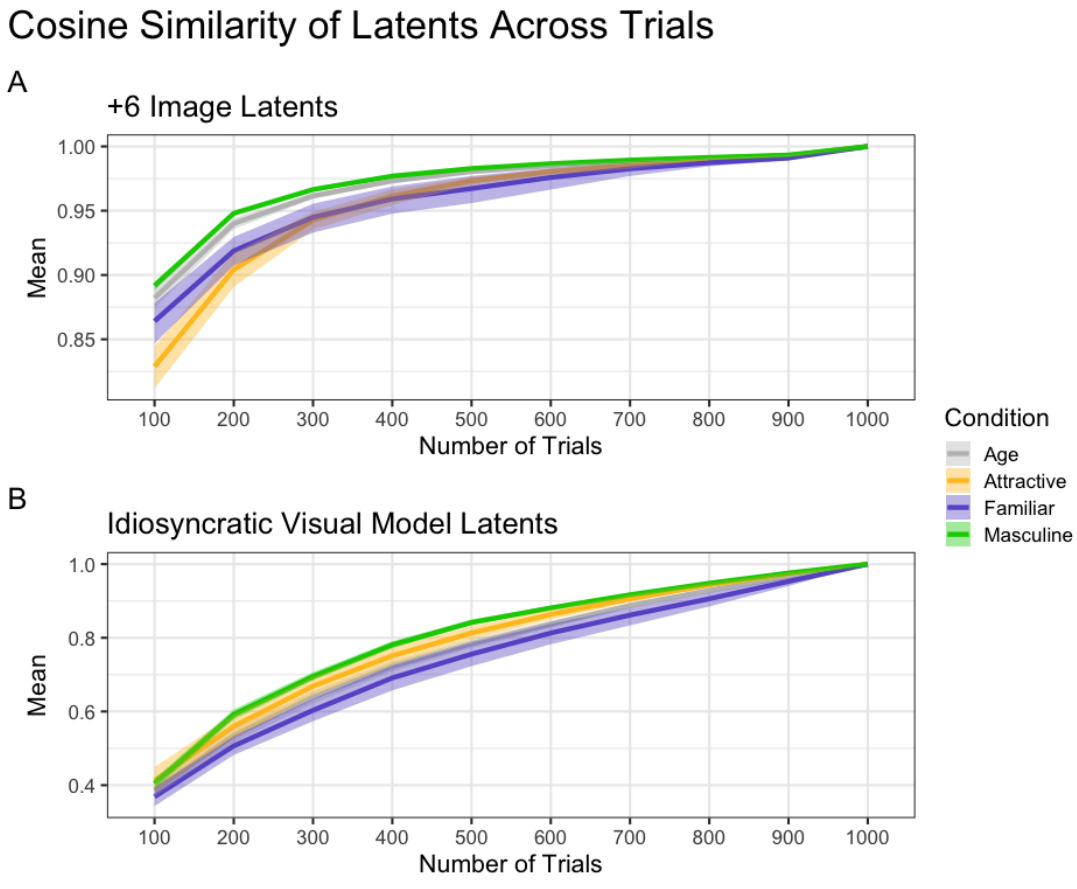

Fig. S5. Comparison of cosine similarities between each condition in Study 2. We compared the latent vectors at each 100 trials to the latent vectors constructed using all 1000 trials in

Study 2. Panel "A" displays the cosine similarity for +6 image latents. Panel "B" displays the cosine similarity of each idiosyncratic visual model.

Visual inspection of the results showed that the cosine similarity of the +6 image latents (Fig. S6A) generally plateau between 300 and 400 trials. However, the cosine similarity of the idiosyncratic model latents (Fig. S6B) increases steadily across all trials. The latter result is not entirely unexpected given how each latent vector is computed. The idiosyncratic model latents are more sensitive to trial-by-trial changes as more novel data is added after each chunk of 100 trials. Thus, the vectors become increasingly more similar to the final vector which includes all data points. On the other hand, the images produced from the +6 image latents vectors are more stable since the vector representing the average of all participants' "neutral" selections are added at each step (refer to equations, above).

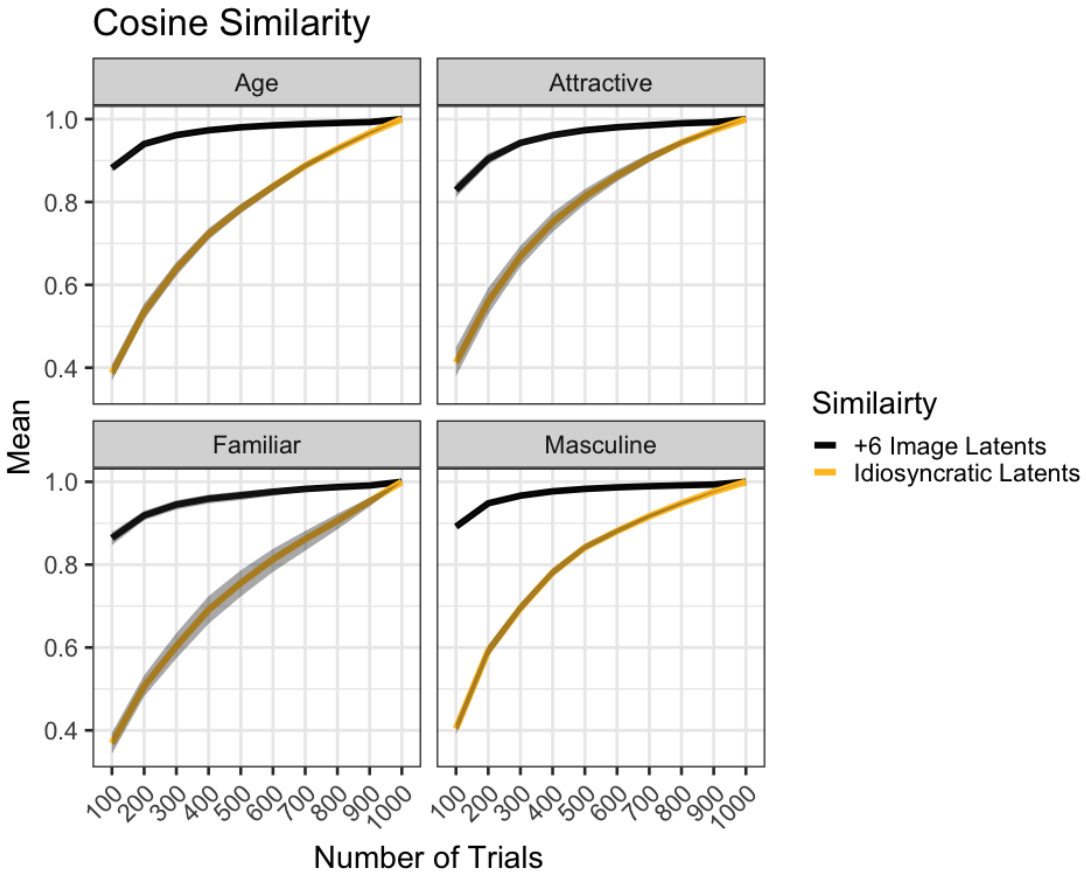

Fig. S6. Direct comparison of the cosine similarities between the latent vectors of the +6 images (black lines) and the idiosyncratic visual models (gold lines) across all 1000 trials. We compared the latent vectors after each group of 100 trials. The comparison was the latent vectors constructed using data from all 1000 trials. Each panel displays the comparison for one of the judgments in Study 2. Shaded areas represent 95% confidence intervals.

Directly comparing the cosine similarity of the idiosyncratic model latents and the +6 image latents across each judgment (Fig. S7), it appears that 300 trials is adequate for obtaining stable visual representations from participants' idiosyncratic models. In other words, a minimum of 300

experimental trials (i.e., 300 stimulus categorizations) is required for typical generative reverse correlation studies if the primary goal is to visualize and compute idiosyncratic representations of the social judgment across model values.

Study 3 Additional Results

Full Linear Mixed-Effects Models for each Judgment in Phase II (Stimulus Ratings)

	Ratings of Masculinity				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	3.76	3.06 – 4.45	11.11	**<0.001**	27.41
Visual Model	0.06	-0.70 – 0.81	0.15	0.879	26.00
Model Value	1.47	0.98 – 1.95	6.22	**<0.001**	26.00
Visual Model × Model Value	-0.06	-0.59 – 0.48	-0.22	0.831	26.00
Random Effects					
σ^2	1.10				
$\tau_{00\ participant2}$	0.14				
$\tau_{00\ face}$	0.53				
ICC	0.38				
$N_{participant}$	45				
N_{face}	30				
Observations	1350				
Marginal R^2 / Conditional R^2	0.695 / 0.810				

Table S7. Linear-mixed effects regression table for the Supplemental Study "masculinity" ratings. The "Visual Model" predictor variable compares images generated from other participants' visual models to images generated from participants' own visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -2 to +2 in increments of one and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

	Ratings of Attractiveness				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	4.27	3.90 – 4.65	22.58	**<0.001**	62.84
Visual Model	-0.02	-0.33 – 0.30	-0.11	0.911	26.00
Model Value	1.09	0.89 – 1.30	11.15	**<0.001**	26.00
Visual Model × Model Value	-0.28	-0.50 – -0.06	-2.59	**0.015**	26.00
Random Effects					
σ^2	1.50				
$\tau_{00\ participant}$	0.85				
$\tau_{00\ face}$	0.07				
ICC	0.38				
$N_{\ participant}$	51				
$N_{\ face}$	30				
Observations	1530				
Marginal R^2 / Conditional R^2	0.385 / 0.618				

Table S8. Linear-mixed effects regression table for the Supplemental Study "attractiveness" ratings. The "Visual Model" predictor variable compares images generated from other participants' visual models to images generated from participants' own visual model. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -2 to +2 in increments of one and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

Study 4 Additional Results

Ratings of "Trust to Fix Your Car" Images

	Ratings of "Trust to Fix Your Car" Images				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	4.15	3.92 – 4.38	35.03	**<0.001**	334.60
Visual Model 1 [Car vs. Child]	-0.45	-0.76 – -0.14	-2.84	**0.005**	258.77
Visual Model 2 [Car vs. Money]	-0.24	-0.55 – 0.07	-1.54	0.126	218.38
Model Value	0.23	0.18 – 0.28	8.53	**<0.001**	479.07
Visual Model 1 × Model Value	-0.32	-0.38 – -0.27	-11.57	**<0.001**	5186.95
Visual Model 2 × Model Value	-0.11	-0.17 – -0.06	-3.97	**<0.001**	4617.50
Random Effects					
σ^2	1.46				
$\tau_{00\ participant}$	0.81				
$\tau_{00\ face}$	0.13				
ICC	0.39				
$N_{participant}$	221				
N_{face}	180				
Observations	4913				
Marginal R^2 / Conditional R^2	0.037 / 0.414				

Table S9. Linear-mixed effects regression table for Study 3 "trust to fix your car" ratings. The "Visual Model 1" predictor variable compares images generated from "trust to fix your car" visual models to images generated from "trust to watch your child" visual models. The "Visual Model 2" predictor variable compares images generated from "trust to fix your car" visual models to images generated from "trust to invest your money" visual models. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -2 to +2 in increments of one and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

Ratings of "Trust to Watch Your Child" Images

	Ratings of "Trust to Watch Your Child" Images				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	3.68	3.47 – 3.89	34.01	**<0.001**	243.13
Visual Model 1 [Child vs. Car]	0.43	0.12 – 0.73	2.78	**0.006**	217.86
Visual Model 2 [Child vs. Money]	0.22	-0.07 – 0.51	1.46	0.145	217.47
Model Value	0.58	0.54 – 0.63	24.18	**<0.001**	392.39
Visual Model 1 × Model Value	-0.97	-1.02 – -0.92	-37.41	**<0.001**	4546.11
Visual Model 2 × Model Value	-0.71	-0.76 – -0.66	-28.54	**<0.001**	4540.58
Random Effects					
σ^2	1.23				
$\tau_{00\ participant}$	0.79				
$\tau_{00\ face}$	0.12				
ICC	0.42				
$N_{\ participant}$	221				
$N_{\ face}$	172				
Observations	4844				
Marginal R^2 / Conditional R^2	0.176 / 0.525				

Table S10. Linear-mixed effects regression table for Study 3 "trust to watch your child" ratings. The "Visual Model 1" predictor variable compares images generated from "trust to watch your child" visual models to images generated from "trust to watch fix your car" visual models. The "Visual Model 2" predictor variable compares images generated from "trust watch your child" visual models to images generated from "trust to invest your money" visual models. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -2 to +2 in increments of one and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

Ratings of "Trust to Invest Your Money" Images

	Ratings of "Trust to Invest Your Money"				
Predictors	*Estimates*	*CI*	*Statistic*	*p*	*df*
(Intercept)	4.08	3.86 – 4.30	36.22	**<0.001**	248.31
Visual Model 1 [Money vs. Child]	-0.33	-0.63 – -0.03	-2.14	**0.034**	217.45
Visual Model 2 [Money vs. Car]	-0.12	-0.43 – 0.20	-0.73	0.463	216.60
Model Value	0.43	0.37 – 0.49	14.46	**<0.001**	281.61
Visual Model 1 × Model Value	-0.55	-0.61 – -0.49	-17.04	**<0.001**	3256.21
Visual Model 2 × Model Value	-0.29	-0.36 – -0.23	-8.88	**<0.001**	3241.00
Random Effects					
σ^2	1.44				
$\tau_{00\ participant}$	0.82				
$\tau_{00\ face}$	0.12				
ICC	0.40				
$N_{\ participant}$	221				
$N_{\ face}$	124				
Observations	3503				
Marginal R^2 / Conditional R^2	0.082 / 0.445				

Table S11. Linear-mixed effects regression table for Study 3 "trust to invest your money" ratings. The "Visual Model 1" predictor variable compares images generated from "trust to invest your money" visual models to images generated from "trust to watch your child" visual models. The "Visual Model 2" predictor variable compares images generated from "trust invest your money" visual models to images generated from "trust to fix your car" visual models. The "Model Value" variable (i.e., the image linear interpolation value) ranged from -2 to +2 in increments of one and was treated as a continuous predictor. The model included random intercepts for each participant and face image.

Exploring Methods for Computing Idiosyncratic Visual Models

As noted above ("Observations and Comparison to Projection Methods" subsection) and in the main text, many of the images generated from the idiosyncratic visual models in Studies 3 and 4 quickly go out of sample at extreme values. This is likely due to the manner in which we generated the stimulus images in these studies. Unlike Studies 1 and 2 where we projected real neutral faces into a pretrained latent space, Studies 3 and 4 sampled directly from the latent space of a new model (see section "Training a New Generative Model"). The former approach artificially shrinks the latent space and is bounded by the neutral faces projected into it, resulting in an effectively smaller latent space to operate in (i.e., to generate the idiosyncratic mental prototype images). In contrast, the latter approach utilizes the whole latent space (if face stimuli are randomly sampled from the latent distribution of the generative model). When each idiosyncratic visual model is multiplied by a constant, the steps between +1 and +2, for example, will be larger for the unconstrained latent space than the artificially constrained latent space. In other words, the length of the computed idiosyncratic visual model vectors when using the entire latent space will be longer than those computed from the artificially restricted latent space.

One way to handle the differences in vector lengths both between different latent spaces and individual visual models is to normalize each idiosyncratic visual model vector before visualizing it such that,

$$\hat{A}_i = \bar{A}_i - \bar{B}_i$$

becomes,

$$\hat{A}_i = \frac{(\bar{A}_i - \bar{B}_i)}{||(\bar{A}_i - \bar{B}_i)||}$$

Likewise, the average "neutral" vector is also normalized,

$$\widehat{N}_i = \frac{\bar{N}_i}{||\bar{N}_i||}$$

Normalizing the idiosyncratic visual model essentially standardizes vector lengths, resulting in an increase of C standard deviations when the vector is multiplied by a constant, C (refer to "Idiosyncratic Visual Model Construction and Visualization" section above for details). It is important to note that the idiosyncratic model vector must first be "unnormalized" before passing it back to the model for visualization.

Fig.* S7. *Example feminine-masculine images generated from a Study 3 participant's idiosyncratic model using non-normalized vectors (top) and normalized vectors (bottom). The center image corresponds to the average of all images selected as "neutral" and each image to the left and right of the center was generated by multiplying the idiosyncratic visual model by a constant up to -/+8 (unnormalized vector; increments of 2 from left to right) or from -/+ 2 (normalized vector; increments of 0.5 from left to right).

In our tests, we observed that this approach works well when applied to participants' visual models computed from latent vectors generated using the full latent space of a generative model, such as in Studies 3 and 4 (Fig. S7). Normalizing the vectors appears to work well up to about +/- 2 or +/- 3 SDs in our sample of participants. Note also how the normalized vectors don't dramatically change the visual models, only how many "steps" it takes to achieve similar visualizations.

We similarly observed that the normalizing process works when applied to idiosyncratic visual models computed using an artificially restricted latent space, such as in Studies 1 and 2 (Fig. S8).

Fig.* S8. *Example trustworthy-untrustworthy images generated from a Study 1 participant's idiosyncratic model using non-normalized vectors (top) and normalized vectors (bottom). The center image corresponds to the average of all images selected as "neutral" and each image to the left and right of the center was generated by multiplying the idiosyncratic visual model by a constant up to -/+8 for unnormalized vector images (by increments of 2) and -/+2 for the normalized vector images (by increments of 0.5).

Given these two results, it is clear that vector normalization stabilizes visualizations generated from idiosyncratic models. Visualizations of idiosyncratic models across both the full latent and artificially constrained latent spaces show that stable results can be achieved within -/+ 2 SDs. Anything more than this results in visualizations that quickly go out of sample or approach caricature representations.

Group-Level and Individual Model Visualizations

Every group-level and individual participant visualization can be viewed online at this study's repository (https://osf.io/aqgfw/) or directly at https://osf.io/g2xdp (separate file).