Supplementary Method

Pilot experiment to evaluate stimuli

Participants

Fourteen participants completed the pilot experiment (8 female, mean age = 25.14, SD = 9.52). All participants received course credit, had normal or corrected-to-normal vision and gave informed consent according to the local ethics guidelines.

Design

We piloted the stimuli using two self-paced tasks, a two-alternative forced-choice task (2AFC) and a ratings task. Participants completed the 2AFC task before the ratings task. Across participants, 16 relevant statements from the mini-IPIP (4 per trait) and all 434 face images were used.

Tasks

2AFC. On each trial, a Low and High trait image on the same ID were presented side by side. Trials were presented in blocks with a trait statement and task instructions presented at the beginning of each block (Supplementary Figure 1). Blocks included 26, 27 or 28 trials depending on the number of stimuli available per trait category and blocks were randomly ordered. Each ID pair was presented twice throughout the experiment, so that two trait relevant statements could be presented per ID per participant. Half of the participants were shown statements from list 1, the other half were shown statements from list 2 (Supplementary Table 2). On each trial, the choice of which stimulus matched the statement was recorded. All 217 ID pairs were presented, each with 2 statements, which made 434 trials in total. For each trait type, a one-tailed one-sample t-test was performed to test if judgments were greater than chance performance (50%) and thereby confirm if stimuli were perceived in

a manner that we expected. Cohen's d_z was also calculated as a standardised measure of effect size.

Ratings. On each trial, a single face image was presented in the centre of the screen and participants were asked to make a judgement based on a statement presented at the start of the block. Trials were presented in blocks with a trait statement and task instructions presented at the beginning of each block (Supplementary Figure 1). Participants responded on a keyboard number pad using the numbers one to nine, where one indicated that the statement presented at the beginning of the block suited the face "not at all well" and nine indicated that it fitted "extremely well". Half of the participants saw the first half of the stimulus IDs warped to a High trait dimension and the second half of the IDs warped to a Low trait dimension. The remaining participants saw the reverse. In addition, the first half of these participant groups were asked to rate faces according to statement list 1, whereas the second half of the participant groups were asked to rate faces according to statement list 2. This resulted in 4 different orders so that one participant sees high with first statement list, next sees low with first statement list, third sees high with second statement list, fourth sees low with second statement list. Thus, across participants, all face IDs were rated on each relevant statement from the mini-IPIP when warped to high or to low of the relevant trait composite. In addition, no participant saw the same ID warped to high in the ratings section if they had seen it warped to low in the ratings section, and vice versa. Each participant was shown each exemplar twice with a different trait statement each time, which produced 434 trials in total. Ratings for high and low faces for each trait were compared using a one-tailed paired samples t-test and Cohen's d_z as a measure of effect size.

Results

2AFC. Except for the agreeableness pairs, all other pairs were discriminated correctly at a level above chance (Supplementary Figure 2A; Supplementary Table 3). Each trait is compared to chance performance (50%) with a one-tailed, one-sample t-test and Cohen's d_z: Extraversion $t(13)= 7.95$, $p<.001$, $d_z = 2.13$; Agreeableness $t(13)= -0.86$, $p=.797$, $dz = -0.23$; Neuroticism $t(13)= 5.53$, $p<.001$, $d_z = 1.48$; Physical Health $t(13)= 3.92$, $p<.001$, $d_z = 1.05$.

Ratings. Except for the agreeableness stimuli, all other high/low trait pairs were rated significantly different from each other, with the high warp rated higher in each trait than the low warp (Supplementary Figure 2B; Supplementary Table 3). Ratings for high and low faces for each trait were compared using a one-tailed paired samples t-test and Cohen's dz: Extraversion $t(13)= 4.24$, $p<.001$, $d_z = 1.13$; Agreeableness $t(13)= -0.98$, $p=.826$, $d_z = -0.26$; Neuroticism $t(13)= 5.22$, $p<.001$, $d_z = 1.39$; Physical Health $t(13)= 3.04$, $p=.005$, $d_z = 0.81$.

S1 Table. Average self-reported trait scores

	High Composite	Low Composite	Individuals
Extraversion	4.69 [4.61, 4.77]	2.08 [1.90, 2.27]	3.54 [3.40, 3.68]
Agreeableness	4.70 [4.62, 4.77]	2.66 [2.51, 2.81]	3.89 [3.79, 3.99]
Neuroticism	4.51 [4.40, 4.62]	1.73 [1.55, 1.90]	3.10 [2.95, 3.25]
Physical Health	59.95 [59.00, 60.89]	40.61 [38.26, 42.95]	52.88 [51.95, 53.81]

Average self-reported trait scores for individuals included in the high and low composites, as well as individuals before transformation. As should be the case, average scores are different for individuals included in the high and low composites (no overlap of 95%CIs, in square brackets). In addition, average scores for individual ratings that would later be transformed do not overlap with average scores from the high or low composites (no overlap of 95% CIs). The lack of overlap between individual images and composite images suggests that prior to transformation individuals are in a neutral position, not especially skewed towards those included in the high or the low composites images.

S2 Table. Statements used in the pilot experiment.

	Trait Type	Reverse Score
Statement List 1		
More sympathetic	Agreeableness	0
Not interested in other people's problems	Agreeableness	1
Is the life of the party	Extraversion	0
Doesn't talk a lot	Extraversion	1
Has frequent mood swings	Neuroticism	0
Is relaxed most of the time	Neuroticism	1
Health is good	Physical Health	0
Accomplishes less due to health problems	Physical Health	1
Statement List 2		
Feels others' emotions	Agreeableness	0
Not really interested in others	Agreeableness	1
Talks to a lot of different people at parties	Extraversion	0
Keeps in the background	Extraversion	1
Gets upset easily	Neuroticism	0
Seldom feels blue	Neuroticism	1
Finds it easy to climb the stairs	Physical Health	0
Pain interferes more with work	Physical Health	1

Note: For items with a number 1 in the reverse score column, subjects' scores were reversed so that a high score represents high trait representation.

S3 Table. Results from the two-alternative forced-choice task (2AFC) and the ratings task.

	2AFC	Rating high	Rating low
Extraversion	81.08 [73.42, 88.75]	5.58 [5.28, 5.88]	4.68 [4.32, 5.04]
Agreeableness	46.90 [39.82, 53.98]	5.07 [4.83, 5.30]	5.19 [5.01, 5.38]
Neuroticism	76.21 [66.92, 85.50]	5.29 [5.11, 5.48]	4.51 [4.31, 4.70]
Physical Health	71.76 [60.87, 82.65]	5.69 [5.39, 6.00]	5.10 [4.68, 5.52]
Overall	69.16 [65.16, 73.15]	5.41 [5.26, 5.56]	4.86 [4.69, 5.04]

Square brackets = 95% Confidence intervals

S4 Table. Exploratory analysis of wider face perception and theory of mind networks.

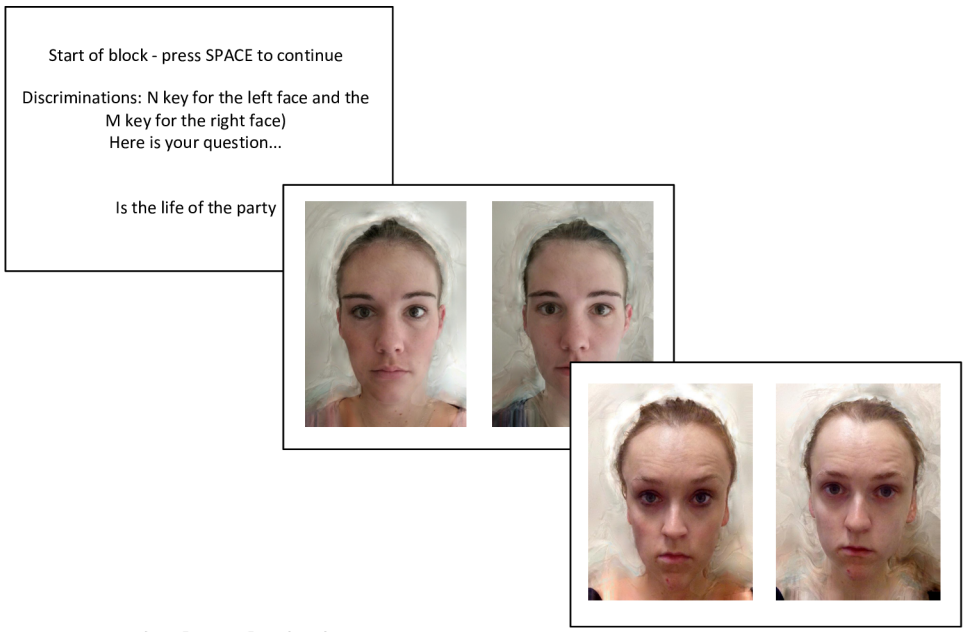
Region				**Novel>Repeated**		
	ROI size (voxels)	**Average localiser mask size (voxels)**	**Inter-subject overlap (%)**	**Percent signal change (SEM)**	**t**	**p(fdr)**
Face localiser						
Right MTG	200	38	82	.135 (.01)	.50	.82
Left OFA	177	32	75	.050 (.21)	.41	.82
Right OT cortex	56	12	54	-.199 (.25)	-.76	.82
Left MTG	141	24	64	.013 (.20)	.07	.82
Left pSTS	140	23	75	-.107 (.19)	-.57	.82
Left FFA	57	9	57	.380 (.22)	1.75	.51
ToM localiser						
Precuneus	870	206	96	-.086 (.14)	-.62	.80
Left TPJ	615	143	100	.129 (.14)	.92	.80
Left ant. temp. cortex	139	27	75	.151 (.11)	1.35	.80
Right MFG	74	13	57	-.049 (.13)	-.37	.80
Left MFG	31	4	54	-.049 (.25)	-.20	.80

Abbreviations: ROI = Region of interest; fdr = false discovery rate; OFA = occipital face area; FFA = right fusiform face area; pSTS = posterior superior temporal sulcus; TPJ = temporoparietal junction; mPFC = medial prefrontal cortex; ant. Temp. = anterior temporal; MTG = middle temporal gyrus; MFG = middle frontal gyrus; OT = occipitotemporal cortex.

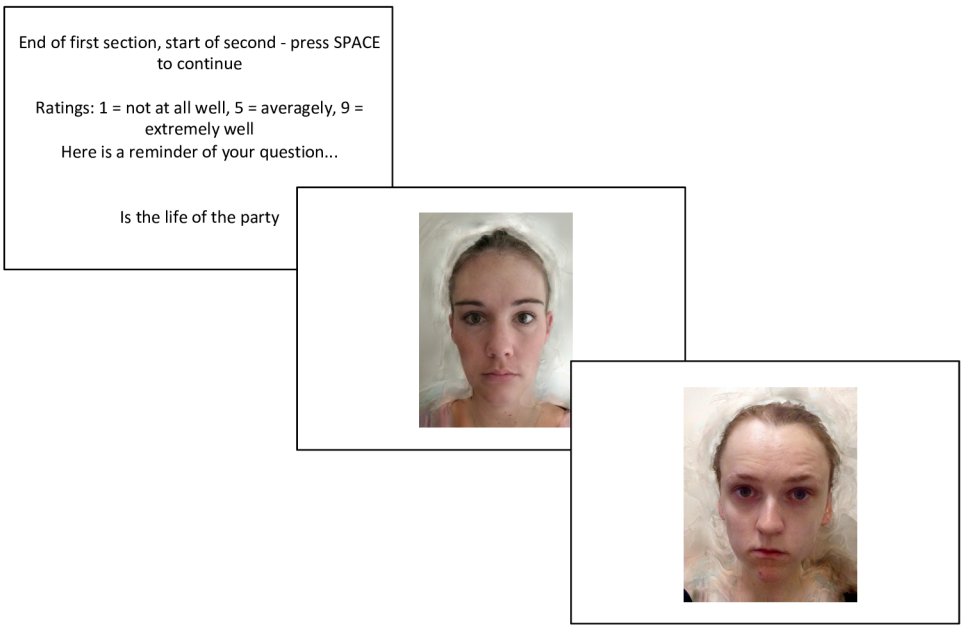
Note: 'ROI size' is the total number of voxels in each ROI based on data from a face perception localiser or a theory-of-mind localiser. 'Average localiser mask size' is the number of voxels that overlap in more than 50% of participants within each ROI. Right MTG, for example, consists of a 200 voxel ROI, with 38 voxels showing overlap in 82% of participants. Analyses were performed on the subset of voxels in each ROI that show overlap in a majority of participants (>50%).

S1 Fig. Methodology for the behavioural pilot experiment.

Example discrimination block beginning

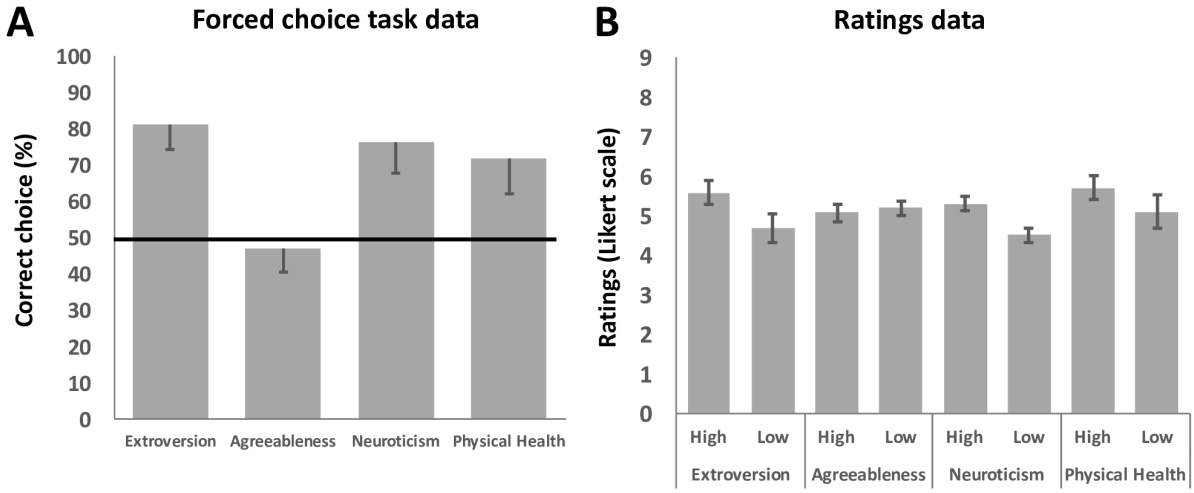


Example rating block beginning



Note: The images used are for illustrative purposes and were not used in the pilot experiment.

S2 Fig. Face rating data for the behavioural pilot experiment.



Face judgment data in a two-alternative forced-choice task (A) and a ratings task (B). The black line at 50% in (A) represents chance performance. Error bars are 95% confidence intervals. One-tailed confidence intervals are displayed in (A) to reflect the one-tailed hypothesis in each comparison.