SUPPLEMENT

Stimulus generalization as a mechanism for learning to trust

SUPPLEMENTAL RESULTS

Experiment 1 Behavioral Results.

Conditioning Phase. Subjects were able to successfully learn which of the three players in the Conditioning Phase could be trusted and which could not (see manuscript). Below we plot the time course across the Conditioning Phase for each player, revealing that subjects began the task entrusting similar amounts of money, and quickly learned to entrust more money in the trustworthy player and less in the untrustworthy player (Fig S1). When we explored response latencies (Table S1A; raw data reported in Table S1B), we observed a main effect of trial type—such that participants responded more quickly as the task progressed. However, we observed no interactive effect between trial and Trust Type, suggesting that people were not faster at learning about an untrustworthy or trustworthy individual.

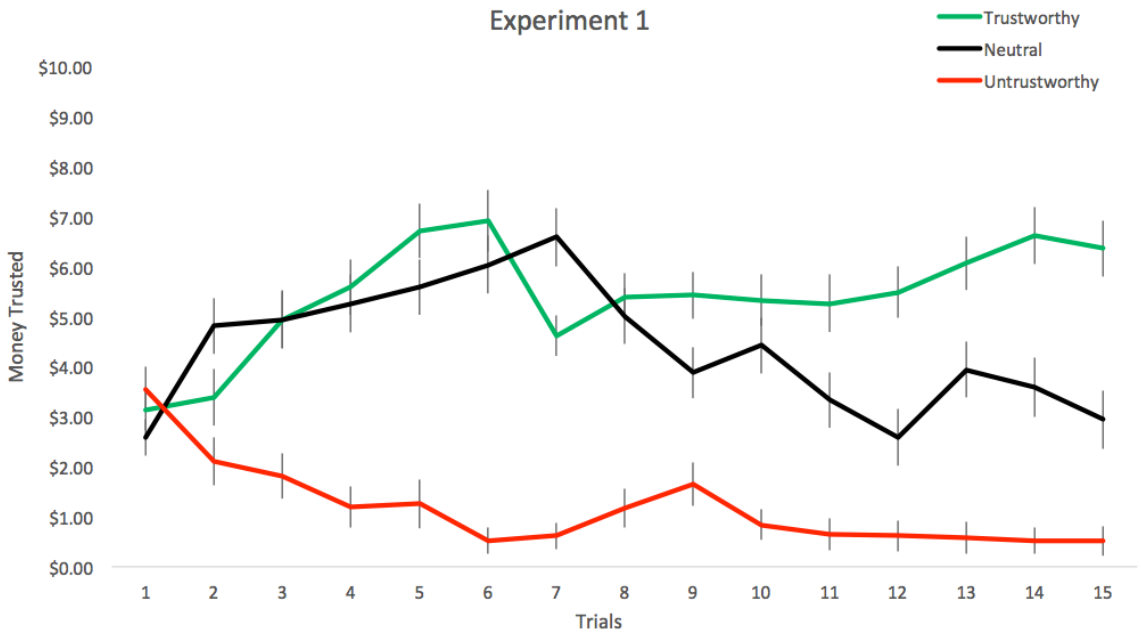

Fig S1 | *Time course data for each player type in the Conditioning Phase in Experiment 1.*

Table S1A Experiment 1:

$Reaction\ Time_{i,t} = \beta_0 + \beta_1 Trial\ Number_{i,t} \times \beta_2\ Trust\ Type_{i,t} + \varepsilon$

DV	***Coefficient*** $(\boldsymbol{\beta})$	***Estimate (SE)***	***t-value***	***P value***
RT	Intercept	*2.14 (.10)*	*20.33*	*<0.001****
	Trial	*-0.02 (.003)*	*-4.31*	*<0.001****
	Trustworthy	*-0.04 (.08)*	*-0.43*	*0.66*
	Untrustworthy	*-0.17 (.09)*	*-2.06*	*0.038**
	Trial × Trustworthy	*0.004 (.003)*	*1.44*	*0.15*
	Trial × Untrustworthy	*-0.0007 (.003)*	*-0.21*	*0.83*

Linear regression where Reaction Time is indexed by subject and trial, and Trust Type is an indicator variable: 0=neutral, -1=untrustworthy, 1=trustworthy; thus neutral serves as the reference category. ***p<0.001, **p<0.01, *p<0.05

Table S1B | Experiment 1

Trust Type	Reaction Times (SD)
Trustworthy	*1.87 (.70)*
Neutral	*1.79 (.69)*
Untrustworthy	*1.51 (.66)*

Experiment 2 Behavioral Results.

Conditioning Phase. In the second experiment, subjects also successfully learned which player could be trusted, as the money entrusted to the trustworthy player ($5.10, SD±1.4) was significantly greater than the amount of money sent to either the neutral ($4.90, SD±1.8) or untrustworthy ($3.14, SD±1.7) players (rmANOVA: $F(2,54)=18.1$, $p<0.001$, $\eta^2=.401$; all pairwise comparisons except trustworthy > neutral: Ps<0.001). The data plotted as a function of trials across the Conditioning Phase for each player reveals that subjects began the task entrusting similar amounts of money, and quickly learned to entrust more money in the trustworthy player and less in the untrustworthy player (Fig S2). Response latencies (Table S2A; raw data reported in Table S2B) analyses replicated the same pattern observed in Experiment 1. There was a main effect of trial type—such that participants responded more quickly across the task, however, there was no interaction between trial and Trust Type.

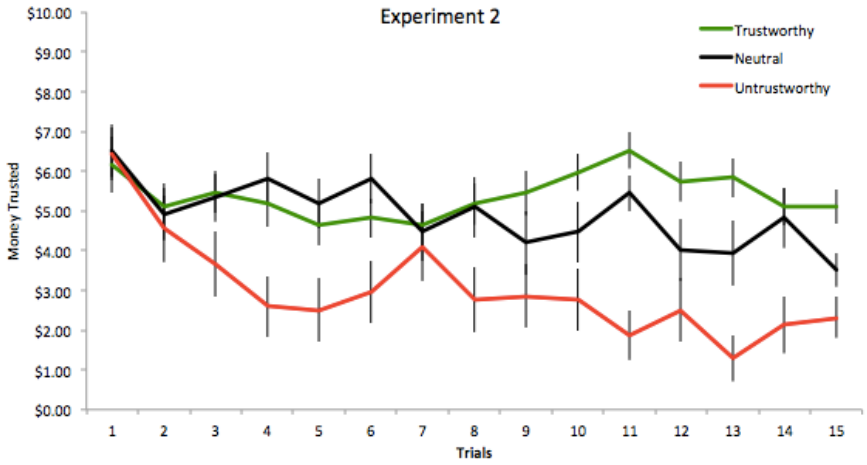

Fig S2 | *Time course data for each player type in the Conditioning Phase in Experiment 2.*

Table S2A Experiment 2:

$Reaction\ Time_{i,t} = \beta_0 + \beta_1 Trial\ Number_{i,t} \times \beta_2\ Trust\ Type_{i,t} + \varepsilon$

DV	***Coefficient*** $(\boldsymbol{\beta})$	***Estimate (SE)***	***t-value***	***P value***
RT	Intercept	*1.67 (.08)*	*19.84*	*<0.001****
	Trial	*-0.007 (.002)*	*-2.49*	*0.01**
	Trustworthy	*-0.06 (.10)*	*-0.55*	*0.58*
	Untrustworthy	*-0.09 (.09)*	*-0.93*	*0.35*
	Trial × Trustworthy	*0.001 (.003)*	*0.29*	*0.77*
	Trial × Untrustworthy	*-0.002 (.004)*	*-0.51*	*0.60*

Linear regression where Reaction Time is indexed by subject and trial, and Trust Type is an indicator variable: 0=neutral, -1=untrustworthy, 1=trustworthy; thus neutral serves as the reference category. ***p<0.001, **p<0.01, *p<0.05

Table S2B | Experiment 2

Trust Type	Reaction Times (SD)
Trustworthy	*1.47 (.67)*
Neutral	*1.52 (.69)*
Untrustworthy	*1.38 (.67)*

Generalization Phase. To probe whether subjects implicitly used perceptual similarity of the morphs to guide novel decisions to trust unfamiliar others in the imaging experiment, we ran a logistic hierarchal regression, where both trustworthiness type (faces morphed with the original trustworthy, untrustworthy, or neutral player), and perceptual similarity (increasing similarity to the original players) were entered as predictors of choosing to play with the morph. We found

that as perceptual resemblance to the original trustworthy player increased, subjects were significantly more likely to choose the morph as a partner for a future Trust Game (Hierarchical Logistic Model where the perceptual similarity variable was mean centered: P<0.001; Table 2).

When we probed whether these generalization gradients evoke structurally similar behavioral tuning profiles, we found that the generalization gradient for untrustworthy morphs was wider than the trustworthy morphs (the slope of the trustworthy and untrustworthy gradients' coefficients from the regression (Table 2) were significantly different from one another: (t(27)=-7.77, p<0.001). This was further confirmed with a rmANOVA Trust Type X Morph interaction $F(10,270)=11.21$, $p<0.001$, $\eta^2=.29$: post hoc t-tests of trustworthy and untrustworthy morph increments against neutral morph increments reveal the untrustworthy morph with 56% similarity is significantly different than the neutral morph with 56% similarity (t(27)=-2.82, p=0.009); same analysis for neutral morph with 56% similarity compared to trustworthy morph with 56% similarity t(27)=2.29, p=0.03; trustworthy and untrustworthy 67% and 78% morph increments against neutral: all Ps<0.05; Fig S3B). Dovetailing with this finding, we observed that subjects made more adaptive choices in the aversive domain (68.3% choosing not to play with a morph who had any perceptual overlap with the original untrustworthy player) compared to the appetitive domain (61% choosing to play with a morph who had any perceptual overlap with the original trustworthy player).

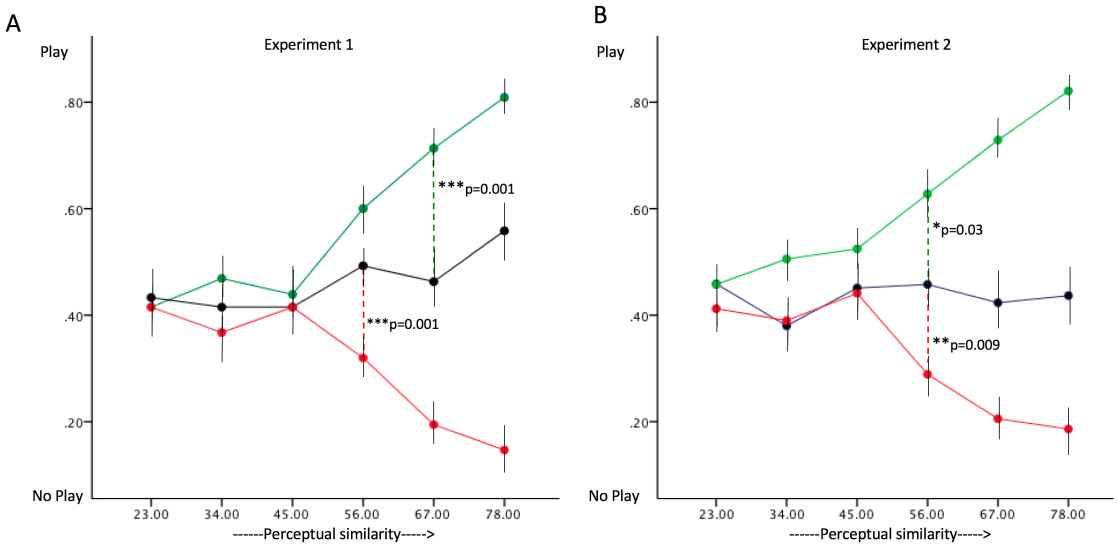

Fig S3 | A) *Raw behavioral data from Experiment 1 reveals asymmetric generalization gradients, such that the untrustworthy gradient is broader and wider than the trustworthy gradient.* ***B)*** *The same pattern of behavior was also observed in the imaging study (Experiment 2).*

Imaging Experiment - Univariate results

Peak voxels reported in the tables were p<0.001 uncorrected, clustered corrected k>20 and are exploratory in nature (univariate analyses), or were *a priori* in nature and thus FWE Bonferroni corrected at p=0.05 for ROI creation (all PSA analyses).

Table S3. Generalization Phase: Face Presentation (parametric untrustworthy gradient > parametric neutral gradient)

Region	***Peak MNI coordinates***			***Z-value***
Amygdala	34	4	-26	3.82
Left AI	-28	36	4	3.48
Left PCC	-14	-42	46	3.68
Right PCC	14	-32	46	3.57
TPJ	48	-34	32	3.36
HPC	-20	-30	8	3.25
Striatum	14	-4	-4	3.27
Caudate	-6	12	6	3.17

Reported at whole brain uncorrected p<0.001.

Table S4. Generalization Phase: Face Presentation (parametric trustworthy gradient > parametric neutral gradient**)**

Region	***Peak MNI coordinates***			***Z-value***
dmPFC	4	42	56	3.24
Anterior TL	-30	-4	-44	3.26
Visual cortex	40	-92	6	3.46

Reported at whole brain uncorrected p<0.001.

ROI Creation

We first ran a conjunction analysis on the three trust types during the Conditioning Phase. ROIs were further selected based on *a priori* hypotheses derived from research on learning in the

nonsocial domain and the extant literature on trust. See manuscript for a more comprehensive explanation of our *a priori* hypotheses.

Table S5. Conditioning Phase:
Conjunction of conditions during presentation of partner's face (trustworthy, untrustworthy, neutral partners)

Region	*Peak MNI coordinates*			*Z-value*
***Left Amygdala**	*-24*	*0*	*-12*	*5.23*
***vmPFC**	*14*	*56*	*-22*	*4.49*
***Right caudate**	*10*	*-6*	*2*	*5.26*
***Right AI**	*38*	*26*	*6*	*6.62*
***Left AI**	*-34*	*18*	*2*	*5.68*
***R Ventral Striatum**	*20*	*12*	*-10*	*5.53*
***L Ventral Striatum**	*-20*	*10*	*-10*	*5.14*
***Right fusiform**	*40*	*-54*	*-18*	*7.19*
***Left Fusiform**	*-40*	*-54*	*-18*	*7.18*
***PCC**	*-2*	*-30*	*26*	*5.90*
***dACC**	*6*	*22*	*30*	*5.68*
***L Hippocampus**	*-20*	*-28*	*-6*	*6.69*
***R Hippocampus**	*24*	*-30*	*-4*	*6.51*
***VTA**	*6*	*-20*	*-14*	*5.62*
***dlPFC**	*-54*	*4*	*46*	*6.14*
***Lateral PFC**	*48*	*44*	*12*	*5.58*

*Reported at FWE corrected p<0.05, *denotes ROIs used for analysis*

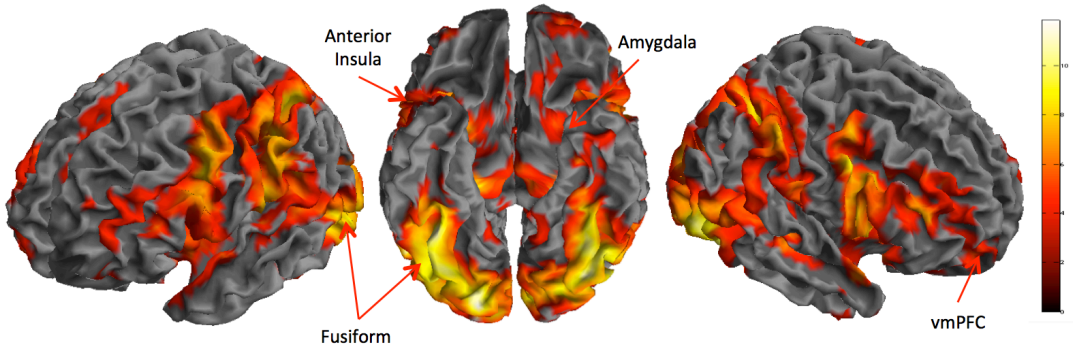

FIG S4| *Conjunction of face presentation for all trust types during learning (Table S5)*

Table S6. Conditioning Phase: Choice epoch (untrustworthy > neutral partner)				
Region	***Peak MNI coordinates***			***Z-value***
Insula	50	-2	-2	4.24
vmPFC/ACC	-8	54	8	4.04
dACC	-2	-6	40	3.40
ATL	-52	4	-34	4.04
TPJ	54	-24	20	4.60

Reported at whole brain uncorrected p<0.001.

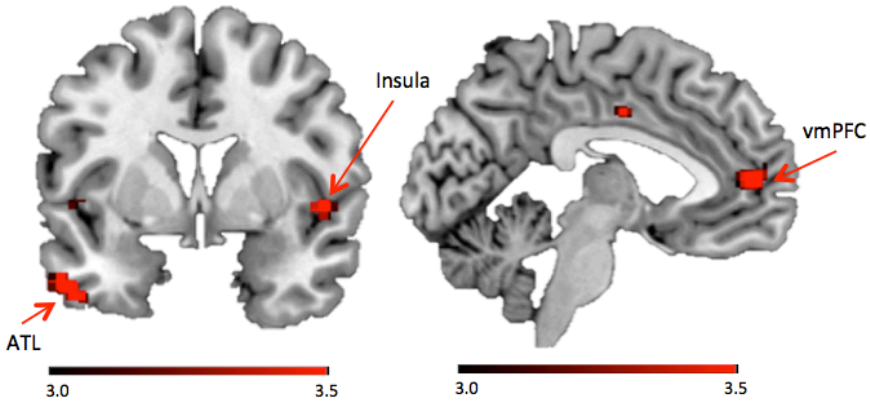

FIG S5 | *Learning about an Untrustworthy Partner: Table S6*

Table S7. Conditioning Task Choice epoch (trustworthy > neutral partner)				
Region	***Peak MNI coordinates***			***Z-value***
PCC	10	-18	36	3.76

ATL	-46	18	-40	3.98
Amygdala	22	0	-26	3.48
dlPFC	56	10	12	3.28

Reported at whole brain uncorrected p<0.001.

Imaging Experiment - Multivariate results

For the pattern similarity (PS) analyses, all mixed effects linear regressions followed the same structure

$$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t} + \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} + \varepsilon$$

where PS is a vector of the Pearson's correlations of the neural pattern similarity between each morph and the corresponding original player per subject; choice is indexed by overall performance at each morph increment (whereby each subject has a composite score between 0 and 1 for each morph increment: 1 indicates choosing morph and 0 indicates not choosing the morph); and trust type is an indicator variable where 0=neutral, -1=untrustworthy, 1=trustworthy.

We confirmed our main perceptual findings (Table 3, Fig 3) with a bilateral anatomical ROI of the amygdala, providing further evidence that similarity to the untrustworthy player scales with the perceptual gradient in the amygdala (Table S8).

TABLE S8: Bilateral amygdala (Anatomical)

$$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t} + \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} + \varepsilon$$

DV	***Coefficient*** $(\boldsymbol{\beta})$	***Estimate (SE)***	***t-value***	***P value***
PS	Intercept	*-0.05 (.024)*	*-3.26*	*0.001***
	Untrustworthy × Perceptual Similarity	*0.01 (.004)*	*2.66*	*0.008***
	Trustworthy × Perceptual Similarity	*-0.01 (.01)*	*-1.33*	*0.18*
	Untrustworthy × Choice	*0.01 (.04)*	*0.35*	*0.72*
	Trustworthy × Choice	*0.07 (.06)*	*1.23*	*0.22*

Anatomical ROI created from WFU pick atlas. ***p<0.001, **p<0.01, *p<0.05

When we add in univariate BOLD activity into our regression—effectively testing whether overall increases in BOLD activity is a better predictor of pattern similarity—we find that perceptual similarity still significantly predicts PS along the untrustworthy gradient (Table S9).

TABLE S9: Amygdala

$$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t} + \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} + \beta_3 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} \times Univariate\ BOLD + \varepsilon$$

DV	*Coefficient* ($\boldsymbol{\beta}$)	*Estimate (SE)*	*t-value*	*P value*
PS	Intercept	*-0.18 (.04)*	*-4.27*	*<0.001****
	Untrustworthy × Perceptual Similarity	*.04 (.01)*	*3.18*	*0.001***
	Untrustworthy × Choice	*0.15 (.09)*	*1.60*	*0.10*
	Untrustworthy × Perceptual Similarity × BOLD	*-0.02 (.01)*	*-1.50*	*0.13*

ROI from conjunction of face presentation across all trust types during the initial learning episode (Conditioning Phase; Table S5) for Untrustworthy compared to Neutral gradient.
***p<0.001, **p<0.01, *p<0.05

Furthermore, we replicate the findings from Table 4 with an anatomical ROI of the vmPFC (Table S10) and report other regions of interest that predict both choice and perceptual similarity within the Untrustworthy condition (AI: Table S11; Ventral Striatum: Table S12). An anatomical ROI of the left Caudate (Table S13)—no left caudate was observed during the Conditioning Phase (Table S5)—further confirmed that that our caudate findings are localized to the right hemisphere (Table 5).

TABLE S10: vmPFC (Anatomical)

$$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t} + \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} + \varepsilon$$

DV	*Coefficient* ($\boldsymbol{\beta}$)	*Estimate (SE)*	*t-value*	*P value*
PS	Intercept	*0.03 (.01)*	*2.18*	*0.29**
	Untrustworthy × Perceptual Similarity	*0.002 (.004)*	*0.35*	*0.73*
	Trustworthy × Perceptual Similarity	*-0.006 (.006)*	*-1.05*	*0.29*
	Untrustworthy × Choice	*-0.08 (.03)*	*-2.43*	*0.01***
	Trustworthy × Choice	*0.006 (.04)*	*0.19*	*0.85*

Anatomical ROI created from WFU pick atlas.
***p<0.001, **p<0.01, *p<0.05

TABLE S11: Anterior Insula

$$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t} + \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t}\ \varepsilon$$

DV	*Coefficient* ($\boldsymbol{\beta}$)	*Estimate (SE)*	*t-value*	*P value*
PS	Intercept	*0.14 (.02)*	*5.81*	*<0.001****
	Untrustworthy × Perceptual Similarity	*-0.02 (.005)*	*-3.88*	*0.001****
	Trustworthy × Perceptual Similarity	*-0.004 (.006)*	*-0.65*	*0.52*
	Untrustworthy × Choice	*-0.10 (.03)*	*-2.60*	*0.009***
	Trustworthy × Choice	*-0.04 (.04)*	*-1.01*	*0.31*

ROI from conjunction of face presentation across all trust types during the initial learning episode (Conditioning Phase; Table S5). ***p<0.001, **p<0.01, *p<0.05

TABLE S12: Ventral Striatum

$$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t} + \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} + \varepsilon$$

DV	*Coefficient* ($\boldsymbol{\beta}$)	*Estimate (SE)*	*t-value*	*P value*
PS	Intercept	*-0.08 (.03)*	*-2.55*	*0.01**
	Untrustworthy × Perceptual Similarity	*0.05 (.01)*	*3.75*	*<0.001***
	Trustworthy × Perceptual Similarity	*-0.01 (.02)*	*-0.43*	*0.66*
	Untrustworthy × Choice	*0.22 (.14)*	*1.47*	*0.14*
	Trustworthy × Choice	*-0.04 (.09)*	*-0.49*	*0.62*

ROI from conjunction of face presentation across all trust types during the initial learning episode (Conditioning Phase; Table S5).
***p<0.001, **p<0.01, *p<0.05

TABLE S13: Anatomical Caudate

$PS_i = \beta_0 + \beta_1 Choice_{i,m} \times Trust\ Type_{i,t}\ \beta_2 Perceptual\ Similarity_{i,t} \times Trust\ Type_{i,t} + \varepsilon$

DV	*Coefficient* ($\boldsymbol{\beta}$)	*Estimate (SE)*	*t-value*	*P value*
PS	Intercept	*-0.06 (.01)*	*-1.35*	*0.18*
	Untrustworthy × Perceptual Similarity	*0.005 (.0004)*	*1.10*	*0.27*
	Trustworthy × Perceptual Similarity	*-0.001 (.006)*	*-0.26*	*0.79*
	Untrustworthy × Choice	*-0.03 (.03)*	*-0.82*	*0.41*
	Trustworthy × Choice	*0.02 (.03)*	*0.62*	*0.53*

Anatomical Caudate created from WFU pick atlas.
***p<0.001, **p<0.01, *p<0.05

We observed choice neural tuning profiles in the vmPFC for adaptively selecting morphs along the untrustworthy gradient and in the caudate for adaptively selecting morphs along the trustworthy gradient (Tables 4-5). Since the visual stimuli of the morphs at the 45-56% increments are identical between the conditions, a highly conservative test would be to probe each subject's data at the trial level to test whether choices still predict pattern similarity when holding perceptual similarity constant. To do this, we extracted neural patterns at each trial during the Generalization Phase for the most ambiguous morphs (45%-56%) and compared these patterns to the average pattern of the trustworthy and untrustworthy players elicited during the Conditioning Phase. We then used the choice data (play=1, no play=0) to predict pattern similarity in trials that presented the option to play with the 45%-56% morph increments, done for each condition separately. Results revealed that decisions to trust predicted increased pattern similarity in the caudate within the trustworthy condition (Table S14). The vmPFC exhibited a similar pattern but failed to reach significance within the Untrustworthy condition (Table S15).

TABLE S14: Caudate ROI

$PS_i = \beta_0 + \beta_1 Choice_{i,m} + \varepsilon$

DV	*Coefficient* ($\boldsymbol{\beta}$)	*Estimate (SE)*	*t-value*	*P value*
PS	Intercept	*0.009 (.01)*	*0.59*	*0.55*
	Choice	*0.03 (.01)*	*1.91*	*0.057^*

Trustworthy Condition for 45-56% Morph Increments (Choice: Play=1, No play=0). ***p<0.001, **p<0.01, *p<0.05, ^trending

TABLE S15: vmPFC ROI

$PS_i = \beta_0 + \beta_1 Choice_{i,m} + \varepsilon$

DV	***Coefficient*** $(\boldsymbol{\beta})$	***Estimate (SE)***	***t-value***	***P value***
PS	Intercept	*0.009 (.01)*	*0.59*	*0.55*
	Choice	*0.05 (.03)*	*1.40*	*0.17*

Untrustworthy Condition for 45-56% Morph Increments (Choice: Play=1, No play=0). ***p<0.001, **p<0.01, *p<0.05

SUPPLEMENTAL METHODS

Subjects. 91 subjects were recruited across all experiments. In Experiment 1, 31 subjects were recruited from the New York University subject pool and surrounding community. Two subjects were not included in the analysis for expressing doubts about the believability of the task. The final sample included 29 subjects (mean age=23.3, SD±4.6, 13 females). However, including all 31 subjects in the analysis, regardless of believability scores (see below), fully replicates all findings. In Experiment 2, 30 subjects were recruited from the New York University subject pool. Two subjects were excluded from analyses because of scanner issues (in one case, the scanner unexpectedly shut down and in the other, for excessive head movement). The final sample for Experiment 2 included 28 subjects (mean age=23.6, SD±4.4, 18 females). In addition, 20 subjects were recruited for piloting purposes, however for the sake of brevity, the results from the pilot study are not included as they fully replicate the findings from both the behavioral and imaging experiments. 10 subjects were also recruited for a perceptual categorization task (see below for details). Subjects were paid $15/hour in Experiment 1 (and in the pilot) and $25/hour in Experiment 2, and received additional compensation based on the result of one randomly selected trial from the first Trust Game.

Behavioral Experiment

Experiment 1 Task Procedures. Task procedures were the same for both experiments, with some minor exceptions that are delineated below. Before starting the experiment, subjects were asked to read instructions about each game. They were also given additional verbal and visual instructions to ensure full comprehension. After completing the instruction phase of the experiment, subjects were photographed in front of a white wall and told that their picture,

along with their responses to "how much of $10 would you split with a future player?" would be used for the next experiment with other subjects. Subjects were then endowed with $10 for investing in the Trust Game.

Conditioning Phase: The Trust Game. In the first task of the experiment, we asked subjects to complete a Trust Game (TG) with three other players. The TG involves a social interaction between two players, an Investor and a Trustee (Figure 1A). The first player, the Investor, is initially faced with a decision to keep a sum of money (in this case, $10) or share part or all of it with the Trustee. If shared, the investment is quadrupled (in the present example, to $40), and the Trustee now faces the decision to repay the trust by sending back half of the increased sum (e.g., $20 for each player, known as reciprocation), or to defect and violate trust by keeping the money (e.g. $40 for the Trustee), leaving the Investor with nothing. By trusting, the Investor can make double their money, however the Investor also risks losing their money if the Trustee decides to defect and keep the increased sum.

Upon arriving at the laboratory, subjects were told that they were randomly selected to be the Investor and that they would play with three other Trustees throughout the course of the task. Although subjects were led to believe they were playing with three other Trustees who had previously come into the lab, in reality, the Trustees were stimuli of three male faces pre-rated to be in neutrally trustworthy and attractive (within one standard deviation from mean trustworthiness and attractiveness ratings). Each Trustee was randomly yoked to a predetermined computer algorithm: the trustworthy Trustee reciprocated an initial decision to trust 93% of the time (80% of the time in the pilot experiment); the neutral Trustee reciprocated 60% of the time[1], and the untrustworthy Trustee reciprocated only 7% of the time (20% of the time in the pilot). Each face stimulus and associated name ("Zach", "Tom" and "Sam") were randomly assigned for every subject, such that while subject 1 might experience a specific face stimulus as trustworthy and know that their name is Sam, subject 2 would experience the same face stimulus as untrustworthy, knowing their name is Tom. In other words, all stimuli and names associated with reciprocation rates (Fig 1A) were fully randomized across all subjects in

[1]During piloting, the algorithm was set so that the neutral Trustee would reciprocate 50% of the time. However, we found that

order to control for possible demand effects.

On each trial, subjects decided how much to trust the Trustee (between $0 and the full $10, in $1 increments for Experiment 1 and in $2.50 increments for Experiment 2). If a decision to trust was made, subjects received feedback of whether the Trustee reciprocated or defected. Subjects played 15 trials with each Trustee, for a total of 45 trials, where each Trustee was pseudo-randomly presented across the experiment, such that each Trustee was never presented more than two times in a row. Subjects were told that one trial would be randomly realized to be paid out. If the subject chose not to trust on that trial, they would be able to keep their initial endowment of $10. If they chose to trust, then whatever the Trustee decided on that trial would be realized. For example, if a subject chose to trust $6 and the Trustee reciprocated, the participant would be paid out $12, whereas if the Trustee defected, the subjects would lose the $6.

To enhance believability of the paradigm—since the task required partners to be yoked to discrete reciprocate and defect algorithms—experimenters took the subjects' picture and asked them if they would complete another version of the task as the Trustee (the second mover) after the experimental session. This was explained as the most efficient way to feed forward subjects' responses so that multiple people do not need to come in for an experimental session at once. Subjects were told that in the event that their decisions as the Trustee were used in subsequent experimental sessions, they would be mailed a check based on that specific Investors decisions to entrust money and their recorded response to reciprocate or defect. In reality, these decisions were never fed forward.

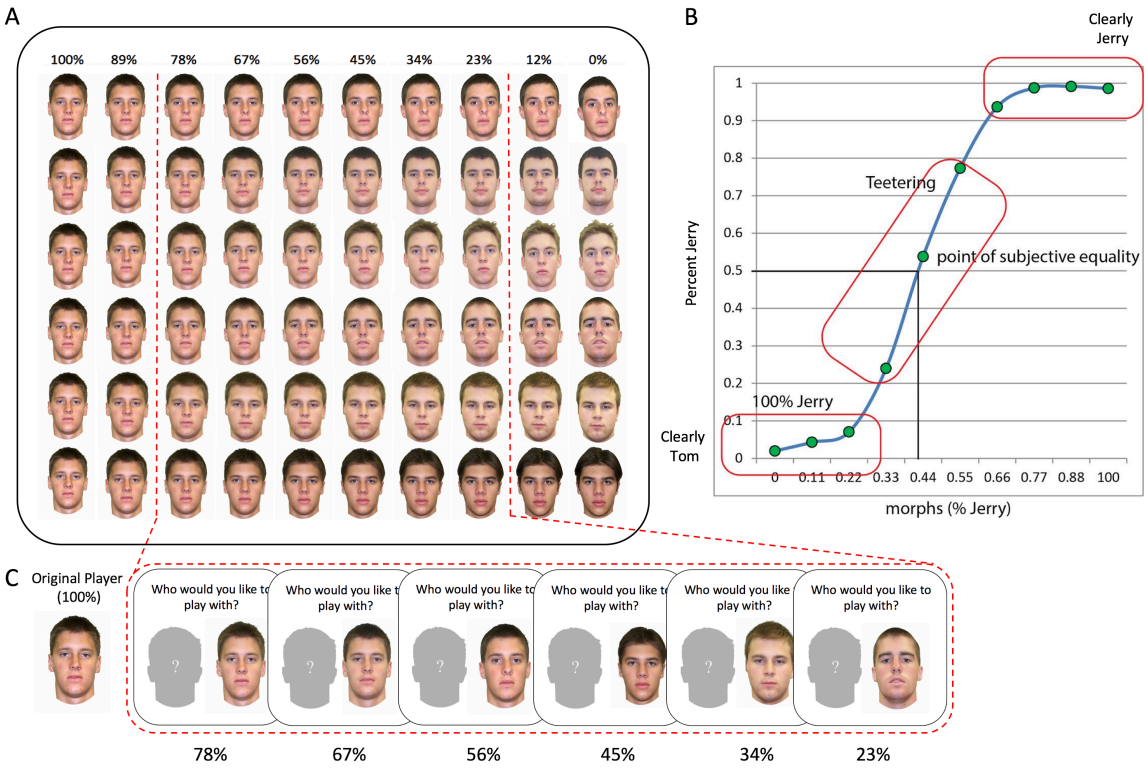

Fig S6 | Morphed Face Stimuli. A) *For the Generalization Phase, face stimuli were generated by morphing the face of each Trustee in the Conditioning Phase with 6 new faces. Morphs were set at 11% increments between the original Trustee and a new neutrally rated face, creating a linear continuum of faces.* ***B)*** *The categorization task—an alternative forced choice task—allowed us to identify at what point perceptual categorization between the two morphed individuals occurs (i.e. Tom and Jerry).* ***C)*** *The final stimuli included morphs that were at least two increments away from one another along the same continuum (i.e. each morph was at least 34% different from a morph on the same continuum).*

Generalization Phase: Choose Your Partner. In the second half of the experiment, we asked subjects to pick which partner they would play with in a subsequent TG. On each trial, of which there were 108, subjects were randomly presented with a picture of a man's face and an image of a silhouette grey face (indicating that the experimenters would select a new person at random; Fig 1C, faces and silhouettes were randomly presented on the left and right). Subjects were asked to decide which person they would prefer to play with. Each morph was presented twice per morph increment (original players were each morphed with six unique faces from which morphs were subsequently selected, Fig S6A). An additional 4 trials consisting of novel faces were added in Experiment 2 (two different faces, each presented twice). Although the 78% morph tautologically bore the greatest resemblance to the original player, subjects were tasked with picking partners across a non-trivial number of trials. Given this, and through our extensive

piloting and debriefing measures, we are confident that subjects did not recognize even the 78% morph as derivative of the original player. Debriefing measures further confirmed that subjects believed they were selecting partners among past real subjects (see below).

Debriefing Procedures. After the experimental session was finished, subjects were funnel debriefed in a manner consistent with effectively probing true believability of the task (1). Subjects answered on a 6-point Likert scale whether they had any doubt as to the veracity of the paradigm (anchored 1 = completely believed, 6 = did not believe). This allowed us to exclude subjects who indicated disbelief that they were playing with real players. For example, in Experiment 1, only two subjects responded with a 5 (mean believability rating 1.89 SD=1.12) and in Experiment 2, two subjects responded with a 5 and one subject with a 6 (mean believability rating 2.38 SD=1.28). During the funnel debriefing, even when explicitly probed about whether the other players might not be real, typical responses included: "*I was a real participant, and you took my picture, so I assumed that the other people were real participants"*; "*You took my picture, and the guys you showed me looked like they did the same thing";* and *"Of course I assumed that they were real. Why would we spend time picking random faces?".*

Stimuli Set. The stimuli used in the Conditioning Phase (and adapted for the Generalization Phase, see below), were taken from pictures of white male faces approximately between the ages of 18-24 (http://iilab.utep.edu/stimuli.htm). Each stimulus featured a unique, yet emotionally neutral face. In order to determine if the stimuli were matched in attractiveness, dominance, and trustworthiness, we asked an independent group (N=30) to rate each stimulus on Amazon Mechanical Turk. This task consisted of 179 trials in which subjects used a sliding bar to make ratings (separate scales of 1 to 10, where 1=not at all and 10=very) along these three dimensions. A subset of stimuli was selected according to the average levels of all three factors (within one standard deviation from mean ratings).

Morphed face stimuli. In the Generalization Phase, face stimuli were generated by morphing faces together using a morphing technology, whereby each Trustee (the original players in the Conditioning Phase) were morphed with six new faces pre-rated to be neutrally trustworthy and attractive. We set the morphs at 11% increments between the original Trustee and a new neutrally rated face, which created a linear continuum of 8 morphed faces (Fig 1C). This resulted

in 48 additional stimuli for each original Trustee, for a total of 144 new stimuli. From these additional stimuli, we removed morphs that were too perceptually similar to the original Trustee or to the entirely novel player (89% and 12% increments, respectively). This resulted in six morph types at the 23%, 34%, 45%, 56%, 67%, and 78% increments from the original Trustee stimuli (Fig S6A). Furthermore, to ensure that subjects did not find the stimuli too similar to one another (after all, they are along a continuum), we only used morphs that were two increments away from one another along the same continuum (i.e. each morph was at least 34% different from a morph on the same continuum). This narrowed our stimuli set to three different morphed faces being presented per morph increment (see Fig 6C for an example of a final morph continuum for one Trustee). The final stimuli (of which there were 18 per trust type), all originating from the original Trustees, were each presented twice to enable greater power for investigating choices along the generalization gradients.

Categorization task to determine final stimuli. Since a linear continuum of morphed faces was generated between two individual face exemplars, it was important to identify at what point perceptual categorization between these two morphed individuals occurs. Accordingly, we generated an alternative forced choice categorization task (using the methods from Beale & Keil 1995 (2)). 10 subjects were first asked to associate both individual exemplars with names (i.e. 'Tom' and 'Jerry'). After learning which individual is Tom and which individual is Jerry, subjects were randomly shown the morphed faces and asked to categorize whether the individual was Tom or Jerry. This enabled us to determine the approximate point of subjective equality for each morph continuum (morph nearest 50%, Fig S6B).

Imaging Experiment

Experiment 2 Procedures. The task procedures were largely the same for the imaging experiment. Subjects completed the tasks in the following order: Conditioning Phase, Generalization Phase (comprised of four short runs rather than one long run) and a face localizer task. In the Conditioning Phase, subjects had up to 3.5s to decide how much, if any, of their money (in increments of $2.50) they would like to transfer to the Trustee. If subjects decided to keep their money, the trial ended (Fig S7A). However, if money was transferred, the transferred amount was presented on the screen for 2s. Subjects then viewed a screen of the Trustee thinking, which was jittered between 1.5-3s, before seeing feedback (presented for 3s) about

whether the Trustee defected or reciprocated. Each inter-trial-interval was jittered around 3s.

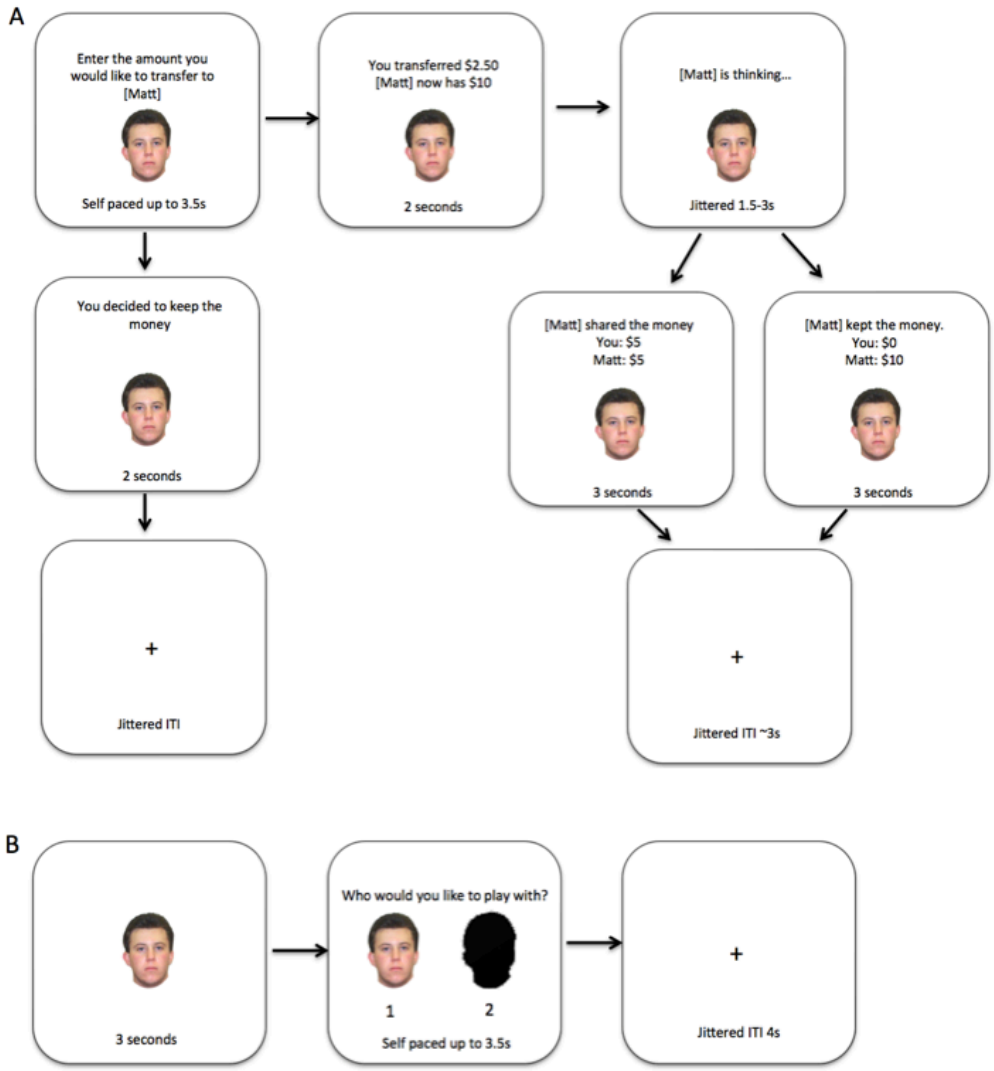

Fig S7 | Timing Structure for Conditioning and Generalization Phases in the scanner. A) *The Conditioning Phase was a classic iterated Trust Game with three players.* ***B)*** *The Generalization Phase was a forced binary decision between a morph and another unknown random player (denoted by a silhouette).*

The Generalization Phase was broken into four, 4-minute runs, each consisting of 28 trials, in order to optimize sequencing for representation similarity analysis (3). In addition to the 108 morphs presented, we also included four additional presentations of completely new, novel faces that were never morphed with any of the original Trustees. At the start of each trial, a face (a morph) was presented on the middle of the screen for 3s, after which the morph moved either left or right and an image of a silhouette also appeared on the opposite side of the screen. These two images (the silhouette and morph) were randomly presented on either the

left or right side (Fig S7B). Subjects then had up to 3.5s to decide which individual they would prefer to play with, and could indicate their response by using the left or right buttons on the button box. To deal with possible expectancies with fixed inter-trial-intervals (ITI), we used an ITI with a Gaussian distribution jittered around 4s.

Face localizer. The face localizer task included four blocks of faces, objects, scrambled objects, and scenes. Each block consisted of 12 pictures presented for 800ms separated by 200ms black screen and followed by a 12s-fixation cross. Data from these scans are not included in the present study.

fMRI acquisition and analysis. Functional imaging was performed using a Siemens Allegra 3T head-only scanner located at the Center for Brain Imaging at New York University. Functional data were collected using an echo-planar (EPI) pulse sequence (36 interleaved slices; TR = 2000ms; TE = 30ms; flip angle = 78º; FOV 192 mm) with slices oriented parallel to the AC-PC axis. Slices were positioned ventrally to provide full coverage of the anterior temporal lobes and prefrontal cortex; this resulted in omission of parts of the superior parietal cortex. A high-resolution T1-weighted anatomical scan (magnetization-prepared rapid-acquisition gradient echo sequence, 1x1x1mm) was also obtained for each subject after the final block of the localizer task.

All pre-processing and data analysis was conducted using SPM8 (Wellcome Trust Centre, www.fil.ion.ucl.ac.uk) implemented in MATLAB. For univariate analyses, images were spatially normalized into Montreal Neurological Institute (MNI) space, voxel size resampled to 3x3x3mm, and smoothed using an isotropic 8mm Gaussian full-width half-maximum kernel. For multivariate analysis (PSA), images were spatially normalized into MNI space, resampled to 3x3x3mm, and smoothed using only an isotropic 2mm Gaussian full-width half-maximum kernel. Functional images were co-registered to each participant's high-resolution T1-weighted structural scan. To account for magnetic equilibrium, the first five functional images were discarded. Images were corrected for head motion using a 3mm movement cutoff in any dimension.

At the first level, activated voxels were identified using an event-related statistical model

representing each presentation of a morph in the Generalization Phase or a decision to invest money (or refrain from investing money) in the Conditioning Phase. These models were convolved with a canonical hemodynamic response function and mean-corrected. Six head-motion parameters defined by the realignment were added to the model as regressors of no interest. In the Generalization Phase, analysis was carried out to establish each subject's voxel-wise activation when subjects made their response or observed the morph (depending on the analysis). For each subject, contrast images were calculated for each morph increment. These first level contrasts were then aggregated into second level full factorial analyses of variance (ANOVAs) in order to compute group statistics. For univariate analyses, we report activity at $p<0.001$ uncorrected for multiple spatial comparisons across the whole-brain, and $p<0.05$ FWE corrected for the following *a priori* regions of interest (ROIs) for all PSA analyses.

Regions of Interest. For the multivariate analyses we employed a classic ROI approach. Functional ROIs were created by running a conjunction analysis on the face presentation epoch for all conditions (Trustworthy, Untrustworthy and Neutral) during the Conditioning Phase. Regions active during initial learning across the three conditions were Bonferroni FWE corrected at $p=.05$. These regions (Table S5) were then further selected based on *a priori* hypotheses (see manuscript) and served as the main ROIs to test whether, at the multivariate level, neural patterns during generalization elicit increasingly similar patterns of neural content compared to those elicited during initial learning. We also used independent anatomical ROIs to verify results within the amygdala. Anatomical ROIs were made using the Wake Forest Pick Atlas.

Pattern Similarity Analysis. Separately for each participant and ROI, we computed neural similarity scores in subject-specific space for each morph increment along the Trustworthy, Untrustworthy and Neutral gradients. Estimates of each face (morphs and original players) were computed in one GLM for the Conditioning Phase and one GLM for the Generalization Phase. Each face was indexed by one regressor (although there were six morph presentations at each increment, each presented two times and original players in the Conditioning Phase were presented 15 times), modeled as an impulse at the start of each of the presentations of the morph (onset) and convolved with a canonical HRF. In each ROI, The neural representation of each morph increment was operationalized as a vector of t-values corresponding to that morph increment (e.g. a vector of all voxel-wise responses to that morph), which was subsequently

compared to a vector of t-values corresponding to the original player in the Conditioning Phase (e.g. a vector of all voxel-wise responses to the original player). Because we were only interested in how perceptual similarity and choice—each as a function of trust type—are instantiated neurally, our ROI specific PSA regressions explicitly tested for the effects of perceptual similarity as a function of trust type, and choice as a function of trust type. In line with both classic statistical principles (4) and more recent theoretical inferential work (5, 6), this theory driven model only captures the variables of interest (i.e. the interactive effects) without compromising the strength of the model or requiring ad hoc interpretations of nuisance variables. Neural similarity scores for each morph were calculated as the mean of the Pearson correlations between the ROI specific t-maps of that morph and the ROI specific t-map corresponding to learning the trustworthiness of the original player during the Conditioning Phase.

Similarity scores were Fisher transformed separately for each subject relative to that subject's within condition similarity score mean and standard deviation. Mixed-effects linear regressions were performed on these similarity scores using maximal hierarchal models (7) in MATLAB 2016a.

An additional similarity analysis was computed to examine trial-by-trial relationships between representational similarity and trust decisions for morphs with the greatest perceptual ambiguity (45-56%). For this analysis, individual estimates of each trial during the Generalization Phase were extracted using a GLM. This is different from the main PSA analysis, where the pattern of activation was extracted by average BOLD responses to all trials that included the same morph. Here, each trial was indexed by one regressor at the start of the trial and convolved with a canonical HRF function. For each ROI, patterns of activation were extracted from resulting t-maps of each trial. These trial-level estimates were input into a mixed-effects regression predicting neural pattern similarity as a function of the decision to either play or not play with the morph presented on each trial.

REFERENCES

1. FeldmanHall O*, et al.* (2012) What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition* 123(3):434-441.
2. Beale JM & Keil FC (1995) Categorical Effects in the Perception of Faces. *Cognition* 57(3):217-239.

3. Mur M, Bandettini PA, & Kriegeskorte N (2009) Revealing representational content with pattern-information fMRIan introductory guide. *Social cognitive and affective neuroscience* 4(1):101-109.
4. Sokal RR & Rohlf FJ (1969) *Biometry; the principles and practice of statistics in biological research* (W. H. Freeman, San Francisco,) pp xxi, 776 p.
5. Crawley MJ (2013) *The R book* (Wiley, Chichester, West Sussex, United Kingdom) Second edition. Ed pp xxiv, 1051 pages.
6. Cleves MA (2008) *An introduction to survival analysis using Stata* (Stata Press, College Station, Tex.) 2nd Ed pp xxiv, 372 p.
7. Barr DJ, Levy R, Scheepers C, & Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang* 68(3):255-278.