Stimulus generalization as a mechanism for learning to trust

Oriel FeldmanHall[a,1], Joseph E. Dunsmoor[b,2], Alexa Tompary[c,2], Lindsay E. Hunter[d], Alexander Todorov[d], and Elizabeth A. Phelps[c,e,f]

[a]Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, RI 02906; [b]Department of Psychiatry, The University of Texas at Austin, Austin, TX 78712; [c]Department of Psychology, New York University, New York, NY 10003; [d]Department of Psychology, Princeton University, Princeton, NJ; [e]Center for Neural Science, New York University, New York, NY 10003; and [f]Emotional Brain Institute, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962

Edited by Dean Mobbs, California Institute of Technology, Pasadena, CA and accepted by Editorial Board Member Marlene Behrmann January 2, 2018 (received for review August 28, 2017)

How do humans learn to trust unfamiliar others? Decisions in the absence of direct knowledge rely on our ability to generalize from past experiences and are often shaped by the degree of similarity between prior experience and novel situations. Here, we leverage a stimulus generalization framework to examine how perceptual similarity between known individuals and unfamiliar strangers shapes social learning. In a behavioral study, subjects play an iterative trust game with three partners who exhibit highly trustworthy, somewhat trustworthy, or highly untrustworthy behavior. After learning who can be trusted, subjects select new partners for a second game. Unbeknownst to subjects, each potential new partner was parametrically morphed with one of the three original players. Results reveal that subjects prefer to play with strangers who implicitly resemble the original player they previously learned was trustworthy and avoid playing with strangers resembling the untrustworthy player. These decisions to trust or distrust strangers formed a generalization gradient that converged toward baseline as perceptual similarity to the original player diminished. In a second imaging experiment we replicate these behavioral gradients and leverage multivariate pattern similarity analyses to reveal that a tuning profile of activation patterns in the amygdala selectively captures increasing perceptions of untrustworthiness. We additionally observe that within the caudate adaptive choices to trust rely on neural activation patterns similar to those elicited when learning about unrelated, but perceptually familiar, individuals. Together, these findings suggest an associative learning mechanism efficiently deploys moral information encoded from past experiences to guide future choice.

trust | social learning | stimulus generalization | amygdala | caudate

The ubiquity of social interaction and collaboration throughout human evolution underscores the importance of adaptive social decision making (1). For instance, deciding whether to trust is a daily activity occurring with varying levels of consequence, from telling trivial secrets to loaning significant amounts of money. Choosing to place one's own well-being into the hands of another typically necessitates first-hand experiences demonstrating the integrity of a partner's reputation (2–4). However, people are frequently confronted with situations in which they must decide whether to trust a stranger in the absence of any prior experience. In contexts void of reputational information or direct prior knowledge, what governs decisions to trust?

Associative learning theory provides useful and straightforward descriptions of how experience shapes learning to guide value-based behaviors (5, 6). One principle of associative learning is that value can spread or transfer between stimuli that perceptually or conceptually resemble one another, known as stimulus generalization (7–9). As stimuli rarely occur in the exact same form from one encounter to the next, similarity-based stimulus generalization mechanisms can be highly adaptive. Given that stimulus generalization is a learning process well-documented across species within the nonsocial domain (10, 11), it is conceivable that a related mechanism operates in highly complex social environments (12), such as deciding when to trust or cooperate with unfamiliar others. To investigate this possibility we designed a task to test whether the transfer of learned value (i.e., stimulus generalization) occurs within the social domain, specifically among moral behaviors exhibited by individuals. In other words, do strangers gain positive and negative social value simply because they resemble another person whose social value is known, and do differences in the strength of this resemblance determine decisions to trust?

In our first experiment, we employed a classic trust game, which can be considered a form of associative conditioning, as subjects learn the social value (in this case, trustworthiness) associated with specific individuals over a series of trials. Subjects were endowed with $10 and acted as the investor in an iterative trust game, deciding whether to entrust their money with three different players (conditioning phase, Fig. 1*A*). On each trial, subjects knew that any money invested would be multiplied four times, and the other player could then either share the money back with the subject (reciprocate) or keep the money for himself (defect). Each player was either highly trustworthy (reciprocation reinforced 93%), somewhat trustworthy (neutral: reciprocation reinforced 60%), or not at all trustworthy (reciprocation reinforced 7%). Subjects successfully learned which player could

Significance

Humans can learn to trust through direct social experiences. In our everyday lives, however, we constantly meet new people where judgments of trustworthiness are blind to reputation. In these cases, what drives decisions to trust? We find a simple learning mechanism observed across species—stimulus generalization—is deployed in complex social learning environments: Individuals distrust strangers who implicitly resemble those known to be untrustworthy. These behavioral findings were mirrored at the neural level, revealing that the amygdala and caudate selectively encode the transfer of social value during moral learning. The results demonstrate a mechanism that draws on prior learning to reduce the uncertainty associated with strangers, ultimately facilitating potentially adaptive decisions to trust, or withhold trust from, unfamiliar others.

Author contributions: O.F.H., J.E.D., and E.A.P. designed research; O.F.H. and L.E.H. performed research; O.F.H., J.E.D., and A. Tompary analyzed data; and O.F.H., J.E.D., A. Tompary, A. Todorov, and E.A.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. D.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. Email: oriel.feldmanhall@brown.edu.

[2]J.E.D. and A. Tompary contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1715227115/-/DCSupplemental.

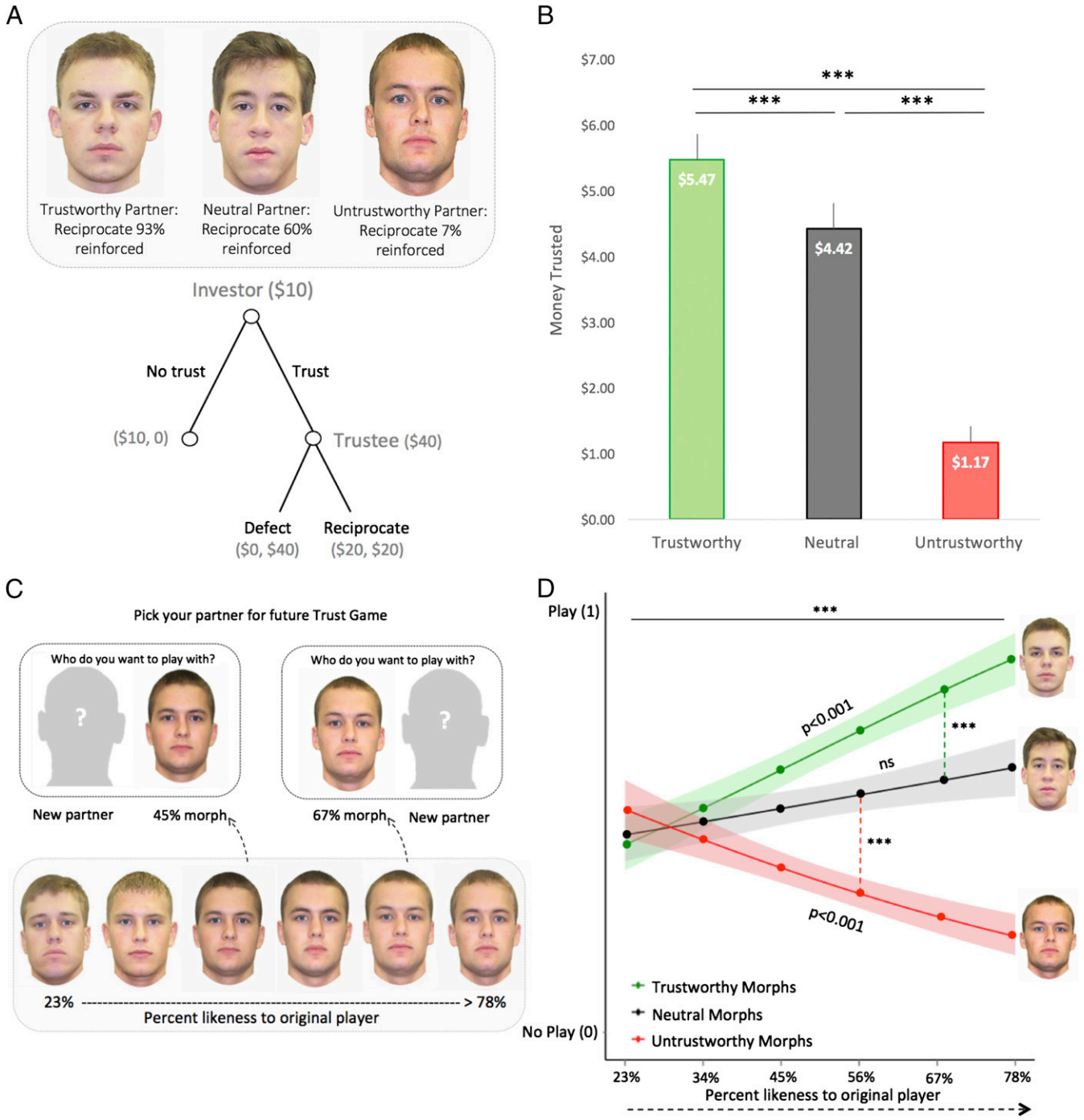

Fig. 1. Conditioning and generalization phases. (*A*) In the conditioning phase subjects played an iterative trust game where they learned about three different partners who were either trustworthy, untrustworthy, or somewhat trustworthy (i.e., neutral). (*B*) Subjects learned to invest the most money in the trustworthy partner and the least amount of money in the untrustworthy player. Error bars reflect 1 SEM. (*C*) In a second task—the generalization phase—subjects selected partners to play with in a subsequent trust game. Unbeknownst to the subject, potential partners were morphed with one of the three original players (in increments of 11%). This allowed us to test whether adaptive choice relies on similarity-based stimulus generalization. The example morph gradient shows six individuals who were morphed with the original untrustworthy partner, resulting in a smooth perceptual continuum (see *SI Appendix*, Fig. S6 for more details). (*D*) Perceived similarity to the original player and trust type predicts choosing to play with the morph, where the *x* axis denotes similarity and the *y* axis denotes the likelihood of a decision to play (1) or refrain from playing (0) with the partner. A hierarchal regression reveals subjects were incrementally biased in selecting morphs that resembled past individuals associated with trustworthy outcomes and averse to selecting morphs that resembled past individuals associated with untrustworthy outcomes (fitted data are plotted above, but raw data can be found in *SI Appendix*). These behavioral tuning profiles were structurally asymmetric, such that subjects were more likely to avoid morphs that even slightly resembled past untrustworthy individuals (tested by comparing slopes of lines and through pairwise comparisons to neutral gradient). ***$P < 0.001$; ns, not significant.

be trusted, as the money entrusted to the trustworthy player ($5.47, SD ± 2.2) was significantly greater than the amount of money sent to either the neutral ($4.42, SD ± 2.2) or untrustworthy ($1.17, SD ± 1.4) players [repeated measures ANOVA: $F(2,56) = 77.8$, $P < 0.001$, $\eta^2 = 0.72$, all pairwise comparisons $P < 0.001$; Fig. 1*B*].

Table 1. Exp. 1: $Choice_{i,t} = \beta_0 + \beta_1 Trust\ Type_{i,t} \times \beta_2 Perceptual\ Similarity_{i,t} + \varepsilon$

Coefficient (β)	Estimate (SE)	*t* value	*P* value
Intercept	0.42 (0.04)	9.15	0.001**
Untrustworthy	−0.13 (0.04)	−3.36	<0.001***
Trustworthy	0.09 (0.03)	2.43	0.02*
Perceptual Similarity	0.22 (0.16)	1.37	0.17
Untrustworthy × Perceptual Similarity	−0.65 (0.16)	−3.85	<0.001***
Trustworthy × Perceptual Similarity	0.40 (0.17)	2.36	0.01*

Hierarchical logistic regression where choice is indexed by subject and trial (0 = morph not selected, 1 = morph selected), trust type (indicator variable: 0 = neutral, such that neutral serves as the reference category, −1 = untrustworthy, 1 = trustworthy), and perceptual similarity (which was mean-centered) is indexed linearly by morph increment. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$.

To test for generalization of learning, in a second task subjects could now select their partners for a future trust game. Each trial presented a picture of a person's face alongside an image of a silhouette (generalization phase, Fig. 1*C*)—indicating a random new partner (13). Unbeknownst to subjects, these faces were generated by morphing each of the original players from the conditioning phase with new, never-before-seen faces (stimuli were matched on multiple dimensions, including perceived attractiveness and trustworthiness). This resulted in a smooth and linear continuum of morphed stimuli at 11% increments (Fig. 1*C*), which served as the final stimuli for the generalization phase (see *SI Appendix* for details on morph creation, normalization of stimuli, and final selection; *SI Appendix*, Fig. S6). Critically, subjects believed the morphs were real people and potential partners for the next trust game, ensuring that the decision to trust was not affected by conscious awareness that the morphs were derivatives of the players from the preceding trust game (e.g., in debriefing subjects did not report being aware that these were morphed faces and believed each face was a unique individual; see *SI Appendix* for details of extensive piloting and debriefing measures). This distinguishes our paradigm from more classic stimulus generalization designs in which simple sensory stimuli (e.g., tones or lights) vary along an explicit unimodal continuum.

This perceptual morph gradient allows us to test whether aversive experiences with one individual (e.g., the untrustworthy partner in the conditioning phase) leads to inferring that related individuals (e.g., the morphs in the generalization phase) are aversive and untrustworthy and thus should be avoided. Since subjects successfully learned which of the three original players could be trusted and which could not, we might expect subjects to be incrementally biased in selecting morphs that more closely resemble past individuals associated with positive outcomes (high rates of reciprocation) and to avoid morphs that more closely resemble past individuals associated with negative outcomes (high rates of defection)—even though morphs were thought to be distinct individuals unrelated to the previous players.

Accordingly, we investigated two key questions: (*i*) Is perceptual similarity of the morphs implicitly used to guide novel decisions to trust unfamiliar others, and (*ii*) do these putative generalization gradients evoke structurally similar behavioral tuning profiles (e.g., adaptively refraining from or choosing to trust at the same rate)? To answer these questions, we ran a hierarchal logistic regression (14), where both trustworthiness type (whether faces were 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 (trust type × perceptual similarity: $P < 0.001$; Fig. 1*D* and Table 1; raw data in *SI Appendix*, Fig. S3). The opposite generalization pattern was observed for untrustworthy morphs; the greater the perceptual similarity to the original untrustworthy player, the less likely subjects were to select the morph for a second trust game. Evidence of a similarity heuristic (15, 16) biasing subsequent choice demonstrates that social learning in the appetitive and aversive domains relies—at least in part—on comparing current experiences with past experiences. Critically, there was no effect of the neutral morph continuum on choice: Subjects were just as likely to choose the morph that most closely resembled the original neutral player as they were the morph that least likely resembled the neutral player ($P > 0.1$).

The shape of the behavioral tuning profiles—choosing to trust or distrust the morphs as a function of increasing perceptual similarity to the original players in the conditioning phase—can be taken as an indication of the strength of generalization, since the rate at which responses increase quantitatively characterizes the strength of the generalization of learned associations. For instance, tuning profiles that are wide and flat depict a slowly decaying gradient reflective of broad overgeneralization of learned stimuli. Given that the conditioning literature illustrates asymmetrical learning in the aversive and appetitive domains (17)—which has been likened to a "better safe than sorry" strategy for aversive phenomena (18)—we further posited that there may be similar asymmetries in tuning profiles between trustworthy and untrustworthy morph gradients. In line with this

Table 2. Exp. 2: $Choice_{i,t} = \beta_0 + \beta_1 Trust\ Type_{i,t} \times \beta_2 Perceptual\ Similarity_{i,t} + \varepsilon$

Coefficient (β)	Estimate (SE)	*t* value	*P* value
Intercept	0.43 (0.08)	5.23	<0.001***
Untrustworthy	−0.11 (0.03)	−3.36	<0.001***
Trustworthy	0.17 (0.04)	4.07	<0.001***
Perceptual Similarity	0.007 (0.15)	0.05	0.96
Untrustworthy × Perceptual Similarity	−0.49 (0.15)	−3.18	<0.001***
Trustworthy × Perceptual Similarity	0.66 (0.14)	4.62	<0.001***

Hierarchical logistic regression where choice is indexed by subject and trial (0 = morph not selected, 1 = morph selected), trust type (indicator variable: 0 = neutral, −1 = untrustworthy, 1 = trustworthy), and perceptual similarity (which was mean-centered) is indexed linearly by morph increment. ***$P < 0.001$.

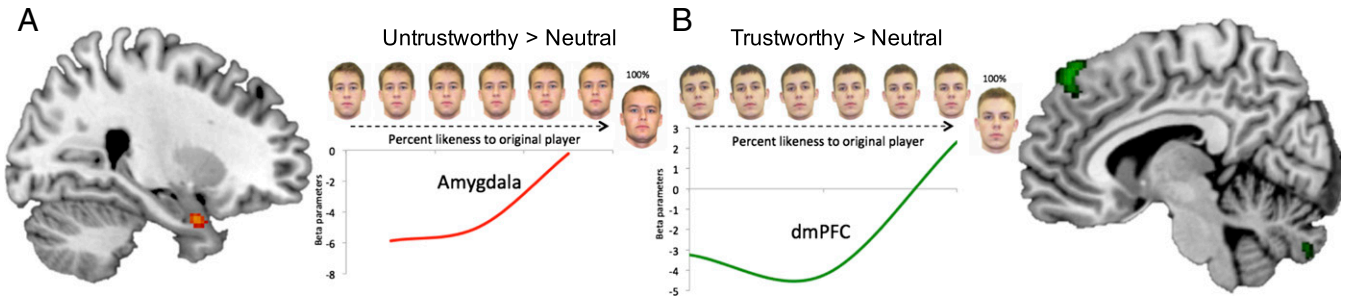

Fig. 2. BOLD activity differentially tracks untrustworthy and trustworthy perceptual similarity gradients. (*A*) Univariate parametric analysis indexing perceptual similarity of the morphs along the untrustworthy gradient > perceptual similarity of the morphs along the neutral gradient in the generalization phase reveal BOLD signal in the amygdala (x = 34, y = 4, z = −26, z-value = 3.82) scales with increasing untrustworthiness. (*B*) The same analysis tracking perceptual similarity of the morphs along the trustworthy gradient > perceptual similarity of the morphs along the neutral gradient illustrate BOLD activity in the dmPFC (x = 4, y = 42, z = 56, z-value = 3.24).

prediction, we observed that the generalization gradient for untrustworthy morphs was wider than that for the trustworthy morphs, despite the original players' being perfectly inversely matched in their reinforcement rates [the slope of the trustworthy and untrustworthy gradients' coefficients from the regression (Table 1) were significantly different from one another (t (28) = −6.13, $P < 0.001$)]. This generalization asymmetry was further confirmed with a repeated measures ANOVA trust type × morph interaction [$F(10,280) = 10.76$, $P < 0.001$, $\eta^2 = 0.26$] (pairwise comparisons for untrustworthy > neutral morph increments were significant until 56% similarity, $P < 0.001$; pairwise comparisons for trustworthy > neutral morph increments were only significant until 67% similarity, $P < 0.001$; Fig. 1*D* and see *SI Appendix*, Fig. S3*A* for raw data). That subjects preferentially avoided the untrustworthy morphs more so than engaging with the trustworthy morphs suggests an asymmetric overgeneralization toward individuals perceived to be morally aversive.

Identifying and characterizing the neural circuitry supporting decisions to trust can further elucidate how this generalization mechanism is precisely deployed and instantiated during social learning. One possibility is that adaptively choosing to trust individuals bearing a greater resemblance to the original trustworthy player should elicit blood-oxygen-level-dependent (BOLD) activation in a network typically associated with trust and social reward [e.g., caudate and dorsal medial prefrontal cortex (dmPFC) (2, 3, 19–23)], while the opposite pattern of behavior should reveal activity in regions critical for processing aversive and socially threatening stimuli [e.g., the amygdala and anterior insula (20, 24–26)].

Therefore, in a second experiment, we sought to replicate our behavioral findings while also probing the underlying BOLD activity supporting adaptive decisions to trust unfamiliar others. While in the scanner, subjects ($n = 28$) followed the same experimental structure described in the first experiment, completing both conditioning and generalization phases. As before, we found that during conditioning subjects learned which players could be trusted and which could not (*SI Appendix*). Furthermore, the asymmetric behavioral tuning profiles were again observed in the generalization phase, replicating the findings from Exp. 1: As perceptual resemblance to the original player increases, subjects are significantly more likely to choose the morph along the trustworthy gradient and not choose the morph along the untrustworthy gradient (Table 2), such that morphs associated with the original player who broke trust and facilitated aversive outcomes were again more broadly and systematically avoided (see *SI Appendix*, Fig. S3*B* for details).

To probe BOLD activity indexing the perceptual similarity of trustworthy and untrustworthy morphs gradients to their respective original players we focused on the presentation of the faces during the choice epoch of the generalization phase (fixed 3 s). We hypothesized that subjects' ability to adaptively refrain from trusting or choose to trust during the generalization phase might rely on discrete BOLD signals that scale with perceptual similarity along the untrustworthy and trustworthy generalization gradient. Using univariate analyses, we applied parametric regressors delineating each perceptual increment along the morph gradient and separately compared activation to trustworthy and

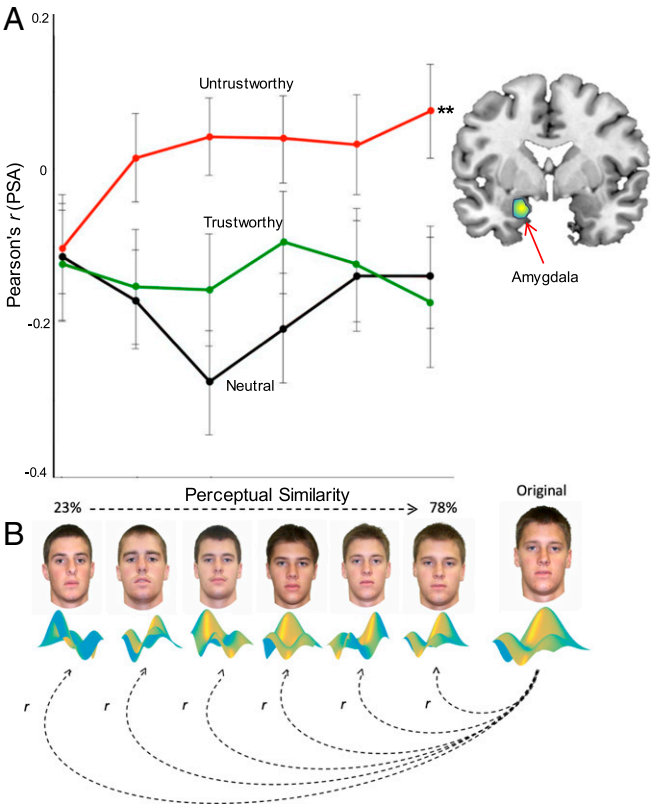

Fig. 3. Amygdala tracks the perceptual tuning profiles of untrustworthy individuals. (*A*) Increasingly similar patterns (derived from PSA) in the left amygdala reveal a neural tuning profile that tracks perceived increases in untrustworthiness. No relationship was found for either the trustworthy or neutral gradients. (*B*) PSA correlated (Pearson's r) activation patterns evoked at each morph increment along the generalization gradient to those evoked when initially learning about the original players (separately for each original trustworthy, untrustworthy, and neutral player).

untrustworthy morph gradients to morphs along the neutral gradient. This enabled a whole-brain examination of the BOLD signal tracking perceptual similarity of trustworthy and untrustworthy partners. Comparison with the neutral gradient serves as a conservative baseline to control for regions that may also generally scale with perceptual similarity. Results revealed that activity in the amygdala tracked morphs that gradually resembled the original untrustworthy player (parametric untrustworthy gradient > parametric neutral gradient; Fig. 2 and *SI Appendix*, Table S3), while the opposite contrast—increasing similarity toward the original trustworthy player compared with increasing similarity of the neutral player—revealed increasing activation in the dmPFC (*SI Appendix*, Table S4), two regions known to be associated with judgments of untrustworthiness and trustworthiness, respectively (20).

Univariate analyses, however, are less well suited for investigating the content of neural representations (27, 28). Accordingly, we used multivariate pattern similarity analysis (PSA) to further characterize how perceptual information and choice are represented in key regions of interest. Probing patterns of BOLD signal across voxels allows us to examine whether social value spreads via an associative learning mechanism at the neural level. In other words, we can test for similarity between the structure of the activation patterns supporting learning about a trustworthy or untrustworthy player in the conditioning phase and the activation patterns elicited at each subsequent morph increment in the generalization phase. If value learned during the initial episode is transferred to subsequent interactions, then morphs that increasingly resemble the original untrustworthy player may evoke activation patterns that are incrementally more similar to the activation patterns that supported learning about the original player's untrustworthiness. These perceptual neural tuning profiles—the correlation between activation patterns during learning about the untrustworthy player and the activation patterns later elicited at each morph increment—should scale with perceptual similarity along the generalization gradient.

Beyond perceptual similarity, there is also the possibility that a discrete generalization mechanism for choice will be reflected at the neural level, which would be consistent with animal lesion work demonstrating an important distinction between the ability to perceptually and behaviorally discriminate along a generalization gradient (29). If this were the case, the observed adaptive behavioral tuning profiles—for example, engaging with morphs that look increasingly more like the original trustworthy player and avoiding morphs that look increasingly more like the original untrustworthy player—should be mirrored at the neural level. This choice model posits that if stimulus generalization is related to the initial learning episode then neural patterns active during learning may also be similarly active when guiding subsequent choice, henceforth referred to as choice neural tuning profiles.

In our PSA used to address these two predictions the activation patterns evoked at each morph increment along the generalization gradient were correlated with the activation patterns evoked from the original player in the initial learning phase (Pearson's; Fig. 3*B*). For each region of interest (ROI), a mixed-effects hierarchical linear regression model was fit with choice, perceptual similarity, and trust type as predictors of the similarity patterns (PS) between each morph increment in the generalization phase and the activation pattern corresponding to that morph's original player in the conditioning phase. This model captures both the effects of perceptual similarity and adaptive choice in predicting the neural similarity of representational content between each morph and the corresponding original player (see *Materials and Methods* and *SI Appendix* for analysis details). ROIs were created from thresholded T maps on a conjunction analysis of all trust types presented during the conditioning phase (*SI Appendix*, Table S5) given our a priori hypothesis about which brain regions would be involved when learning to trust (discussed below).

Based on prior work (30–34) and the results from our univariate analysis, there are candidate brain regions, including the amygdala, that should be more sensitive to exhibiting an incremental change in pattern similarity along the perceptual untrustworthy gradient. For example, much work has characterized the role of the amygdala during emotional learning (35–37), such as fear generalization (10, 38) and social avoidance (39), which provides a convincing case for exploring whether aversive social phenomena, such as perceiving distrustful individuals, also evoke a similar perceptual neural tuning profile in the amygdala. Results reveal that at the perceptual level the amygdala exhibits a neural tuning profile—increasingly similar patterns—that tracks perceived increases in untrustworthiness ($P < 0.001$; Fig. 3*A* and Table 3, further confirmed with an anatomical ROI: *SI Appendix*, Table S8). This relationship between perceptual similarity with the original untrustworthy player and patterns in the amygdala remains robust when accounting for overall univariate BOLD activity (*SI Appendix*, Table S9). Simply put, the activation patterns within the amygdala recruited when initially learning about the untrustworthy player were increasingly correlated with the activation patterns recruited as the morphs progressively resembled the original untrustworthy player. This relationship did not extend to either the trustworthy or neutral gradients.

We further theorized that when adaptively choosing which morph to play with (Fig. 4*A*) a network involved in trust and distrust (26, 40–42), which includes the ventromedial prefrontal cortex (vmPFC) and anterior insula (AI), may index decisions to refrain from playing with morphs that resemble untrustworthy past players. Conversely, given the abundance of evidence documenting the role the caudate plays in encoding reward (43, 44)—especially when deciding whether to trust (2, 3)—we posited that activation patterns within the caudate might selectively scale along the trustworthy gradient.

Consistent with the notion that there are distinct perceptual and behavioral generalization gradients reflected at the neural level, we found that refraining from engaging with the untrustworthy morph (a putatively adaptive choice) evoked a neural tuning profile in the vmPFC (even when controlling for perceptual similarity, Fig. 4*C* and Table 4; see *SI Appendix* for AI data). These choice neural tuning profiles in the vmPFC were only observed for the untrustworthy gradient (results replicate when choice is the dependent variable predicted by the interaction between perceptual similarity and neural pattern similarity, $P = 0.01$). In contrast, when selecting morphs along the trustworthy gradient the caudate

Table 3. Amygdala: $PS_i = \beta_0 + \beta_1 \text{Choice}_{i,m} \times \text{Trust Type} + \beta_2 \text{Perceptual Similarity}_{i,t} \times \text{Trust Type} + \varepsilon$

Coefficient (β)	Estimate (SE)	*t* value	*P* value
Intercept	−0.18 (0.04)	−4.27	<0.001***
Untrustworthy × Perceptual Similarity	0.05 (0.01)	3.64	0.003**
Trustworthy × Perceptual Similarity	−0.017 (0.01)	−0.98	0.32
Untrustworthy × Choice	0.15 (0.08)	1.70	0.09
Trustworthy × Choice	0.143 (0.11)	1.24	0.21

ROI from conjunction of face presentation across all trust types during the initial learning episode (conditioning phase; *SI Appendix*, Table S5). ***$P < 0.001$, **$P < 0.01$.

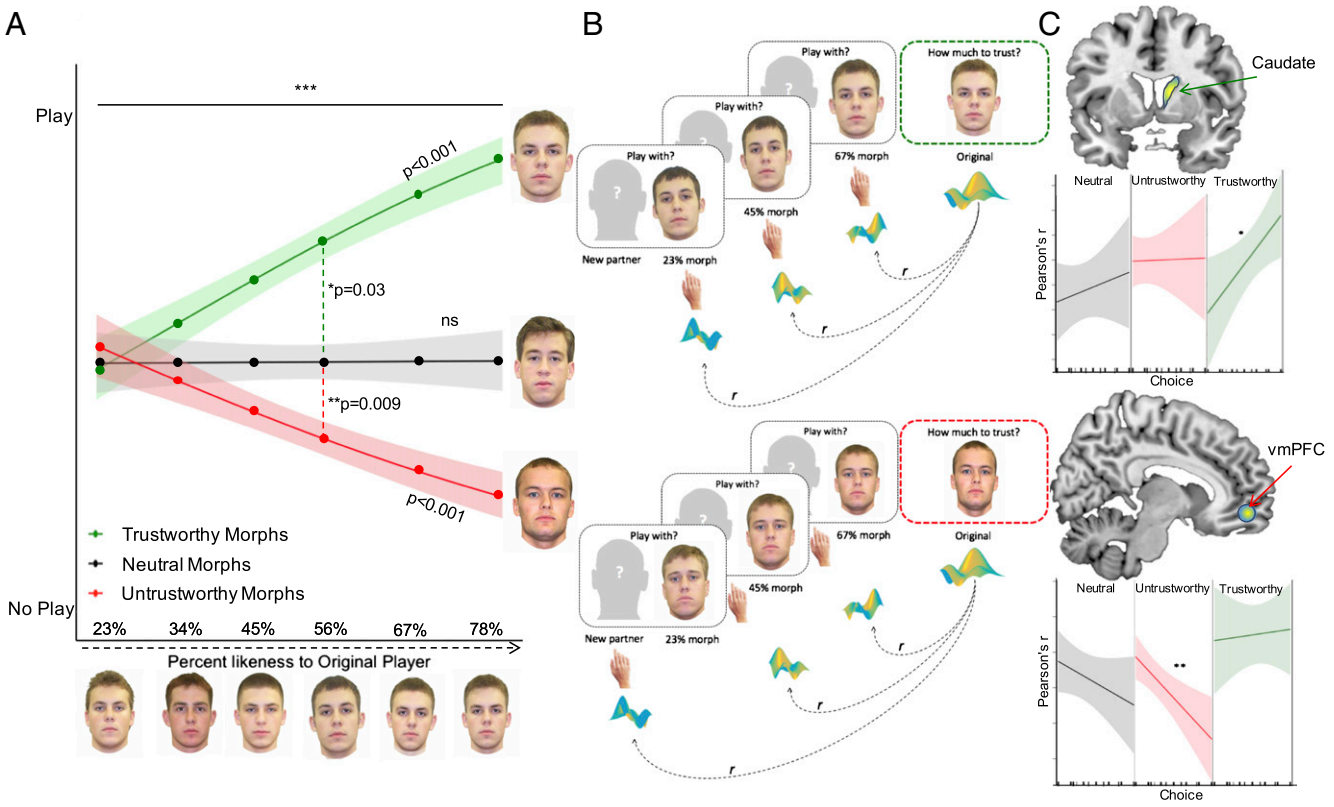

Fig. 4. Neural regions supporting adaptive choice tuning profiles. (*A*) In the imaging experiment, behavioral tuning profiles reveal that subjects incrementally engaged with morphs that look increasingly more like the original trustworthy player and incrementally avoided morphs that look increasingly more like the original untrustworthy player. There was no effect of the neutral gradient ($P > 0.1$). Plotted raw data can be found in *SI Appendix*. Morph gradient is shown with six individuals who were morphed with the trustworthy partner. (*B*) To probe whether adaptive choices relied on similar patterns of activation to the original learning episode we correlated patterns of activity when learning about the trustworthy and untrustworthy players in the conditioning phase with decisions to adaptively engage in, or refrain from, playing with trustworthy or untrustworthy morphs. (*C*) Patterns of activation (derived from PSA) in the caudate support choosing morphs that look like the original trustworthy player. This relationship was only observed in the trustworthy condition. We further observed that patterns of activation in the vmPFC selectively support adaptively refraining from choosing morphs that look like the original untrustworthy player. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$.

selectively exhibited increasingly similar neural patterns that mirrored the observed incremental increases in trusting behavior (Fig. 4*C* and Table 5; results replicate when choice is the dependent variable predicted by the interaction between perceptual similarity and neural pattern similarity, $P = 0.04$). Further interrogating each subject's data at the trial-by-trial level for the morphs that have the greatest perceptual ambiguity (45–56% morph increments, which corresponds to identical visual stimuli across all conditions) provides converging evidence for choices predicting pattern similarity when holding perceptual similarity constant. This highly conservative analysis revealed that deciding to play with the morph predicted increasing pattern similarity in the caudate in the trustworthy condition (*SI Appendix*, Table S14; the vmPFC exhibited a similar pattern but failed to reach significance within the untrustworthy condition, *SI Appendix*, Table S15: see *SI Appendix*).

Discussion

Trust is a basic component of human social life (45). However, little is understood about the cognitive and neural mechanisms supporting learning to trust in the absence of explicit reputational information. Here we demonstrate that a mechanism related to generalizing associative value—a process documented across species—is also deployed in humans in highly complex social decision-making environments. We find that strangers bearing greater resemblance to past individuals previously known to be trustworthy are trusted more and those who resemble individuals previously known to be untrustworthy are trusted less. These behavioral tuning profiles are asymmetrically deployed, whereby individuals are distrusted more when even minimally resembling someone previously associated with untrustworthy and aversive

Table 4. vmPFC: $PS_i = \beta_0 + \beta_1 \text{Choice}_{i,m} \times \text{Trust Type} + \beta_2 \text{Perceptual Similarity}_{i,t} \times \text{Trust Type} + \varepsilon$

Coefficient (β)	Estimate (SE)	*t* value	*P* value
Intercept	0.01 (0.02)	0.77	0.44
Untrustworthy × Perceptual Similarity	0.007 (0.004)	1.31	0.19
Trustworthy × Perceptual Similarity	0.007 (0.01)	0.65	0.51
Untrustworthy × Choice	−0.101 (0.04)	−2.78	0.005**
Trustworthy × Choice	0.01 (0.06)	0.21	0.84

ROI from conjunction of face presentation across all trust types during the initial learning episode (conditioning phase; *SI Appendix*, Table S5). **$P < 0.01$.

Table 5. Caudate: $PS_i = \beta_0 + \beta_1 \text{Choice}_{i,m} \times \text{Trust Type} + \beta_2 \text{Perceptual Similarity}_{i,t} \times \text{Trust Type} + \varepsilon$

Coefficient (β)	Estimate (SE)	*t* value	*P* value
Intercept	−0.06 (0.02)	−2.35	0.019*
Untrustworthy × Perceptual Similarity	0.01 (0.01)	1.41	0.15
Trustworthy × Perceptual Similarity	−0.014 (0.02)	−1.10	0.28
Untrustworthy × Choice	0.046 (0.06)	0.71	0.47
Trustworthy × Choice	0.19 (0.08)	2.48	0.01**

ROI from conjunction of face presentation across all trust types during the initial learning episode (conditioning phase; *SI Appendix*, Table S5). $^{**}P < 0.01$, $^{*}P < 0.05$.

outcomes. Wider generalization in the aversive domain accords with the idea that incorrectly identifying a dangerous stimulus as safe is more costly than treating a safe stimulus as a threat (46). These behavioral tuning profiles were mirrored at the neural level. Within the perceptual domain, the amygdala's functional role in tracking untrustworthiness was supported by both increasing univariate BOLD activity and multivariate activation pattern similarity, such that patterns of neural representation were elicited in a graded fashion along the untrustworthy gradient.

These findings extend previous work linking the amygdala to threat processing and judgments of trustworthiness (32–34), revealing that not only does its hemodynamic activity reflect involvement in extracting information about an individual's untrustworthiness (20), but it also functionally encodes information about untrustworthiness on a representational level. This higher-level functional analysis permits a fine-grained approach for inferring mental states, demonstrating that the amygdala represents critical information about potentially threatening individuals. When perceiving a stranger who looks similar to a past untrustworthy individual the amygdala evokes a pattern of BOLD representations similar to the representations that support initial learning. Evidence of a neural tuning profile tracking perceptual similarity in a graded fashion (15) implies that the amygdala selectively encodes the transfer of negative social value between individuals perceptually resembling one another.

We further observed that neural tuning profiles in the caudate and vmPFC track the behavioral patterns to choose partners who might procure positive outcomes and avoid partners who might yield negative outcomes, respectively. These findings indicate that in the absence of direct and explicit information about an individual's reputation, adaptive decisions to trust or withhold trust rely on activation patterns similar to those elicited when learning about other unrelated, but perceptually familiar, individuals. These choice neural tuning profiles capture behavior above and beyond mere tracking of perceptual similarity, revealing that at the neural level an associative learning mechanism efficiently deploys moral information encoded from past experiences to guide future choice.

Together these findings suggest that when deciding whom to trust humans rely on an efficient, albeit rudimentary, learning heuristic that facilitates adaptive engagement. A similarity-based generalization mechanism can be highly adaptive because it enables many stimuli—in this case, unfamiliar individuals—to acquire value from minimal learning. Even without any direct experience of untrustworthiness individuals implicitly deemed as potentially untrustworthy are systematically avoided. Future work that probes whether initial learning rates bias the efficacy of generalization would be an important next step for understanding whether specific contexts facilitate rapid learning. Importantly, our task is a departure from canonical stimulus generalization tasks in which novel stimuli are often perceived to explicitly overlap with the initial stimulus. Subjects in our experiments believed they were selecting real partners for the next game and therefore treated each potential partner as unique. This indicates that even in the absence of conscious awareness social learning relies on neural processes that make comparisons between current and past experiences to bias decisions about who can be trusted. Ultimately, the finding that complex, dyadic social choices seem to be buoyed by behavioral and neural mechanisms that operate across domains and species suggests that a domain-general system governs many types of emotional learning.

Materials and Methods

Across all experiments 91 subjects were recruited and brought into the laboratory to partake in either a behavioral or imaging task. Subject size of each experiment was based on extant research (13). All subjects were paid $15/h for the behavioral study or $25/h for the imaging study and could make up to an additional $20 based on their decisions during the task. Subjects provided written consent and all experiments were approved by the New York University Committee on Activities Involving Human Subjects. Experimental procedures—which included an iterative trust game with three trustees and a task where subjects could select their own partners for a second trust game—were similar across experiments, with minor exceptions (*SI Appendix*). Individuals presented for the second trust game were morphs of the original trustees, set at 11% increments between the original trustee and a new, neutrally rated face [for a total of eight new morphs, although the morphs closest to the original player (89% increment) and novel individual (12% increment) were not included in the final gradient], which created a linear continuum of increasing perceptual similarity along six morph increments (Fig. 1*C* and *SI Appendix*, Fig. S6). For the imaging experiment, subjects also completed a face localizer task and anatomical scan.

Multivariate PSA was computed in each participant's native space separately for each ROI. Specifically, we assessed the neural similarity between the trustworthy, untrustworthy, and neutral players in the conditioning phase to their respective morphed faces in the generalization phase (see *SI Appendix* for details on how pattern similarity was assessed). This allowed us to probe the representational structure of the activation patterns elicited at each morph increment along the trustworthy, untrustworthy, and neutral gradients, testing whether morphs bearing greater perceptual similarity to the original players would elicit activation patterns that were increasingly similar to those observed during initial learning. Thus, the neural representation of each original player and each morph increment was operationalized as a vector of *t* statistics corresponding to the voxelwise responses in a given ROI. Neural similarity scores for each morph were calculated as the Pearson correlations between that morph's vector and the vector corresponding to the original player during the conditioning phase. Each ROI-specific PS mixed effects linear regression had the same structure and parameters:

$$PS_i = \beta_0 + \beta_1 \text{Choice}_{i,m} \times \text{Trust Type}_t + \beta_2 \text{Perceptual Similarity}_{i,t} \times \text{Trust Type}_t + \varepsilon,$$

where PS is a vector of the correlations (Fisher-transformed values) of the neural pattern similarity between each of the six morph increments and the corresponding original player, for the three trust types, per subject. Choice is indexed by the overall performance at each morph increment (whereby each subject has a composite score of choices across the same trial type, lying between 0 and 1 for each morph increment: 1 indicates choosing the morph and 0 indicates not choosing the morph) and trust type is an indicator variable where 0 = neutral, −1 = untrustworthy, and 1 = trustworthy. This theory driven model captures both the effects of perceptual similarity and adaptive choice in predicting the neural similarity of representational content between each morph and the corresponding original player (ref. 47, p. 776). For further details see *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the National Institute on Aging.

1. Ruff CC, Fehr E (2014) The neurobiology of rewards and values in social decision making. *Nat Rev Neurosci* 15:549–562.
2. Delgado MR, Frank RH, Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci* 8:1611–1618.
3. King-Casas B, et al. (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308:78–83.
4. Chang LJ, Doll BB, van 't Wout M, Frank MJ, Sanfey AG (2010) Seeing is believing: Trustworthiness as a dynamic belief. *Cognit Psychol* 61:87–105.
5. Dayan P, Berridge KC (2014) Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cogn Affect Behav Neurosci* 14:473–492.
6. Rescorla RA, Solomon RL (1967) Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychol Rev* 74:151–182.
7. Rescorla RA (1976) Stimulus generalization: Some predictions from a model of Pavlovian conditioning. *J Exp Psychol Anim Behav Process* 2:88–96.
8. Rescorla RA, Furrow DR (1977) Stimulus similarity as a determinant of Pavlovian conditioning. *J Exp Psychol Anim Behav Process* 3:203–215.
9. Verosky SC, Todorov A (2010) Generalization of affective learning about faces to perceptually similar faces. *Psychol Sci* 21:779–785.
10. Dunsmoor JE, Paz R (2015) Fear generalization and anxiety: Behavioral and neural mechanisms. *Biol Psychiatry* 78:336–343.
11. Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* 237:1317–1323.
12. Andersen SM, Baum A (1994) Transference in interpersonal relations: Inferences and affect based on significant-other representations. *J Pers* 62:459–497.
13. Murty VP, FeldmanHall O, Hunter LE, Phelps EA, Davachi L (2016) Episodic memories predict adaptive value-based decision-making. *J Exp Psychol Gen* 145:548–558.
14. Vanbrabant K, et al. (2015) A new approach for modeling generalization gradients: A case for hierarchical models. *Front Psychol* 6:652.
15. Kahneman D, Tversky A (1972) Subjective probability: A judgment of representativeness. *Cognit Psychol* 3:430–454.
16. Rozin P, Nemeroff C (2007) Sympathetic magical thinking: The contagion and similarity heuristics. *Heuristics and Biases: The Psychology of Intuitive Judgment*, eds Gilovich G, Griffin D, Kahneman D (Cambridge Univ Press, New York).
17. Bouton ME, Doyle-Burr C, Vurbic D (2012) Asymmetrical generalization of conditioning and extinction from compound to element and element to compound. *J Exp Psychol Anim Behav Process* 38:381–393.
18. Schechtman E, Laufer O, Paz R (2010) Negative valence widens generalization of learning. *J Neurosci* 30:10460–10464.
19. Fareri DS, Chang LJ, Delgado MR (2012) Effects of direct social experience on trust decisions and neural reward circuitry. *Front Neurosci* 6:148.
20. Baron SG, Gobbini MI, Engell AD, Todorov A (2011) Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Soc Cogn Affect Neurosci* 6:572–581.
21. Majdandžić J, Amashaufer S, Hummer A, Windischberger C, Lamm C (2016) The selfless mind: How prefrontal involvement in mentalizing with similar and dissimilar others shapes empathy and prosocial behavior. *Cognition* 157:24–38.
22. Waytz A, Zaki J, Mitchell JP (2012) Response of dorsomedial prefrontal cortex predicts altruistic behavior. *J Neurosci* 32:7646–7650.
23. Krueger F, et al. (2007) Neural correlates of trust. *Proc Natl Acad Sci USA* 104: 20084–20089.
24. Baumgartner T, Heinrichs M, Vonlanthen A, Fischbacher U, Fehr E (2008) Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58: 639–650.
25. Castle E, et al. (2012) Neural and behavioral bases of age differences in perceptions of trust. *Proc Natl Acad Sci USA* 109:20848–20852.
26. Rilling JK, Sanfey AG (2011) The neuroscience of social decision-making. *Annu Rev Psychol* 62:23–48.
27. Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis–Connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
28. Dunsmoor JE, Kragel PA, Martin A, LaBar KS (2014) Aversive learning modulates cortical representations of object categories. *Cereb Cortex* 24:2859–2872.
29. Butter R (1965) Stimulus generalization in monkeys with inferotemporal lesions and lateral occipital lesions. *Stimulus Generalization*, ed Mostofsky DI (Stanford Univ Press, Stanford, CA), pp 119–133.
30. Adolphs R, Tranel D, Damasio AR (1998) The human amygdala in social judgment. *Nature* 393:470–474.
31. Adolphs R, Baron-Cohen S, Tranel D (2002) Impaired recognition of social emotions following amygdala damage. *J Cogn Neurosci* 14:1264–1274.
32. Todorov A (2012) The role of the amygdala in face perception and evaluation. *Motiv Emot* 36:16–26.
33. Winston JS, Strange BA, O'Doherty J, Dolan RJ (2002) Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat Neurosci* 5:277–283.
34. Mende-Siedlecki P, Said CP, Todorov A (2013) The social evaluation of faces: A meta-analysis of functional neuroimaging studies. *Soc Cogn Affect Neurosci* 8:285–299.
35. LeDoux JE (2000) Emotion circuits in the brain. *Annu Rev Neurosci* 23:155–184.
36. Morris JS, Ohman A, Dolan RJ (1998) Conscious and unconscious emotional learning in the human amygdala. *Nature* 393:467–470.
37. Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48:175–187.
38. Resnik J, Paz R (2015) Fear generalization in the primate amygdala. *Nat Neurosci* 18: 188–190.
39. Stein MB, Goldin PR, Sareen J, Zorrilla LT, Brown GG (2002) Increased amygdala activation to angry and contemptuous faces in generalized social phobia. *Arch Gen Psychiatry* 59:1027–1034.
40. Haas BW, Ishak A, Anderson IW, Filkowski MM (2015) The tendency to trust is reflected in human brain structure. *Neuroimage* 107:175–181.
41. Tabibnia G, Lieberman MD (2007) Fairness and cooperation are rewarding: Evidence from social cognitive neuroscience. *Ann N Y Acad Sci* 1118:90–101.
42. van den Bos W, van Dijk E, Westenberg M, Rombouts SARB, Crone EA (2009) What motivates repayment? Neural correlates of reciprocity in the trust game. *Soc Cogn Affect Neurosci* 4:294–304.
43. Levy DJ, Glimcher PW (2012) The root of all value: A neural common currency for choice. *Curr Opin Neurobiol* 22:1027–1038.
44. Grahn JA, Parkinson JA, Owen AM (2008) The cognitive functions of the caudate nucleus. *Prog Neurobiol* 86:141–155.
45. Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum Nat* 13:1–25.
46. Bateson M, Brilot B, Nettle D (2011) Anxiety: An evolutionary approach. *Can J Psychiatry* 56:707–715.
47. Sokal RR, Rohlf FJ (1969) *Biometry; The Principles and Practice of Statistics in Biological Research* (Freeman, San Francisco), p xxi.