Confirmation of interpersonal expectations is intrinsically rewarding

Niv Reggev* [1,2,3], Anoushka Chowdhary[1], and Jason P. Mitchell[1]

[1] Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA, US, 0218

[2] Department of Psychology, Ben-Gurion University of the Negev, P.O. Box 653, Be'er-Sheva, Israel, 84105

[3] Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, P.O. Box 653, Be'er-Sheva, Israel, 84105

ORCiD identifiers:

Niv Reggev: 0000-0002-5734-7457

Corresponding author:

Niv Reggev, Department of Psychology, Ben Gurion University of the Negev, P.O. Box 653,

Be'er-Sheva, Israel, 84105

Phone number: +972-8-67472030

Email address: reggevn@bgu.ac.il

This article is a preprint, not yet published in a peer-reviewed journal. We welcome any comments or feedback - please send these to the corresponding author

Abstract

Despite the inherent sociality of human nature, other people pose some of the most difficult challenges to the mind. To successfully interact with other individuals, we need to predict their future responses, a computationally-vexing problem given the enormous range of behaviors in which other people can engage. Decades of research have demonstrated that to simplify this task, perceivers routinely draw on prior beliefs—that is, rather than wait to construct social predictions solely on relevant incoming information, people regularly use prior knowledge, stereotypes, and other sources of information to proactively predict the traits and behaviors of other people. Such research has also demonstrated that once formed, these predictions strongly influence social interactions even when people attempt to change or ignore them. Here, we test the hypothesis that our social predictions resist change because perceivers place high subjective value on having their expectations of others confirmed. Across four studies, we report data consistent with this hypothesis, both when perceivers' expectations derive from gender stereotypes and when they derive from knowledge of familiar individuals. Specifically, in two neuroimaging experiments (n = 58), we observed increased activation in brain regions associated with reward processing—including the nucleus accumbens—when social expectations were confirmed. In two additional behavioral experiments (n = 704), we observed that perceivers were willing to forgo money to encounter an expectation-confirming target and avoid an expectation-violating target. Together, these findings suggest that perceivers value having their social expectations confirmed, much like other primary or secondary rewards.

Keywords: Stereotypes, Expectations, Reward, fMRI, NAcc, Value

Significance statement

People want to interact successfully with other individuals, and they invest significant efforts in attempting to do so. To simplify the dauntingly complex task of interpersonal communication, perceivers use stereotypes and other sources of prior knowledge to predict the responses of individuals in their environment. Here, we show that these top-down expectations can also shape the subjective value of expectation-confirming and expectation-violating targets. To manipulate expectations, we used gender stereotypes and knowledge about US presidents. We observed that participants forgo money to see expectation-confirming targets and that brain regions associated with reward processing show increased activity for such targets. These findings suggest that confirmation of social expectations is intrinsically rewarding, similar to food or monetary rewards.

People dedicate a substantial portion of their time to interacting with other individuals. When we succeed in doing so, we feel better physically and psychologically (1–4). At the same time, other people also present one of the most complicated challenges we have to face. Understanding another person involves inferring hidden states based on fragmentary sensory, verbal and visceral cues, each of which conveys only a small amount of information. To simplify the highly demanding challenge of social cognition, perceivers use top-down predictions (e.g., stereotypes) that help make sense of others in a rapid fashion (5–8). Perceivers can thus seamlessly interact with their environment while refraining from the effortful construction of elaborative representations for each individual they encounter (9, 10). For example, on her first day of school, a freshman might assume that her female peers may be interested in conversing about shopping. Utilizing these predictions, the freshman could effortlessly engage in spontaneous conversations with her new female peers, potentially facilitating her social bonds.

However, although they often facilitate interpersonal interaction, social predictions also impose a cost on perceivers. Individuals tend to adhere to predictions they have previously formed and fail to modify them even in the face of contradictory evidence (11–15). For instance, when perceivers first learn that someone is ‘intelligent’ and then subsequently learn that he is also ‘envious’, they form a favorably-skewed impression of that person; on the other hand, when perceivers learn about these same two character traits in reverse order, they form an unfavorable impression of the target (16, 17). To date, research has not been able to identify the sources that support the persistence of these initial predictions about other individuals. Here, we integrate insights from social psychology and neuroscience to explore the idea that perceivers prefer to ‘stick’ with their initial predictions because they attribute subjective value to the confirmation of these predictions. That is, we posit that targets who confirm our expectations about them (such as stereotype-confirming targets) will trigger a reward-like response similar to food, sex, or chocolate.

Several lines of research already hint at such an effect. Perceivers generally like individuals who conform to expectations more than individuals who violate them (18–20); for example, observers typically prefer female teachers to male teachers, but like male leaders better than their equally competent female peers (21, 22). Similarly, participants express greater trust in targets that fit gender-based predictions (23, 24). Moreover, perceivers demonstrate similar effects for emotion-based expectations, regardless of the valence of the emotion (25). Several theorists have suggested that perceivers may gradually develop a habitual hedonic response for targets conforming to normative expectations (26–29). These suggestions dovetail with the well-documented aversive reactions people experience when confronted with violations of predictions and the uncertainty associated with such violations (15, 29–32). Put together, these positive and aversive responses motivate perceivers to seek expectation-confirming information.

In spite of ample evidence for perceivers' motivation to confirm their social expectations, scholars are still debating the mechanisms supporting the persistence of this motivation. Here we suggest that neural activity can offer a novel insight on this topic. In recent years, scholars have identified the involuntary effects of motivation and expectation in several neural systems, most notably in the mesolimbic dopaminergic system (33). Animal models suggest that midbrain dopaminergic activity signals one's internal desire to obtain a goal (34). In humans, dopamine modulates the motivation to experience positive effects by adjusting midbrain and striatal responses to better- or worse-than-expected information (35–37). Likewise, participants expecting a painful stimulus demonstrate increased striatal activity while experiencing pain, compared with participants who do not expect to feel pain (38, 39). Similarly, stigmatized participants demonstrate different striatal activity compared to non-stigmatized participants (40). Finally, a recent meta-analysis reported that when perceivers agreed with expected group opinions, they demonstrated robust striatal activity, as compared with times in which they deviated from the group consensus (41). These studies suggest that the striatum responds to events that align with perceivers' motivation.

Notably, the involvement of the striatum hints at a potential mechanism mediating the effects of expectation and motivation. Researchers repeatedly identify striatal activity, and most prominently activity in its ventral portion, in anticipation and receipt of various types of reward (42–44). For example, the nucleus accumbens (NAcc), located at the ventral-rostral tip of the striatum, responds both to primary rewards (e.g., food or erotic) and secondary rewards (e.g., money or positive feedback) (45–47). The NAcc also responds to social experiences, such as engagement with attractive or smiling faces, prosocial actions, or placing one's trust in peers (48–51).

Together, these studies suggest that confirmation of stereotypes and other forms of interpersonal predictions is intrinsically rewarding. To test this hypothesis, we first measured NAcc activity in response to confirmation or violation of social expectations. Using functional magnetic resonance imaging (fMRI), we scanned participants while they viewed target individuals who either conformed to or violated interpersonal expectations. As the NAcc is consistently involved in rewarding experiences, activity in this region can serve as a marker of a neural reward response. If perceivers value having their social expectations confirmed, we should observe increased NAcc activity for trials in which targets confirm expectations compared to trials in which targets violate them.

In addition, we assessed whether perceivers actively prefer expectancy-confirming social information by creating experimental situations in which participants could trade money for the chance to view stereotypic targets. To do so, we relied on a modified version of a "pay-per-view" task, previously used with human and non-human primates, to measure the monetary value associated with expectancy-confirming and expectancy-violating stimuli (52, 53). If perceivers experience expectancy-confirming information as intrinsically more valuable, we expected participants to forgo money to interact with expectancy-confirming individuals rather than with expectancy-violating individuals.

Interestingly, social expectations can take multiple forms. For example, stereotypes relate specific attributes to social groups regardless of personal knowledge about group members. Conversely, we can construct detailed individuated expectations about familiar individuals, such as personally familiar others or famous people (10, 14). Accordingly, we also probed whether predictions from these two distinct sources evoked qualitatively different reward responses. Together, the results of four studies support the hypothesis that perceivers are motivated to reaffirm their interpersonal forecasts, in part because they experience the confirmation of social expectations as a powerful form of subjective reward.

Results

Study 1: Neural response to stereotype confirmation vs. violation

In Study 1 we observed greater NAcc activity when participants saw gender-stereotype-confirming targets than when they saw stereotype-violating targets. We scanned participants (*n* = 28) while they formed impressions about targets that varied in the degree to which they confirmed gender stereotypes. Gender stereotypes included various characteristics typically associated with men (e.g., “emotionally closed” or “CEO of a big company”) or women (e.g., “loves children” or “an admired preschool teacher”). In each of 204 trials, participants first read a short description for 1.5 seconds and then saw the face of a target man or woman (see Fig S1A). Participants rated how likely the target was to have the presented characteristic (see SI results and Table S1 for behavioral results). We conducted three parallel analyses to examine whether the neural region most associated with reward — namely, the NAcc — was more engaged when the target confirmed stereotype-derived expectations compared to when the target violated them. First, a whole-brain random-effects contrast identified regions that were more active for *stereotypical* > *counter-stereotypical* trials ($p < 0.05$, corrected; see Table S2 for full results). This analysis indicated a significantly greater response in the NAcc when the presented target matched the stereotypical expectation set by the preceding statement than when the target violated that expectation (Fig. 1A). Parallel results were obtained when we modeled expectation confirmation as a continuous rather than dichotomized predictor (see Table S2 and SI Materials and Methods for full details).

Second, to confirm that this region overlapped with those responsive to rewards, we independently defined a neural regions of interest (ROI) based on spheres around peak voxels identified in a comprehensive meta-analysis (47). Comparing the confirmation and violation of stereotypes in this independently defined region revealed significantly greater activity for confirmation of stereotypes compared to their violation [one-sided test: $t_{(54)} = 3.53$, $p = 0.0004$, *Hedges's g* = 0.37 [95% confidence intervals: 0.1-0.64], Fig. 1B]. Third, we corroborated this finding by defining ROIs from a task in which participants received monetary rewards based on their performance [the Monetary Incentive Delay (MID) task; see Materials and Methods] (54). This analysis yielded similar results [$t_{(54)} = 2.87$, $p = 0.0029$, $g = 0.30$ [0.03-0.59]]. Together, these patterns suggest that seeing a person who confirms a stereotypical expectation triggers activation in the very same region that responds to primary and secondary reinforcers, highlighting the intrinsic value of stereotype confirmation.

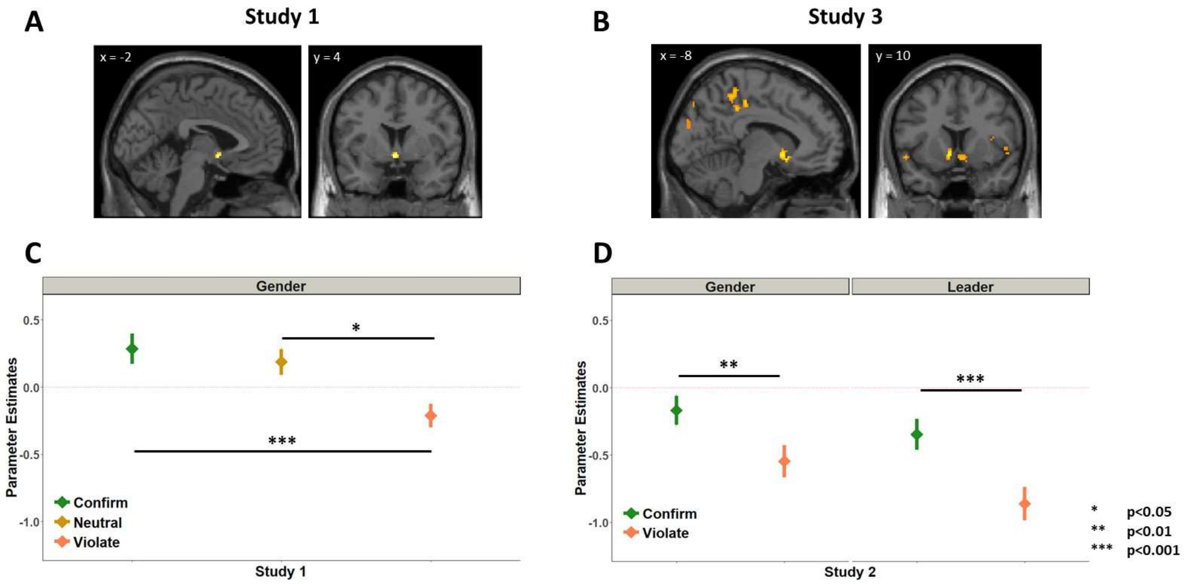


Figure 1. Neural responses associated with the confirmation of expectations about other individuals. (a) Whole-brain random-effects contrasts comparing confirming > violating *trials revealed activity in the nucleus accumbens (NAcc) in (A) Study 1 and (B) Study 3. (C) We independently defined a region of interest in the NAcc using a comprehensive meta-analysis (MNI coordinates: -6, 10, -6; 10, 12, -6). Analysis of parameter estimates in this region confirmed that the bilateral NAcc showed a stronger response during confirming than during violating trials in Study 1 and (D) Study 3. Here and in subsequent figures, error bars depict SE calculated for within-subject designs.*

Additionally, to investigate whether this NAcc activation is limited to stereotype-derived expectations, we included a third type of statements in our study: stereotype-neutral statements (e.g., “drinks coffee every morning”). We found that the overall NAcc response to stereotype-neutral targets was higher than stereotype-violating targets and not different from stereotype-confirming targets [Fig. 1B; Bonferroni corrected comparisons: neutral versus violating: Meta-analysis ROIs: $t_{(54)} = 2.83$, $p = 0.0195$, $g = 0.29$ [0.09-0.51]; MID ROIs: $t_{(54)} = 2.38$, $p = 0.062$, $g = 0.25$ [0.04-0.47]; neutral versus confirming: Meta-analysis ROIs: $t_{(54)} = 0.7$, $p = 0.76$, $g = 0.07$

[-0.19-0.34]; MID ROIs: $t_{(54)} = 0.49$, $p = 0.63$, $g = 0.05$ [-0.18-0.28]]. This finding is consistent with at least two non-mutually-exclusive interpretations. One possible interpretation would suggest that our findings reflect an aversive response to stereotype-violating targets rather than a positive reward value for expectation-confirming targets. Alternatively, our findings could indicate a more general expectation-confirming effect, whereby stereotype-neutral statements created target-specific expectations, expectations that were by and large confirmed. For instance, a statement such as "drinks coffee every morning" could have formed an expectation for a certain type of target (e.g., a young to middle-aged professional). The increased reward response for stereotype-neutral targets described above might indicate that participants in our specific sample had found such targets to generally confirm their idiosyncratic expectations.

To explore these alternative accounts, we used participants' own ratings about the targets as a modulator variable. On each trial, participants had rated the likelihood that the target could be described by the accompanying characteristic (e.g., "enjoys shopping for shoes"). Although gender-relevant characteristics were pretested to be consistently gender-stereotypical, participants naturally did not apply every characteristic to every target (see Table S1 and SI results). This variability provided an opportunity to test whether increased NAcc activation corresponds more closely with the nature of the information that perceivers view or their subjective endorsement (or rejection) of how well that expectation applies to a particular target. Consistent with the second possibility, participants' judgments of specific stereotype-neutral targets were closely related to NAcc response: we observed greater NAcc response when participants endorsed a neutral characteristic ("drinks coffee every morning") than when they rejected its applicability to a target. (Meta-analysis ROIs: $F_{(1,76.18)} = 15.67$, $p = 0.0002$; $\eta_p^2 = 0.17$ [0.06, 0.29]; MID ROIs: $F_{(1,76.5)} = 12.66$, $p = 0.0006$; $\eta_p^2 = 0.14$ [0.04, 0.26]). Targets associated with a stereotype-violating characteristic (e.g., a man who “enjoys shopping for shoes”) elicited a similar pattern of activity in the NAcc (Meta-analysis ROIs: $F_{(1,76.18)} = 17.84$, $p = 0.0001$; $\eta_p^2 = 0.19$ [0.07, 0.31]; MID ROIs: $F_{(1,76.5)} = 15.76$, $p = 0.0002$; $\eta_p^2 = 0.17$ [0.06, 0.29]). Overall, participants’ response significantly modulated the activity of the NAcc for stereotype-violating and stereotype-neutral targets.

Neural activity for stereotype-confirming targets, on the other hand, displayed a different pattern. NAcc response for stereotype-confirming targets was similar regardless of whether or not participants judged the target as likely to possess the characteristic (interaction between response and condition: Meta-analysis ROIs: $F_{(1.77,47.68)} = 5.4$, $p = 0.01$, $\eta_p^2 = 0.17$ [0.02, 0.3]; MID ROIs: $F_{(1.87,50.4)} = 4.01$, $p = 0.03$, $\eta_p^2 = 0.13$ [0.01, 0.26]; see Fig. 2. Stereotype-confirming trials: meta-analysis ROIs: $F_{(1,76.18)} = 0.2$, $p > 0.5$; $\eta_p^2 = 0.003$ [0, 0.05]; MID ROIs: $F_{(1,76.5)} = 0.63$, $p > 0.5$; $\eta_p^2 = 0.005$ [0, 0.06]). This pattern suggests that stereotype-confirming information triggers reward-related neural activity even when one rejects such an association for a specific target individual, hinting at the uncontrollable nature of this response. Thus, NAcc activation was sensitive to the degree to which a participant judges information to apply to a target, but only for stereotype-neutral information or stereotype-violating targets. In contrast, stereotype-confirming information was associated with neural reward regardless of one’s personal willingness to endorse or reject the specific application of the stereotype (see Table S3 and SI results for complimentary exploratory whole-brain analyses).

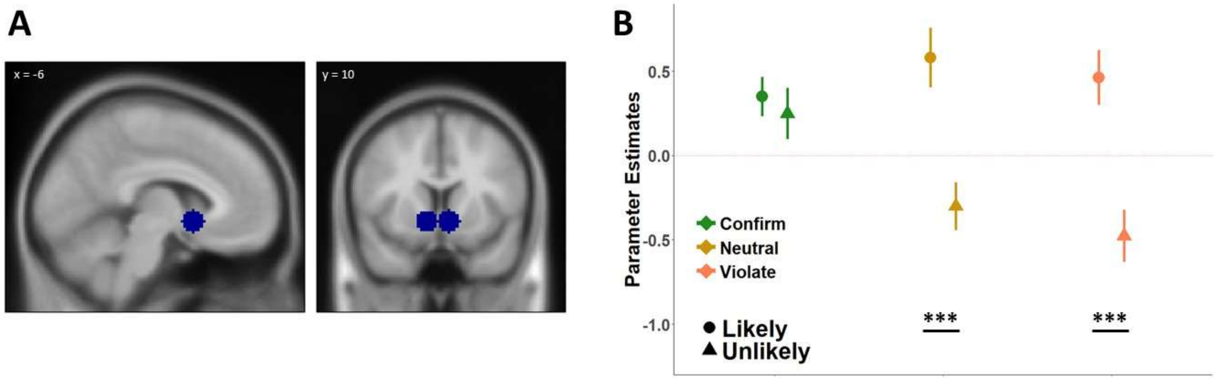


Figure 2. The effect of behavioral ratings of trials on neural responses. In each trial, participants indicated whether they thought the presented target was likely or unlikely to be associated with the presented statement. (A) The independently defined regions of interest in the NAcc. (B) We observed a significant interaction in the independently defined NAcc in Study 1: Participants' ratings modulated the neural response only for stereotype-neutral and stereotype-violating targets, suggesting that stereotype-confirming targets are involuntarily rewarding.

Study 2: The monetary value of stereotype confirmation

Although activation of the NAcc often reflects the presence of rewarding stimuli (55), this region can also respond to non-value related processes including information coding or salience effects (56). To complement our initial neural results, in the preregistered Study 2 we examined a behavioral measure of the value associated with rating expectancy-confirming targets. Specifically, we tested how much money participants were willing to forgo to view stereotype-confirming instead of seeing stereotype-violating targets.

Participants on Amazon Mechanical Turk (*n*s = 174 and 169 in Study 2a and 2b, respectively) made a series of choices to rate one of two target types: a stereotype-confirming target (e.g., a man who enjoys riding motorcycles) or a stereotype-violating target (e.g., a man who enjoys shopping for shoes), designated as "typical" and "atypical" respectively. After each choice, participants saw a target accompanied by a statement and rated the likelihood that the target would be associated with the statement on a 0-100 scale. Participants had up to 5 seconds for each phase of the task (see Fig S2). To avoid potentially different responses to male and female targets, Study 2a included only male faces and Study 2b included only female faces. On each of the 25 trials, small monetary payoffs ($0.03-$0.09 in increments of 2 cents) were associated with each target choice. Participants received a subset of these payoffs as a monetary bonus for the task. Payoff amounts for each target choice varied across trials (and were occasionally equal), as did the location of the option for which participants received the larger amount. If confirmation of stereotypes is intrinsically rewarding, participants should be willing to forgo money — that is, choose the lower-paying option — to see stereotype-confirming individuals. On the other hand, a participant seeking to maximize monetary payoff should consistently choose the higher paying option regardless of the stereotypicality of the information that follows.

We modeled the relative value of each target type by calculating the point of subjective equivalence (PSE) between the two options. This value was derived by fitting a cumulative normal distribution curve to participants' choices (Fig. 3A) and finding the monetary value at which participants effectively chose arbitrarily between the two target types (57). Thus, the PSE represents the relative monetary value of one target type over another.

As predicted, participants demonstrated a significant preference for seeing stereotype-confirming targets over stereotype-violating targets. When the two trial types shared the same payoff amounts, participants chose the stereotypical targets 58% and 57% of the time for male and female targets, respectively (significantly more than chance, as indicated in a generalized mixed model analysis by an odds ratio of 1.4, $Z = 3.11$, $p = 0.0019$ and an odds ratio of 1.41, $Z = 2.91$, $p = 0.00369$ for the zero-centered intercept in Study 2a and 2b, respectively). Moreover, the calculated PSE indicated that participants forewent an average of 0.34 and 0.335 cents per trial to rate a stereotype-confirming over a stereotyple-violating target in Study 2a and 2b, respectively

(95% Confidence Interval [CI]: 0.21-0.47, $t_{(173)} = 5.17$, $p < 0.0001$, *Cohen's d* = 0.39 [0.24-0.55] and CI: 0.25-0.43, $t_{(168)} = 7.3$, $p < 0.0001$, $d = 0.56$ [0.4-0.72] for the two studies, respectively; Fig 3b). This PSE resulted in an average loss of 10% of potential earnings, as participants chose lower monetary amounts to view stereotype-confirming targets. Just as non-human primates prefer to view dominant groupmates over receiving juice (52) and students are willing to forgo money to talk about themselves (53) or to view attractive members of the opposite sex (58), our participants gave up money to view information that was in line with their stereotypical expectations.

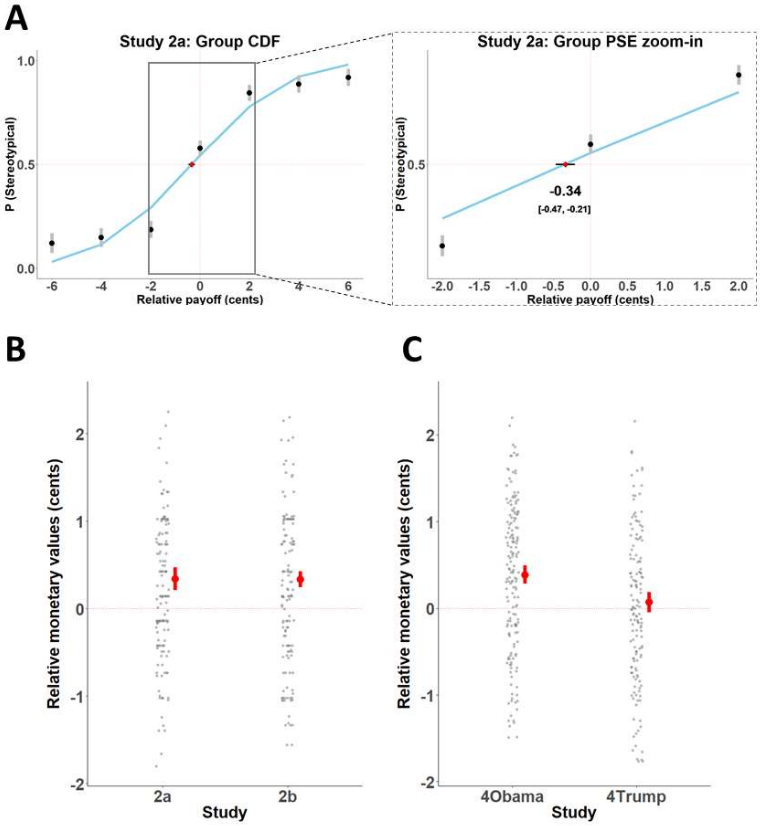


Figure 3. The monetary values of confirmation of stereotype and person-specific expectations. (A) Visualization of the cumulative distribution function we used to calculate the Point of Subjective Equivalence (PSE), illustrated by group data from Study 2a. The x-axis represents the difference between the monetary values associated with the two target types presented in each trial. Each dot indicates the proportion of trials in which participants chose to rate a stereotype-confirming over a stereotype-violating target. The PSE was calculated as the point at which a cumulative normal distribution function, fit to these responses, passes 50%. This point represents the relative monetary value associated with one target type over the other. Negative values indicate that participants preferred to incur a relative monetary loss to rate a confirming target. Error bars depict 95% confidence intervals. (B) Distribution of individual PSE values for Study 2. Rating stereotype-confirming targets in Studies 2a and 2b was associated with significantly higher subjective value than rating stereotype-violating targets. Each gray dot depicts PSE for a specific participant. Red dots indicate the sample mean. Error bars depict 95% confidence intervals. (C) Distribution of individual PSE values for Study 4. Rating Obama-confirming trials was associated with significantly higher subjective value than rating Obama-violating trials. However, rating Trump-confirming trials did not significantly differ from rating Trump-violating trials.

Study 3: Neural response to interpersonal expectations

Together, Studies 1 and 2 suggest that perceivers experience the confirmation of social expectations as intrinsically rewarding. However, we designed these studies primarily to test the effects of a specific type of social expectations – gender-based stereotypes. Although stereotypes are a significant source of interpersonal expectations, perceivers routinely make use of additional, idiosyncratic sources of information, especially for individuals with whom they are highly familiar. Does the reward value of expectancy-confirming information extend to person-specific predictions?

To examine this question, Study 3 assessed the responses of the neural reward system to the confirmation and violation of expectations regarding two highly familiar targets, the current and previous presidents of the United States at the time of the study: Donald Trump and Barack Obama, respectively. To facilitate comparison to Study 1, the preregistered Study 3 also included stereotype-derived expectations about unfamiliar targets. As in Study 1, participants (*n* = 30) rated how likely each of 240 specific statements described a specific man or woman or described Donald Trump or Barack Obama. We presented stereotype-related and person-specific statements in blocks of 15 trials per content domain for a total of 60 trials per condition (confirming/violating targets in the stereotype/person-specific domain). Person-specific statements included various characteristics typically associated with one – but not the other – leader (e.g., “Supports a wall along the borders” versus “Acts to support women’s rights”). Participants first read a short statement for 1.5 seconds and then saw the face of a target (a man, a woman, Trump or Obama; see Fig S1C). Participants rated how likely the statement was to describe the target (see Table S4 and SI Results for full details) .

We conducted two parallel analyses to examine whether the NAcc was more engaged when presented with information that confirmed expectations than with information that violated them. First, a whole-brain random-effects contrast identified regions that were more active for *expectation-confirming* > *expectation-violating* trials ($p < 0.05$, corrected; see Table S5 for full results). This analysis indicated significantly greater response in several regions, including the NAcc, when the target confirmed the expectation set by the preceding statement than when the target violated that expectation (Fig. 1B).

In a second analysis we defined the NAcc via two independent procedures, first as bilateral spheres around peak voxels identified in a meta-analysis, and then by examining participants' neural responses during the MID task. Across both procedures the NAcc demonstrated a robust difference in activity between expectation-confirming and expectation-violating targets (Fig. 1D). However, we did not observe any difference between stereotypes and person-specific trials in any of the analyses. Specifically, a 2 X 2 repeated measures ANOVA over activity in bilateral NAcc revealed a main effect of expectation-confirmation, with higher activation associated with expectation-confirmation compared to expectation-violation (meta-analysis ROIs: $F_{(1,29)} = 21.89$, $p < 0.0001$, $\eta^2_p = 0.43$ [0.19-0.58]; MID ROIs: $F_{(1,29)} = 11.2$, $p = 0.002$, $\eta^2_p = 0.28$ [0.07-0.46]). Activation did not significantly differ between stereotype and person-specific content (meta-analysis ROIs: $F_{(1,29)} = 3.64$, $p = 0.07$, $\eta^2_p = 0.11$ [0-0.29]; MID ROIs: $F_{(1,29)} = 1.41$, $p = 0.24$, $\eta^2_p = 0.05$ [0-0.2]) and no interaction was observed between the factors (meta-analysis ROIs: $F_{(1,29)} = 0.31$, $p = 0.58$, $\eta^2_p = 0.01$ [0-0.13]; MID ROIs: $F_{(1,29)} < 0.01$, $p = 0.99$, $\eta^2_p < 0.0001$). Accordingly, expectation-confirmation yielded more neural activity for each of the two content domains when examined separately (meta-analysis ROIs: $F_{(1,54.8)} = 5.96$, $p = 0.0179$, $\eta^2_p = 0.1$ [0.01-0.23] and $F_{(1,54.8)} = 11.03$, $p = 0.0016$, $\eta^2_p = 0.17$ [0.04-0.31] for gender stereotypes and person-specific expectations, respectively; MID ROIs: $F_{(1,55.47)} = 4.36$, $p = 0.04$, $\eta^2_p = 0.07$ [0.001-0.2] and $F_{(1,55.47)} = 4.45$, $p = 0.04$, $\eta^2_p = 0.07$ [0.002-0.2], respectively). Thus, confirmation of social expectations triggered more activation in the NAcc than violation of such expectations, regardless of whether source of the expectation was general social knowledge (stereotypes) or person-specific knowledge.

Finally, similar to Study 1, variability in behavioral responses in Study 3 enabled us to test the effects of subjective endorsement of characteristics for stereotype-based characteristics (but not for person-specific-expectations; see SI results). As in Study 1, NAcc activity was not modulated by subjective endorsement of characteristics for stereotype-confirming targets (simple effects analyses: meta-analysis ROIs: $F_{(1,58)} = 3.99$, $p = 0.0503$; $\eta_p^2 = 0.06$ [0, 0.18]; MID ROIs: $F_{(1,56.5)} = 3.20$, $p = 0.08$; $\eta_p^2 = 0.05$ [0, 0.17]). However, we did not observe an interaction between response and condition (meta-analysis ROIs: $F_{(1,29)} = 0.08$, $p > 0.5$, $\eta_p^2 = 0.003$ [0, 0.09]; MID ROIs: $F_{(1,29)} = 0.40$, $p > 0.5$, $\eta_p^2 = 0.01$ [0, 0.14]). In other words, unlike in Study 1, NAcc activation was *in*sensitive to the degree to which a participant judges information to apply to a target, regardless of the stereotypicality of the target. Together, these results suggest that confirmation of expectations about other targets is valuable for the two most dominant sources of interpersonal expectations – group-based stereotypes and person-specific knowledge.

Study 4: The monetary value of person-specific expectation-confirmation

In Study 2, we observed that perceivers are willing to forgo money to view stereotype-confirming (rather than stereotype-violating) information. To examine whether this behavioral effect extends to expectations about specific individuals, Study 4 replicated the procedure from Study 2 using familiar individuals (Obama and Trump). On each of 32 trials, participants on Prolific Academic (n = 189 and 172 in Study 4a and 4b, respectively) first chose between seeing either an expectation-confirming or an expectation-violating target and then rated the target. Study 4a included only Obama as the target of statements and Study 4b included only Trump. After choosing the type of content they would like to see, participants rated the likelihood that a specific statement would be associated with the target on a 0-100 scale. Payoff amounts for each choice varied across trials (and were occasionally equal), as did the option for which participants received the larger amount. As in Study 2, we quantified the subjective monetary value of each option by calculating the PSE between the two display types by fitting a cumulative normal distribution curve to participants' choices and finding the monetary value at which participants were indifferent to the two options. If participants experience confirmation of expectation as equally rewarding regardless of the target of expectations, then they should choose to forgo money to rate statements consistent with Obama and Trump in Studies 4a and 4b, respectively.

The results from Study 4a indicate that participants preferred to see expectation-confirming statements of Barack Obama. When the payoff amounts were equal for confirming and violating statements, participants chose the confirming statements 60% of the time (significantly more than chance, as indicated in a generalized mixed model analysis by an odds ratio of 1.42, Z = 4.74, $p < 0.0001$ for the zero-centered intercept in Study 4a). Moreover, the calculated PSE indicated that, on average, participants gave up 0.39 cents per trial to rate an expectancy-confirming over an expectancy-violating statement about Obama (95% [CI]: 0.28-0.49, $t_{(188)}$ = 7.21, $p < 0.0001$, d = 0.52 [0.37-0.68]; Fig 3C). However, the same was not true for Trump-related statements in Study 4b. At equal payoff amounts, participants had no preference between expectancy-confirming and -violating statements (choosing the confirming options 49% of the time; odds ratio of 1.01, z = 0.13, p = 0.9), and the calculated PSE was not different from 0 (0.07 cents; [-0.05-0.19], $t_{(173)}$ = 1.17, p = 0.12, d = 0.09 [-0.06-0.24]; Fig 3C). The difference between the studies was significant ($t_{(328.29)} = 2.28$, $p = 0.023$, $d = 0.25$ [0.03-0.46]). Together, these findings suggest that the rewarding effect of expectation-confirmation is not limited to stereotypes, but also applies to knowledge about specific individuals. Notably, however, not all sources of knowledge equally contribute to the reward value; people were not willing to forgo monetary amounts to interact with content confirming their expectations about Donald Trump (see SI results for potentially related findings in Study 3).

Discussion

The human preference for consistent and predictable social interactions has long been acknowledged as a core motivational component driving everyday behavior (31, 32). To predict the behavior of others, perceivers regularly employ biased strategies to collect and interpret information that corresponds to their expectations (59–63). Here we provide evidence to suggest that humans associate expectation-confirmation with intrinsic value, much like other forms of reward such as food or money. Our findings suggest that this reward value is generated regardless of the source of the social expectation. Participants were willing to forgo money to rate an expectation-confirming target rather than its expectancy-violating counterpart. Moreover, doing so was associated with increased activity in a brain region in the neural reward circuitry, regardless of whether the target confirmed gender stereotypes or knowledge about US presidents. Put simply: people find it rewarding to have their expectations (stereotypical or idiosyncratic) confirmed.

This line of research coincides with emerging theories that highlight the instrumental value of behaviors and perceptions that fall in line with our expectations. In an early example of this value, Allport (64) described a mental process in which ‘A Scotsman who is penurious delights us because he vindicates our prejudgment’ (p.22). Some recent theories suggest that, because most of our social expectations are anchored in our social environment, repeated interaction with expectancy-confirming information leads to continuous reinforcement of our expectations (27, 63, 65). Once established, these expectations induce motivations and cognitive representations that persist even in the face of disconfirmation (28, 66, 67). One theory further suggests that the metabolic costs associated with the violation of expectations increase the desirability of expectation-confirming behavior from an evolutionary standpoint (29). Consistent with these theories, the current studies demonstrate that the confirmation of social expectations is indeed associated with subjective value.

The current findings provide a neural extension to prominent accounts of implicit (i.e., involuntary) stereotyping and prejudice (68, 69). Group-based stereotypes typically draw on categorical distinctions to facilitate easier decision making by enabling faster and more efficient processing of stereotype-confirming information (15). Downstream, perceivers evaluate expectation-confirming individuals more positively, allocate more economic resources to them, and judge them as more hirable (18, 19, 23). Complementarily, violations of social expectations pose a threat to individuals and social structures alike, which, in turn, often try to eliminate the threat and reinforce the original expectation (70). Our results provide a candidate mechanism for these effects, whereby the preference of expectation-confirming information translates into a subjective value that shapes how we evaluate specific individuals (cf., 71).

Finally, our results also hint at when perceivers can assign value to expectation-violating information. In our task, behavior asymmetrically affected the neural response in the nucleus accumbens. Whereas stereotype-confirming targets always evoked the same level of neural activity regardless of participants' responses, in Study 1 stereotype-violating targets elicited enhanced NAcc activity only if perceivers judged them as likely to be associated with the expectancy. This pattern suggests that participants can assign value to stereotype-violating targets, depending on the believability of the counter-stereotypical judgment. The exploratory whole-brain analysis (see SI Results) raised the possibility that for these targets, participants experienced reward when they invoked control to modulate their initial responses. In line with this suggestion, recent findings highlight the intrinsic subjective value of exerting control over overt responses (72, 73). If the value of control can be associated with stereotype-violating targets, then future interactions with those targets might lead to a change in the expectancies they were originally violating.

These findings join a growing body of literature that characterizes how our prior beliefs modulate information-processing to fortify a world view and protect established expectations (36, 39, 66, 74, 75). We suggest that the subjective value imbued upon targets who conform to societal expectations may serve to sustain multiple stereotype- and expectation-induced biases. To mitigate the negative implications associated with these expectations, society will need to acknowledge the subjective value associated with their confirmation.

Materials and Method

Materials. To create expectation-setting statements we generated a list of verbal statements that described relevant individual preferences, traits, behaviors or professions. Studies 1 and 2 included 136 gender-related and 68 gender-neutral statements (see SI Materials and Methods). We verified the stereotypicality of these statements in a pilot study (*n* = 78) in which participants from the local community indicated how typical the characteristic was for a specific gender on a visual scale of 0 (”very untypical”) to 100 (”very typical”; the scale had no other tick marks). Each participant was randomly assigned to rate each statement either for men or for women. Participants were instructed to base their ratings on how they thought the average person would respond. This verification procedure was successful; men were associated with men-stereotypic statements more than women (mean difference: 25.7) and women were associated with women-stereotypic statements more than men (mean difference: 25.1). Overall, the statements contained 2-9 words (mean: 4.69, S.D: 1.44; no difference between experimental conditions, $p > 0.2$) and 9-45 characters (mean: 26.68, S.D.: 7.69; $p > 0.18$). Studies 3 and 4 further included 120 person-specific statements pertaining to Barack Obama and Donald Trump. A total of 243 participants from Amazon Mechanical Turk rated a sample of 60 of these statements, randomly determined per participant. On each trial participants indicated how typical the presented characteristic was for the two targets (a separate scale for each target; the two scales were presented simultaneously with a randomly determined order). Obama was associated with Obama-related statements more than Trump (mean difference: 56.1) and Trump was associated with Trump-related statements more than Obama (mean difference: 54.5). Overall, the statements contained 2-9 words (mean: 4.65, S.D: 1.52; no difference between experimental conditions, $p > 0.5$) and 11-45 characters (mean: 28.33, S.D.: 8.43; $p > 0.5$).

Studies 1 and 3. Twenty-eight individuals participated in Study 1 and 30 individuals participated in Study 3. Additional participants were excluded due to excessive motion, technical issues or lack of response to more than 20% of trials (3 and 6 participants from Studies 1 and 3, respectively). All participants provided informed consent in a manner approved by the Committee on the Use of Human Subjects in Research at Harvard University. Study 3 was pre-registered (https://osf.io/h9c6x/?view_only=6fa03fc0ceb04e2083db9e485dbe6615). See SI for demographic information. The current sample size allowed a power of 0.8 to detect a medium effect size (*Cohen's d* = 0.5) in the planned one-tailed analysis at the region of interest. In both studies participants formed impressions about target individuals. On each trial, participants first saw a statement for 1.5 seconds. The statements in Study 1 described a stereotypically neutral, stereotypically male, or stereotypically female characteristic; statements in Study 3 described a characteristic which was either stereotypically male or female or closely associated with Barack Obama or Donald Trump (see https://osf.io/tgja3/?view_only=d849bad1b606474f80c1a3e40e740875 for open materials). Next, participants saw the statement with a face of a man or a woman (in Study 1; Study 3 also included face images of the relevant leaders). The statement-face pair appeared on screen for 4 additional seconds (3.5 seconds in Study 3) for a total of 5.5 seconds (5 seconds) per trial. Each trial ended with a 0.5 second fixation crosshair. Participants used their left hand to indicate how likely the presented target was to be described by the specific characteristic using a 4-point scale (1- ‘very unlikely’; 4 -‘very likely’). Participants responded while the pair appeared on screen. Trials in both studies were separated by variable intertrial intervals of 0-9s (76) optimized for our contrast of interest (see SI Materials and Methods for details).

To localize brain regions associated with the processing of rewarding stimuli, we defined 8-mm spheres around peak coordinates drawn from a comprehensive meta-analysis (47). To functionally identify these brain regions, participants in both studies completed a Monetary Incentive Delay (MID) task (54) immediately after the impression formation task (see Fig. S1B and SI Materials and Methods for details). This task allows the identification of monetary-reward-sensitive ROIs by comparing trials in which participants won money to trials in which participants could not earn any reward. We extracted and averaged parameter estimates across voxels in each ROI per condition of interest and analyzed them in a repeated-measures ANOVA.

We collected neuroimaging data with a 3T Siemens Prisma scanner system (Siemens Medical Systems, Erlangen, Germany). First, we acquired high-resolution anatomical images using a T1-

weighted 3D MPRAGE sequence. Next, whole brain functional images were collected using a simultaneous multi-slice (multiband) T2*-weighted gradient echo sequence (TR = 2000 msec, TE = 30 msec, voxel size = 2 × 2 × 2 mm3, 75 slices auto-aligned to -25 degrees of the AC-PC line). Participants completed four impression formation task runs consisting of 229 volumes each (245 volumes in Study 3). Finally, participants completed the MID task in a single run consisting of 110 volumes using identical parameters to those mentioned above. We used SPM12 version 6225 (Wellcome Department of Cognitive Neurology, London, UK) to process and analyze the fMRI data. Data were corrected for differences in acquisition time between slices, corrected for inhomogeneities in the magnetic field using fieldmap (77), realigned to the first image to correct for head movement, unwarped to account for residual movement-related variance and co-registered with each participant's anatomical data. Functional data were then transformed into a standard anatomical space (2 mm isotropic voxels) based on the ICBM152 brain template (Montreal Neurological Institute). Normalized data were then spatially smoothed (6 mm full-width at half-maximum, FWHM) using a Gaussian Kernel (see SI Materials and Methods for full details of scanning and analysis procedures). We analyzed preprocessed data using a general linear model in which we modeled trials as boxcar functions with an onset at face presentation (1.5s after statement presentation) and with variable duration determined per trial by reaction time to control for effects of reaction time on the neural response (78). Our main analysis included a model in which we conditionalized trials based on trial type (stereotype-confirming, stereotype-neutral or stereotype-violating trials in Study 1; stereotype-confirming, stereotype-violating, person-specific-confirming and person-specific-violating trials in Study 3). In our secondary analysis (Fig 2) we split each of the trial types included in the main analysis with two regressors, one in which participants provided a "Likely" rating and one in which they provided an "Unlikely" rating. We convolved events with a canonical hemodynamic response function and its temporal derivative and included additional covariates of no interest (session mean, no response trials, six motion parameters and their temporal derivative). The final first-level GLM was high-pass filtered at 128 s. Analyses were performed individually for each participant, and contrast images were subsequently entered into a second-level analysis treating participants as a random effect. We report activations that survived a threshold of p<0.001 (uncorrected) at the voxel level and (cluster-size) corrected to p<0.05 at the cluster level using Monte Carlo simulations (1,000 iterations) with the current imaging and analysis parameters (79).

Studies 2 and 4. A total of 343 participants were included in Study 2 (174 in Study 2a) and 361 in Study 4 (189 in Study 4a). Additional participants were excluded by criteria set in the pre-registered protocols for each study (see SI Materials and Methods for details; see also https://aspredicted.org/blind.php?x=rt3wi6 (Study 2a), https://aspredicted.org/blind.php?x=db3gz4 (Study 2b), and https://aspredicted.org/blind.php?x=9z83yb (Studies 4a and 4b)). Sample size was set to allow sufficient power (0.8) to detect a small effect size (*Cohen's d* = 0.2) in a one-sample t-test for each study. Informed consent was obtained from all participants in a manner approved by the Committee on the Use of Human Subjects at Harvard University. We calculated the PSE (and the related SEs) for the difference between the two target types for each study by the Delta Method as implemented in the MixedPsy package (57, 80) for R. The Delta Method relies on responses to all trials aggregated across participants in a generalized linear model to approximate the PSE with a Gaussian distribution (57) and to plot the cumulative distribution function. The model included the difference between the two target types as the predictor value and a binary outcome (stereotypical/knowledge-confirming option chosen) as the predicted value, with the probit link.

Acknowledgements

This work was partially supported by the Israeli Science Foundation (No. 79/18, to N.R.). N.R was supported by Yad Ha'Nadiv (Rothschild) Foundation and the Mind, Brain and Behavior interfaculty program at Harvard University. Parts of this work were conducted at the Harvard Center for Brain Science, which is supported by the NIH Shared Instrumentation Grant Program (No. S10OD020039).

References

1. L. Tay, K. Tan, E. Diener, E. Gonzalez, Social relations, health behaviors, and health outcomes: A survey and synthesis. *Appl. Psychol. Heal. Well-Being* **5**, 28–78 (2013).

2. R. F. Baumeister, M. R. Leary, The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* **117**, 497–529 (1995).

3. J. T. Cacioppo, S. Cacioppo, D. I. Boomsma, Evolutionary mechanisms for loneliness. *Cogn. Emot.* **28**, 1–22 (2014).

4. G. A. Matthews, K. M. Tye, Neural mechanisms of social homeostasis. *Ann. N. Y. Acad. Sci.*, 1–21 (2019).

5. M. Otten, A. K. Seth, Y. Pinto, A social Bayesian brain: How social knowledge can shape visual perception. *Brain Cogn.* **112**, 69–77 (2017).

6. J. B. Hutchinson, L. F. Barrett, The power of predictions: An emerging paradigm for psychological research. *Curr. Dir. Psychol. Sci.* **28**, 280–291 (2019).

7. D. I. Tamir, M. A. Thornton, Modeling the predictive social mind. *Trends Cogn. Sci.* **22**, 201–212 (2018).

8. J. B. Freeman, K. L. Johnson, More than meets the eye: Split-second social perception. *Trends Cogn. Sci.* **20**, 362–374 (2016).

9. C. N. Macrae, G. V. Bodenhausen, Social cognition: Thinking categorically about others. *Annu. Rev. Psychol.* **51**, 93–120 (2000).

10. S. T. Fiske, S. L. Neuberg, “A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation.” in *Advances in Experimental Social Psychology*, M. P. Zanna, Ed. (Academic Press, 1990), pp. 1–74.

11. A. P. Gregg, B. Seibt, M. R. Banaji, Easier done than undone: Asymmetry in the malleability of implicit ireferences. *J. Pers. Soc. Psychol.* **90**, 1–20 (2006).

12. J. E. Dunsmoor, J. T. Kubota, J. Li, C. A. O. Coelho, E. A. Phelps, Racial stereotypes impair flexibility of emotional learning. *Soc. Cogn. Affect. Neurosci.* **11**, 1363–1373 (2016).

13. N. A. Wyer, You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Soc. Cogn.* **28**, 1–19 (2010).

14. D. L. Hamilton, S. J. Sherman, Perceiving persons and groups. *Psychol. Rev.* **103**, 336–355 (1996).

15. N. J. Roese, J. W. Sherman, “Expectancy” in *Social Psychology: Handbook of Basic Principles*, 2nd Ed., E. T. Higgins, A. W. Kruglanski, Eds. (The Guilford Press, 2007), pp. 91–115.

16. S. E. Asch, Forming impressions of personality. *J. Abnorm. Soc. Psychol.* **41**, 258–290 (1946).

17. J. Sullivan, The primacy effect in impression formation: Some replications and extensions. *Soc. Psychol. Personal. Sci.* **10**, 432–439 (2019).

18. J. E. Phelan, L. A. Rudman, Reactions to ethnic deviance: The role of backlash in racial stereotype maintenance. *J. Pers. Soc. Psychol.* **99**, 265–281 (2010).

19. C. Stern, T. V. West, N. O. Rule, Conservatives negatively evaluate counterstereotypical people to maintain a sense of certainty. *Proc. Natl. Acad. Sci.* **112**, 15337–15342 (2015).

20. A. H. Eagly, S. J. Karau, Role congruity theory of prejudice toward female leaders. *Psychol. Rev.* **109**, 573–598 (2002).

21. L. A. Rudman, C. A. Moss-Racusin, J. E. Phelan, S. Nauts, Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *J. Exp. Soc. Psychol.* **48**, 165–179 (2012).

22. C. A. Moss-Racusin, E. R. Johnson, Backlash against male elementary educators. *J. Appl. Soc. Psychol.* **46**, 379–393 (2016).

23. C. Stern, N. O. Rule, Physical androgyny and categorization difficulty shape political conservatives' attitudes toward transgender people. *Soc. Psychol. Personal. Sci.* **9**, 24–31 (2018).

24. M. Olszanowski, O. K. Kaminska, P. Winkielman, Mixed matters: fluency impacts trust ratings when faces range on valence but not on motivational implications. *Cogn. Emot.* **32**, 1032–1051 (2018).

25. L. Chanes, J. B. Wormwood, N. Betz, L. F. Barrett, Facial expression predictions as drivers of social perception. *J. Pers. Soc. Psychol.* **114**, 380–396 (2018).

26. J. Jost, O. Hunyady, The psychology of system justification and the palliative function of ideology. *Eur. Rev. Soc. Psychol.* **13**, 111–153 (2003).

27. B. Huebner, "Implicit bias, reinforcement learning, and scaffolded moral cognition" in *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, M. Brownstein, J. Saul, Eds. (Oxford University Press, 2016), pp. 47–79.

28. K. C. Berridge, From prediction error to incentive salience: Mesolimbic computation of reward motivation. *Eur. J. Neurosci.* **35**, 1124–1143 (2012).

29. J. E. Theriault, L. Young, L. F. Barrett, The sense of should: A biologically-based framework for modeling social pressure. *Phys. Life Rev.* (2020) https:/doi.org/10.1016/j.plrev.2020.01.004.

30. O. Feldmanhall, A. Shenhav, Resolving uncertainty in a social world. *Nat. Hum. Behav.* **3**, 426–435 (2019).

31. L. Festinger, *A theory of cognitive dissonance (Vol. 2)* (Stanford university press, 1957).

32. B. Gawronski, Back to the future of dissonance theory: Cognitive consistency as a core

motive. *Soc. Cogn.* **30**, 652–668 (2012).

33. A. Kohli, *et al.*, Using Expectancy Theory to quantitatively dissociate the neural representation of motivation from its influential factors in the human brain: An fMRI study. *Neuroimage* **178**, 552–561 (2018).

34. K. C. Berridge, From prediction error to incentive salience : mesolimbic computation of reward motivation. *Eur. J. Neurosci.* **35**, 1124–1143 (2012).

35. T. Sharot, M. Guitart-masip, C. W. Korn, R. Chowdhury, R. J. Dolan, How Dopamine enhances an optimism bias in humans. *Curr. Biol.* **22**, 1477–1481 (2012).

36. C. J. Charpentier, E. S. Bromberg-Martin, T. Sharot, Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proc. Natl. Acad. Sci.* **115**, E7255–E7264 (2018).

37. G. Lefebvre, M. Lebreton, F. Meyniel, S. Bourgeois-Gironde, S. Palminteri, Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1**, 1–9 (2017).

38. K. A. Schwarz, *et al.*, How stereotypes affect pain. *Sci. Rep.* **9**, 8626 (2019).

39. M. Jepma, L. Koban, J. Doorn, M. Jones, T. D. Wager, Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nat. Hum. Behav.* **2**, 838–855 (2018).

40. B. L. Welborn, Y. Hong, K. G. Ratner, Exposure to negative stereotypes influences representations of monetary incentives in the nucleus accumbens. *Soc. Cogn. Affect. Neurosci.*, 261–271 (2020).

41. H. Wu, Y. Luo, C. Feng, Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* **71**, 101–111 (2016).

42. W. Schultz, Multiple reward signals in the brain. *Nat. Rev. Neurosci.* **1**, 199–207 (2000).

43. S. N. Haber, B. Knutson, The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology* **35**, 4–26 (2010).

44. T. A. Hare, J. P. O'Doherty, C. F. Camerer, W. Schultz, A. Rangel, Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J. Neurosci.* **28**, 5623–5630 (2008).

45. G. Sescousse, X. Caldú, B. Segura, J.-C. Dreher, Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neurosci. Biobehav. Rev.* **37**, 681–696 (2013).

46. J. Peters, C. Büchel, Neural representations of subjective reward value. *Behav. Bran Res.* **213**, 135–141 (2010).

47. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).

48. A. Lin, R. Adolphs, A. Rangel, Social and monetary reward learning engage overlapping neural substrates. *Soc. Cogn. Affect. Neurosci.* **7**, 274–281 (2012).

49. L. M. Hackel, B. B. Doll, D. M. Amodio, Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nat. Neurosci.* **18**, 1233–1240 (2015).

50. L. T. Harris, S. T. Fiske, Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Soc. Neurosci.* **5**, 76–91 (2010).

51. A. R. Krosch, D. M. Amodio, Scarcity disrupts the neural encoding of Black faces: A aocioperceptual pathway to discrimination. *J. Pers. Soc. Psychol.* **117**, 859–875 (2019).

52. R. O. Deaner, A. V. Khera, M. L. Platt, Monkeys pay per view: Adaptive valuation of social images by rhesus macaques. *Curr. Biol.* **15**, 543–548 (2005).

53. D. I. Tamir, J. P. Mitchell, Disclosing information about the self is intrinsically rewarding. *Proc. Natl. Acad. Sci.* **109**, 8038–8043 (2012).

54. B. Knutson, A. Westdorp, E. Kaiser, D. Hommer, FMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* **12**, 20–27 (2000).

55. J. P. Bhanji, M. R. Delgado, The social brain and reward: Social information processing in the human striatum. *Wiley Interdiscip. Rev. Cogn. Sci.* **5**, 61–73 (2014).

56. J. P. O'Doherty, The problem with value. *Neurosci. Biobehav. Rev.* **43**, 259–268 (2014).

57. A. Moscatelli, M. Mezzetti, F. Lacquaniti, Modeling psychophysical data at the population-level: The generalized linear mixed model. *J. Vis.* **12**, 26–26 (2012).

58. B. Y. Hayden, P. C. Parikh, R. O. Deaner, M. L. Platt, Economic principles motivating social attention in humans. *Proc. R. Soc. B* **274**, 1751–1756 (2007).

59. L. C. Johnston, C. N. Macrae, Changing social stereotypes: The case of the information seeker. *Eur. J. Soc. Psychol.* **24**, 581–592 (1994).

60. J. A. Frimer, L. J. Skitka, M. Motyl, Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *J. Exp. Soc. Psychol.* **72**, 1–12 (2017).

61. J. Falben, *et al.*, Predictably confirmatory: The influence of stereotypes during decisional processing. *Q. J. Exp. Psychol.* **72**, 2437–2451 (2019).

62. C. G. Lord, L. Ross, M. R. Lepper, Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **37**, 2098–2109 (1979).

63. D. Oyserman, V. X. Yan, "Making meaning: A culture-as-situated-cognition approach to the consequences of cultural fluency and disfluency" in *Handbook of Cultural Psychology*, Second, D. Cohen, S. Kitayama, Eds. (Guilford Press, 2019), pp. 536–565.

64. G. W. Allport, *The nature of prejudice* (Addison–Wesley, 1954).

65. U. Peters, What is the function of confirmation bias? *Erkenntnis* (2020) https:/doi.org/10.1007/s10670-020-00252-1.

66. D. Yon, F. P. De Lange, C. Press, The predictive brain as a stubborn scientist. *Trends Cogn. Sci.* **23**, 6–8 (2019).

67. A. Uusberg, G. Suri, C. S. Dweck, J. J. Gross, “Motivation: A valuation systems perspective” in *Emotion in the Mind and Body. Nebraska Symposium on Motivation. Vol. 66*, M. Neta, I. J. Haas, Eds. (Springer, Cham, 2019), pp. 161–192.

68. H. Tibboel, J. De Houwer, B. Van Bockstaele, Implicit measures of “wanting” and “liking” in humans. *Neurosci. Biobehav. Rev.* **57**, 350–364 (2015).

69. A. G. Greenwald, C. K. Lai, Implicit social cognition. *Annu. Rev. Psychol.* **71** (2020).

70. T. Morgenroth, M. K. Ryan, The Effects of Gender Trouble: An Integrative Theoretical Framework of the Perpetuation and Disruption of the Gender/Sex Binary. *Perspect. Psychol. Sci.* (2020) https:/doi.org/10.1177/1745691620902442.

71. D. M. Amodio, P. G. Devine, Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *J. Pers. Soc. Psychol.* **91**, 652–661 (2006).

72. A. Westbrook, B. Lamichhane, T. S. Braver, The subjective value of cognitive effort is encoded by a domain-general valuation network. *J. Neurosci.* **39**, 3934–3947 (2019).

73. K. S. Wang, M. R. Delgado, Corticostriatal circuits encode the subjective value of perceived control. *Cereb. Cortex* **29**, 5047–5054 (2019).

74. S. J. Gershman, How to never be wrong. *Psychon. Bull. Rev.* **26**, 13–28 (2019).

75. R. Golman, D. Hagmann, G. Loewenstein, Information avoidance. *J. Econ. Lit.* **55**, 96–135 (2017).

76. A. M. Dale, Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* **8**, 109–114 (1999).

77. R. Cusack, N. Papadakis, New robust 3-D phase unwrapping algorithms: Application to magnetic field mapping and undistorting echoplanar images. *Neuroimage* **16**, 754–764 (2002).

78. J. Grinband, T. D. Wager, M. Lindquist, V. P. Ferrera, J. Hirsch, Detection of time-varying signals in event-related fMRI designs. *Neuroimage* **43**, 509–520 (2008).

79. S. D. Slotnick, Cluster success: fMRI inferences for spatial extent have acceptable false-positive rates. *Cogn. Neurosci.* **8**, 150–155 (2017).

80. A. Moscatelli, P. Balestrucci, Psychophysics with R: The R package MixedPsy (2017).

Figure Legends

Figure 1. Neural responses associated with the confirmation of expectations about other individuals. (a) Whole-brain random-effects contrasts comparing *confirming* > *violating* trials revealed activity in the nucleus accumbens (NAcc) in (A) Study 1 and (B) Study 3. (C) We independently defined a region of interest in the NAcc using a comprehensive meta-analysis (MNI coordinates: -6, 10, -6; 10, 12, -6). Analysis of parameter estimates in this region confirmed that the bilateral NAcc showed a stronger response during confirming than during violating trials in Study 1 and (D) Study 3. Here and in subsequent figures, error bars depict SE calculated for within-subject designs.

Figure 2. The effect of behavioral ratings of trials on neural responses. In each trial, participants indicated whether they thought the presented target was likely or unlikely to be associated with the presented statement. (A) The independently defined regions of interest in the NAcc. (B) We observed a significant interaction in the independently defined NAcc in Study 1: Participants' ratings modulated the neural response only for stereotype-neutral and stereotype-violating targets, suggesting that stereotype-confirming targets are involuntarily rewarding.

Figure 3. The monetary values of confirmation of stereotype and person-specific expectations. (A) Visualization of the cumulative distribution function we used to calculate the Point of Subjective Equivalence (PSE), illustrated by group data from Study 2a. The x-axis represents the difference between the monetary values associated with the two target types presented in each trial. Each dot indicates the proportion of trials in which participants chose to rate a stereotype-*confirming* over a stereotype-*violating* target. The PSE was calculated as the point at which a cumulative normal distribution function, fit to these responses, passes 50%. This point represents the relative monetary value associated with one target type over the other. Negative values indicate that participants preferred to incur a relative monetary loss to rate a *confirming* target. Error bars depict 95% confidence intervals. (B) Distribution of individual PSE values for Study 2. Rating *stereotype-confirming* targets in Studies 2a and 2b was associated with significantly higher subjective value than rating *stereotype-violating* targets. Each gray dot depicts PSE for a specific participant. Red dots indicate the sample mean. Error bars depict 95% confidence intervals. (C) Distribution of individual PSE values for Study 4. Rating *Obama-confirming* trials was associated with significantly higher subjective value than rating *Obama-violating* trials. However, rating *Trump-confirming* trials did not significantly differ from rating *Trump-violating* trials.

Supplementary Information for

Confirmation of interpersonal expectations is intrinsically rewarding

Niv Reggev, Anoushka Chowdhary and Jason P. Mitchell, Harvard University.

Corresponding author: Niv Reggev

Email: reggevn@bgu.ac.il

This PDF file includes:

Supplementary text

Figures S1 to S4

Tables S1 to S6

SI References

SI Materials and Methods

Below we provide a detailed description of all the materials and methods used including all measures and data exclusions.

Participants

Studies 1 and 3. Twenty-eight individuals participated in Study 1 (mean age: 21.85, S.D.: 2.95, range: 18-30, 21 females, 15 Caucasian, 4 Hispanic, 4 mixed, 3 Asian, 2 African American) and 30 individuals participated in Study 3 (mean age: 22.37, S.D.: 2.62, range: 18-30, 17 females, 14 Caucasian, 7 Asian, 4 African American, 3 mixed, 1 Hispanic, 1 did not self-identify). Participants were recruited from Harvard University and its surroundings using Harvard's Psychology Study Pool website. Additional participants were excluded due to technical issues (1 participant from Study 3), lack of response to more than 20% of trials (2 participants from each study) or excessive movement (more than 1mm – half-width of the acquired voxel size) in more than one functional run in the scanner (1 participant in Study 1, 3 participants in Study 3). All participants were healthy, right-handed, native English speakers with normal or corrected-to-normal vision and no history of neurological or psychiatric conditions. Participants were compensated with $65 in Study 1 and $55 in Study 3. We determined supplemental compensation (up to $30) based on participants' performance in the monetary incentive delay task (see below). Pilot versions of the imaging tasks were conducted outside the scanner to ensure the functionality of the task.

Studies 2 and 4. Three-hundred-ninety-six individuals from Amazon Mechanical Turk completed Study 2 (in Study 2a, 198 out of 216 individuals who started the study; in Study 2b, 198 out of 223 individuals). Study 2a was conducted in March 2018, and Study 2b was conducted in April 2018. Four-hundred-seventy individuals from Prolific Academic completed Studies 4a and 4b in January 2019 (235 in each study; a total of 503 participants started the study). All participants were 18 years old or older, had an approval rate of 95% or higher, held a US nationality and participated only in one of the studies. In addition, participants from Amazon Mechanical Turk were required to have completed at least 100 tasks to be eligible for Study 2. The final sample size we report in the main text incorporates the following pre-registered exclusion criteria for participants who (1) answered incorrectly at least two of the manipulation or attention checks, (2) completed the survey in more than 20 minutes or less than 2.5 standard deviations below the sample mean duration, (3) failed to answer more than 15% of the survey, (4) provided similar ratings (on the 0-100 scale) on all trials; specifically, participants whose standard variation of the rating was 2 standard deviations below the sample mean standard deviation (not applied in Study 2a), (5) provided an identical response in all 2AFC trials (applied only in Study 4) and (6) indicated in their debriefing that they had understood the goal of the task and had acted per this understanding at any point during the task. Participants in all studies were similarly distributed on gender (Study 2a: 51.72% males, 45.98% females; Study 2b: 50.3% males, 48.52% females; Study 4a: 51.32% males, 46.56% females; Study 4b: 41.86% males, 56.98% females; across studies 1%-3% of participants self-identified as gender-nonconforming or other identification), age (Study 2a: Mean [S.D.]: 37.47 [11.75]; Study 2b: 35.13 [10.04]; Study 4a: 31.22 [10.32]; Study 4b: 31.99 [10.97]) and ethnicity (Study 2a: 78.16% White/European American; Study 2b: 74.56% White/European American; Study 4a: 69.84% White/European American; Study 4b: 69.19% White/European American). Study 4 measured political orientation on a 1-9 scale (see below). We observed no differences on this measure between Study 4a and Study 4b (Study 4a: 3.79 [2.13], Study 4b: 3.45 [1.96]).

Stimuli

All studies included face-statement pairs. We selected faces from the 10k US Adult Faces Database (a large-scale database of natural face photographs of the U.S. adult population (1)). The current investigation focused on Caucasian faces to avoid intrusion of racial or intersectional stereotypes. We further restricted the faces to be moderately memorable (0.4 – 0.6 hit rate in the 10k database, computed over an average of 81.7 people per photo) and excluded faces with any kind of distinguishable accessories (e.g. hats, big necklaces, etc.). This resulted in a total of 204 faces, 102 per gender. Three additional photos per gender were used in the practice sessions. All face pictures had neutral to mildly positive expressions. We resized all pictures to 240 by 256 pixels and presented them in color on a gray background (see Fig S1). For Studies 3 and 4 we also included images portraying the faces of Barack Obama and Donald Trump (a single image per leader). These images were taken from open access sources and cropped to be identical in dimensions to the rest of the face stimuli.

We generated gender-stereotype-relevant statements in the form of individual preference, trait, behavior or profession (e.g., “Loves taking risks”, “Doesn’t cry”, “Swears a lot”, “Is a truck driver”, etc.; see https://osf.io/tgja3/?view_only=d849bad1b606474f80c1a3e40e740875). We piloted the verbal statements in a two-phase procedure using non-overlapping groups of participants that did not participate in the reported studies. In the first session, 23 participants rated 160 gender-related statements generated by the authors based on various sources (e.g., 2). Participants saw each statement once and indicated how typical the characteristic was for a specific gender on a visual scale of 0 (= ”very untypical”) to 100 (= ”very typical”; the scale had no other tick marks). Each participant rated each statement either for men or for women, never for both, randomly determined per participant. Participants were instructed to base their ratings on how they thought the average person would respond. Based on the ratings from this first phase, we selected 136 gender-related statements (68 per gender) for which the average rating was at least 65 for one of the genders and less than 35 for the other gender. In a second session of the pilot testing, we verified the stereotypicality of these statements with a total of 78 participants who used the same rating procedure to rate the selected 136 statements and an additional 68 gender-neutral statements (e.g., “Exercises regularly”, “Thinks positively”, “Is a TV reporter”). The statements included in the studies contained 2-9 words (mean: 4.69, S.D: 1.44; no difference between experimental conditions, $p > 0.2$) and 9-45 characters (mean: 26.68, S.D.: 7.69; $p > 0.18$).

We utilized a similar procedure for leader-relevant statements. First, we generated 160 statements based on common knowledge about the selected leaders. We constructed the statements such that a statement confirming knowledge about one leader would violate knowledge about the other leader. We piloted these statements with 243 Amazon Mechanical Turk participants. Each participant saw a randomly selected subset of 40 or 60 statements and rated them on two scales: “how likely is the statement to be attributed to Barack Obama” and “how likely is the statement to be attributed to Donald Trump” on a visual scale of 0 (= ”very unlikely”) to 100 (= ”very likely”). Participants were instructed to base their ratings on how they thought the average person would respond. Each statement had an average of 81.75 raters (S.D.: 6.23; range: 67-100). We selected the 120 statements that differed the most between ratings for the two leaders. Overall, the statements contained 2-9 words (mean: 4.65, S.D: 1.52; no difference between experimental conditions, $p > 0.5$) and 11-45 characters (mean: 28.33, S.D.: 8.43; $p > 0.5$).

Behavioral procedure

Studies 1 and 3. The impression formation task included 204 face-statement pairs in Study 1 and 240 face-statement pairs in Study 3. To create stereotype-confirming and stereotype-violating pairs, we yoked half of the 136 stereotypical statements (120 statements in Study 3) with faces from the gender corresponding to the stereotype and the other half with faces from the mismatching gender (yoking of statements was randomized across participants; e.g., the statement "Doesn't cry" could be paired with a male face for one participant and with a female face for another participant to create a stereotype-confirming or a stereotype–violating pair, respectively). Gender-neutral statements (presented only in Study 1; 68 statements) were randomly and evenly yoked with male and female faces. We randomized the specific face identity paired with each statement across participants. We used a similar procedure to create leader-specific pairs (60 statements per leader), with the exception that a single photo was used per leader.

Impression Formation Task. Each trial in the Impression Formation task (see Fig S1) started with a statement (describing a neutral, stereotypically male or stereotypically female characteristic in Study 1; stereotypically male, stereotypically female, Obama-specific or Trump-specific in Study 3) presented for 1.5 seconds. Then, a face joined the statement to form a face-statement pair that was either consistent or inconsistent (or, in Study 1, neutral) with gender stereotypes or (in Study 3) person-specific knowledge. The pair appeared on screen for 4 seconds for a total of 5.5 seconds per trial in Study 1, and 3.5 seconds for a total of 5 seconds in Study 3. Each trial ended with a 0.5-second fixation crosshair. For each pair, participants used their left hand to indicate how likely the presented target was to be described by the specific characteristic using a 4-point scale (1- ‘very unlikely’; 4 -‘very likely’). Participants were to provide a response while the pair appeared on screen.

Each of four functional runs included 51 (in Study 1) or 60 (in Study 3) unique face-statement pairs (17 per condition in Study 1, 15 per condition in Study 3). In Study 3 we presented statements in mini-blocks by knowledge domain (stereotypes versus person-specific), 2 blocks per type per run; each mini-block contained 15 trials. Order of the mini-blocks was randomly determined within each run with a limitation that consecutive blocks never displayed the same type of content. To optimize estimation of the event-related fMRI response, we intermixed conditions in a pseudorandom order and separated trials by a variable interstimulus interval (Study1: 0–9 seconds, mean: 3.03, S.D.: 2.25; Study 3: 0-7 seconds, mean: 1.24, S.D: 1.84). We used OptSeq2 (3) to generate sequences optimized for the efficiency of a 3-conditions design for a first-order counterbalanced event sequence in Study 1 and for the efficiency of the (expectancy-confirming versus expectancy-negating) contrast for a first-order counterbalanced event sequence in Study 3. Of these sequences, we selected 6 sequences that contained no more than 5 consecutive events of the same condition (separate sequences were generated and selected per Study). We randomly assigned (with replacement) an event sequence for each functional run to avoid spurious results attributable to differences between conditions in one specific event sequence (4). Within conditions, trials were presented in random order. To facilitate familiarization with the task, participants completed a brief practice session before entering the MRI machine. This practice session included six statement-face pairs in Study 1 and 12 pairs in Study 3; the stimuli used in the practice were not used in any other phase of the experiment.

Monetary Incentive Delay (MID) Task. After completing the impression formation task, participants in Studies 1 and 3 completed the Monetary-Incentive Delay (MID) task (5) to allow us to localize brain regions associated with the processing of rewarding stimuli in a non-social context. Participants were not informed about this task before its execution to prevent them from forming an association between the main task and reward processing.

The MID task included a series of trials in which participants attempted to respond, via a button press, to a briefly presented target (a white rectangle) (see Fig S1B). Each trial started with a cue (a blue circle or a green circle) that was presented for 0.5 seconds. The green circle predicted a modest monetary reward ($1) upon a successful response to the target, whereas the blue circle predicted no reward. Nevertheless, participants were instructed to respond to both cue types. Cues were followed by a delay interval randomly varying in duration between 2 and 2.5 seconds. The target was then briefly presented for a duration varied between 130 and 350 milliseconds. Duration varied as a function of the participants' performance. Specifically, we implemented a 2-down 1-up staircase procedure to create a level of difficulty that would allow participants to successfully respond to the target on two-thirds of the trials. This algorithm succeeded; on average, participants were rewarded on approximately 20 of the 30 trials (Study 1 mean: 20.26; Study 3: 20.21). At the end of each trial, participants saw the amount of money they had earned on that trial along with the total amount they had earned during the task up to that point (presented for 0.5 seconds). The task included 45 trials (with 30 green cue trials). We added participants' gains in this task to their overall compensation.

Memory Test. Once outside the scanner, participants completed a surprise associative memory task that will be reported elsewhere (Reggev & Mitchell, in preparation). Briefly, none of the results reported in the current manuscript were affected by including memory in the analyses.

Additional measures. Next, participants completed several individual differences and explicit attitudes scales measuring beliefs about sexism (Ambivalent Sexism Inventory – ASI) (6), social dominance orientation (SDO) (7), motivation to control sexism (MCS) (8) and need for cognitive closure (NFC) (9). Order of the scales was randomized between participants. We included these scales to facilitate future individual differences analyses. Individual differences scores were not used in the current manuscript in any of the analyses due to insufficient power and are reported solely for full disclosure's sake.

Finally, participants in Study 1 indicated the extent to which they thought the different statements presented during the impression formation task were associated with women and men using the procedure used for piloting the statements. We presented each statement with the gender to which it was yoked in the impression formation task. Participants had up to 10 seconds per trial and were told that they should base their judgments on their own beliefs about men and women, rather than on what the "average" person in the population thinks.

After completing these tasks, participants provided demographic details (age, self-identified gender and self-identified race in an open response format, and in Study 3 political affiliation on a 1-9 scale, 1 = ”extremely liberal” and 9 = ”extremely conservative”). Then, we probed participants for their understanding of the goal of the study and asked whether they had suspected a memory test. Lastly, we paid and fully debriefed them.

Imaging procedure

Images were collected with a 3T Siemens Prisma scanner system (Siemens Medical Systems, Erlangen, Germany) using a 64-channel radiofrequency head coil. Stimuli were projected onto a screen at the end of the magnet bore that participants viewed via a mirror mounted on the head coil. Stimulus presentation was controlled by PsychoPy v1.84.2(10) running under Windows 7. Prior to entering the scanner participants were extensively briefed by one of the authors about potential movements that can occur in the scanner and ways to mitigate them. Participants were then set up in the scanner, head first and supine in the scanner bore, with a response box in their left hand. Foam cushions were placed within the head coil to minimize head movements. First, high-resolution anatomical images were acquired using a T1-weighted 3D MPRAGE sequence (TR = 2200 msec, TI = 1100 msec, acquisition matrix = 256 × 256 × 176, flip angle = 7, voxel size = 1 × 1 × 1 mm^3). Second, a fieldmap was acquired in the same plane as the functional images (see below) to correct for inhomogeneities in the magnetic field (11). Next, whole-brain functional images were collected using a simultaneous multi-slice (multiband) T2*-weighted gradient echo sequence, sensitive to BOLD contrast, developed at the Center for Magnetic Resonance Research (CMRR) at University of Minnesota (12–14) (TR = 2000 msec, TE = 30 msec, voxel size = 2 × 2 × 2 mm^3, 75 slices auto-aligned to -25 degrees of the AC-PC line, image matrix = 104 × 104, FOV = 208 * 208 mm^2, flip angle = 75°, GRAPPA acceleration factor = 2, multiband factor = 3, phase encoding direction = A -> P). After a brief practice run (identical in content to the practice session completed before entering the scanner), participants completed four impression formation task runs consisting of 229 volumes each in Study 1 and 245 volumes each in Study 3; all runs were complemented by two additional dummy scans and an initial period of approximately 26 s dedicated to references for the GRAPPA procedure. The first four volumes from each run (i.e., in addition to dummy scans) were discarded to ensure T1 equilibrium. The last 5 volumes from each run always included a crosshair fixation to ensure the appropriate estimation of the hemodynamic function for the last events in the run. Finally, participants completed the MID task in a single run consisting of 110 volumes using identical parameters to those mentioned above.

Imaging analysis

We processed and analyzed the fMRI data using SPM12 version 6225 (Wellcome Department of Cognitive Neurology, London, UK) on a 2015b MATLAB platform (Mathworks, Natick, MA, USA). Functional data were corrected for differences in acquisition time between slices, corrected for inhomogeneities in the magnetic field using the fieldmap (11), realigned to the first image to correct for head movement using a 2nd degree B-spline interpolation, unwarped to account for residual movement-related variance using a 4th degree B-spline interpolation and co-registered with each participant's anatomical data. Then, the functional data were transformed into standard anatomical space (2 mm isotropic voxels) based on the ICBM152 brain template (Montreal Neurological Institute). Normalized data were spatially smoothed (6 mm full-width at half-maximum, FWHM) using a Gaussian Kernel. In addition to the GLM models reported in the main text, in Study 1 we also examined a model in which the stereotypicality of trials was modeled continuously rather than with the dichotomous binning approach. Specifically, this additional model included two regressors - one for trials including a woman's face and another for trials including a man's face. We included a separate parametric modulator for each regressor to model the extent of the stereotypicality of the statement included in that trial based on our pilot ratings. For example, the statement "Can lift heavy things" was rated as related more to men than to women in our pilot studies with a 28-points difference on the 0-100 scale. Consequently, the parametric modulation value was 0.28 for trials in which this statement was presented with a man's face, and -0.28 for trials in which this statement was presented with a woman's face. Similar to the models reported in the main text, in this additional model we convolved events with a canonical hemodynamic response function and its temporal derivative and included additional covariates of no interest (session mean, no response trials, six motion parameters, and their temporal derivative).

Regions of interest (ROIs)

We used two complementary approaches to localize regions involved in the processing of rewards. For independently defined ROIs, we defined 8-mm spheres around peak coordinates drawn from a recent meta-analysis (15). Specifically, we utilized the peaks of the region identified as supporting the processing of both monetary and primary incentives: bilateral ventral striatum (x=-6, y=10, z=-6 and x=10, y=12, z=-6).

To functionally locate these regions, we examined the MID task to identify voxels that responded more to rewarded trials (i.e., trials in which participants successfully responded to the target) than to no-reward trials (i.e., trials in which no reward was available). Whole-brain corrected clusters (using the procedure described in the main text) were defined as independent ROIs.

As we had no prediction about laterality of the hypothesized effects, we collapsed across hemispheres to create a single ROI. We extracted and averaged parameter estimates across voxels and analyzed them with planned contrasts in a repeated-measures ANOVA context using p<0.05 as a threshold.

Studies 2 and 4. In each trial, participants chose one of two decks of cards (see Fig S2). We instructed participants to choose based on their preferences regarding the information they had available for each trial. Participants had two sources of information to rely on. First, each deck was associated with a small monetary payoff (ranging from $0.03 to $0.09 in 2 cents increments). Participants were told that a subset of their choices (5 trials in Study 2, 7 trials in Study 4) would be added to their final compensation for the study. Payoff amounts for each choice varied across trials (and were occasionally equal). Second, each deck was associated with a specific label (“Typical” versus “Atypical” in Study 2, “Common” versus “Uncommon” in Study 4) that determined the content presented when that deck is selected (stereotypical or counter-stereotypical targets in Study 2, knowledge-confirming or knowledge-violating targets in Study 4). After making their selection, participants saw a face-statement pair corresponding to their choice. For example, if a participant in Study 2 selected the “Typical” deck, they would be presented with a stereotypical target (e.g., in Study 2a a male associated with the statement “CEO of a big company”). Participants then indicated how likely that target was to possess that characteristic on a visual scale of 0 (= ”Not at all likely”) to 100 (= ”Very likely”; the scale had no other tick marks). The location of the specific labels (right deck versus left deck) was counterbalanced between participants. Study 2a included only faces of men, Study 2b included only faces of women, Study 4a included only the face of Barack Obama and Study 4b included only the face of Donald Trump.

All studies included a demo trial, 4 practice trials, and 25 deck selection trials (32 trials in Study 4). We also included several catch trials to detect if participants were responding without considering the statement presented. The catch trials prompted the participants to slide the bar to the right tick mark or to the left tick mark. Participants had up to 4s to select a card and up to 5s to rate an individual. Participants that did not respond fast enough to more than 20% of the trials got their survey terminated midway through the task. The specific amounts associated with each trial and the face-statement pairs presented per trial were randomized. Following these trials, participants were presented with 4 final manipulation-check trials in which they were asked to select the card with the higher value.

After completing the main task, participants responded to the individual difference scales mentioned above – ASI, SDO, MCP, and NFC. Order of the scales was randomized between participants. No significant correlations between Point of Subjective Equivalence (PSE) and the individual scores on these scales were consistently detected across studies. Finally, participants supplied demographic information (including age, self-reported gender identity, self-reported race with multiple answers enabled and whether they were born in the US). Study 4 also probed participants' political affiliation (as in Study 3), how much they liked Barack Obama and how much they liked Donald Trump on two separate 0-100 scales. Then, we probed for participants' intuitions about the goal of the task and fully debriefed them.

SI Results

Behavioral analyses

Table S1 presents the summaries of participants' ratings and reaction times in the impression formation task in Study 1. We analyzed rating data with mixed models as implemented in the lme4 package version 1.1-14 (16) and the 'ordinal' package version 2018.4-19 (17) for R version 3.4.2 (R Core Team, 2017). As the behavioral data obtained in the impression formation task were ordinal, we analyzed them using cumulative link mixed models (CLMM) with the logit link. To avoid the transformation of raw reaction time data, we used generalized linear models (gLMMs) with the inverse Gaussian identity link (18). We included random effects for the intercepts for participants and statements, as well as by-participant random slopes for the fixed effect of stereotypicality. Trials that elicited no response (<1.5% of all trials; no difference between conditions) were excluded from all analyses.

To examine whether our stereotype-confirming and stereotype-violating targets were indeed perceived differently by our participants in Study 1, we tested the effect of our a-priori categorization on behavioral ratings in a CLMM. Stereotypicality was dummy coded with stereotype- neutral trials as the intercept and behavioral ratings were centered on 0. Overall stereotypicality affected ratings, as indicated by leave-one-out model comparison, comparing our model to an intercept only model ($\chi^2_{(2)} = 44.45$, $p < 0.001$). The manipulation worked as anticipated: Stereotype-confirming targets received higher ratings than stereotype-neutral targets ($\beta \pm SE = 0.42 \pm 0.14$, $Z = 2.99$, $p = 0.003$), whereas stereotype-violating targets received lower ratings ($\beta \pm SE = -1.22 \pm 0.15$, $Z = -8.31$, $p < 0.001$).

Stereotypicality did not have a main effect on reaction time for the behavioral ratings ($\chi^2_{(2)} = 0.35$, $p > 0.8$). However, reaction time did vary by the interaction of stereotypicality and behavioral ratings, as indicated by comparing a model with an interaction term to a model without it, $\chi^2_{(2)} = 133.88$, $p < 0.001$); participants were slower to provide responses that were not in line with our a-priori definitions (e.g., indicating that a woman is very likely to be a firefighter or saying that a man is very unlikely to be a CEO); see Table S1 for full descriptive results.

Behavioral results in Study 3 generally replicated the behavioral results we obtained in Study 1 (see Table S4). The outcome of the expectation (whether the target confirmed or violated the expectation) significantly affected participants' ratings ($\chi^2_{(1)} = 73.61$, $p < 0.001$). Domain (stereotype-based or person-specific-based knowledge) also affected participants' ratings ($\chi^2_{(1)} = 27.58$, $p < 0.001$). Interestingly, participants distributed their expectation-based responses differently between the domains (outcome by domain interaction: $\chi^2_{(1)} = 862.52$, $p < 0.001$). Participants used more extreme outcome-congruent ratings for person-specific expectations compared to stereotype-based expectations (see Table S4).

Similar to Study 1, participants' reaction time did not differ between expectation-confirming or expectation-violating trials (t = -0.29, p = 0.77), nor between person-specific or stereotype-based expectations (t = 1.89, p = 0.059). Comparable to Study 1, we observed a significant interaction between outcome and behavioral ratings (t = -15.11, $p < 0.001$), an interaction that was qualified by a 3-way interaction with domain (t = 6.36, $p < 0.001$). To interpret this interaction, we analyzed responses separately for the two knowledge domains. In both knowledge domains, participants responded faster when their responses were in line with our a-priori definitions (t = -9.72 and t = -13.81 for stereotype- and person-specific-based expectations, respectively; p's < 0.001).

Neuroimaging

In the main text we report, in addition to the main findings (see Fig 1 and Table S2) that activity in the NAcc in Study 1 increased when participants indicated that a target was likely to have a counter-stereotypical characteristic, compared to when they rated a target as unlikely to have such a trait (see Fig 2). To examine the mechanisms driving this response, we conducted an exploratory whole-brain analysis comparing stereotypical and counter-stereotypical targets that received a 'Likely' rating. This analysis revealed significant differences in multiple brain regions, including the bilateral insula, dorsal anterior cingulate cortex, and right temporal-parietal junction (see Fig S3 and Table S3 for full details), possibly indicating that the attribution of counter-stereotypical traits to targets triggers unique processes. These results suggest that expectancy-violating targets can be associated with increased reward-related neural activity – if we can engage multiple additional neural processes to overcome the decreased activity typically associated with them. Conversely, when a target confirmed participants' stereotypical expectations, the NAcc increased its activity regardless of how participants behaviorally responded to it, stressing the potency of these expectations.

In Study 3, in addition to the main findings (see Fig 1 and Table S5), the design allowed us to compare the effects of expectation confirmation to different specific targets, namely, men, women, Trump or Obama. Analyzing activity in the NAcc in a 2 (expectation result) X 2 (content domain) X 2 (specific target) ANOVA yielded, in addition to the main effect of expectation confirmation ($F_{(1,29)} = 21.41$, $p < 0.0001$, $\eta^2_p = 0.42$ [0.19-0.58], a main effect of content domain ($F_{(1,29)} = 4.56$, $p = 0.04$, $\eta^2_p = 0.14$ [0.003-0.32]; all other effects p>0.1) such that gender-related trials were associated with increased activity. To complement the findings, we conducted a whole-brain interaction analysis to examine whether any neural regions responded differently to the confirmation or violation of expectations between specific targets. Given the different experimental context between the two content domains, we first examined the interaction of expectations and specific targets within each content domain separately. This analysis yielded 3 regions (p<0.05, FWE-corrected; Table S6 and Figure S4), including the left inferior frontal gyrus (LIFG) and left superior temporal gyrus (LSTG), in which confirmation of expectations yielded more activity than their violation only when the statements pertained to Trump (LIFG results: $F_{(1,114.24)} = 40.12$, $p < 0.0001$, $\eta^2_p = 0.26$ [0.15-0.36]) and not to any of the other targets (triple interaction: $F_{(1,29)} = 33.32$, $p < 0.0001$, $\eta^2_p = 0.53$ [0.30-0.66; similar patterns were observed in LSTG]. Disambiguating this triple interaction, we verified the interaction between expectation results and specific targets for leaders: $F_{(1,29)} = 42.73$, $p < 0.0001$, $\eta^2_p = 0.6$ [0.37-0.71]; the parallel interaction within gender content was not significant: $F_{(1,29)} = 0.55$, $p > 0.4$, $\eta^2_p = 0.02$ [0-0.03]). Notably, no main effect of expectation results was observed in these regions (all $p's > 0.05$). Together, these findings suggest that the confirmation of expectations pertaining to Donald Trump evoke an additional process which is not triggered for confirmation of other types of expectations.

SI References

1. W. A. Bainbridge, P. Isola, A. Oliva, The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.* **142**, 1323–1334 (2013).

2. D. A. Prentice, E. Carranza, What women and men should be, shouldn't be, are allowed to be, and don't have to be: the contents of prescriptive gender stereotypes. *Psychol. Women Q.* **26**, 269–281 (2002).

3. A. M. Dale, Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* **8**, 109–114 (1999).

4. J. A. Mumford, T. Davis, R. A. Poldrack, The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage* **103**, 130–138 (2014).

5. B. Knutson, A. Westdorp, E. Kaiser, D. Hommer, FMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* **12**, 20–27 (2000).

6. P. Glick, S. T. Fiske, The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *J. Pers. Soc. Psychol.* **70**, 491–512 (1996).

7. A. K. Ho, *et al.*, The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *J. Pers. Soc. Psychol.* **109**, 1003–1028 (2015).

8. S. C. Klonis, E. A. Plant, P. G. Devine, Internal and external motivation to respond without sexism. *Personal. Soc. Psychol. Bull.* **31**, 1237–1249 (2005).

9. A. W. Kruglanski, D. M. Webster, A. Klem, Motivated resistance and openness to persuasion in the presence or absence of prior information. *J. Pers. Soc. Psychol.* **65**, 861–876 (1993).

10. J. W. Peirce, PsychoPy-Psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).

11. R. Cusack, N. Papadakis, New robust 3-D phase unwrapping algorithms: Application to magnetic field mapping and undistorting echoplanar images. *Neuroimage* **16**, 754–764 (2002).

12. S. Moeller, *et al.*, Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn. Reson. Med.* **63**, 1144–1153 (2010).

13. D. A. Feinberg, *et al.*, Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. *PLoS One* **5**, e15710 (2010).

14. J. Xu, *et al.*, Evaluation of slice accelerations using multiband echo planar imaging at 3 T. *Neuroimage* **83**, 991–1001 (2013).

15. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).

16. D. M. Bates, M. Maechler, B. Bolker, S. Walker, lme4: Linear mixed-effects models using Eigen and S4 (2014).

17. R. H. B. Christensen, ordinal---Regression Models for Ordinal Data (2018).

18. S. Lo, S. Andrews, To transform or not to transform : Using Generalized Linear Mixed Models to analyse reaction time data. *Front. Psychol.* **6**, 1–16 (2015).

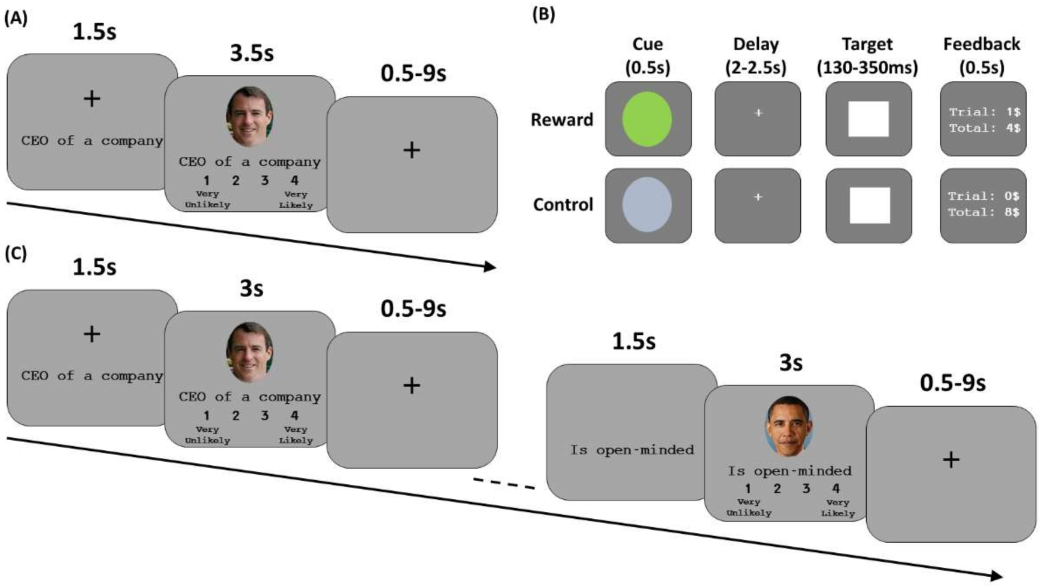


Fig S1.

Experimental designs of Studies 1 and 3. (A) In Study 1, participants saw 204 unique trials. Each started with a gender-relevant or irrelevant statement, followed by a target face confirming, violating or neutral in respect to the displayed statement. Participants indicated how likely the presented person was to have the characteristic described in the statement on a 1 (“very unlikely”) to 4 (“very likely”) scale. (B) Following the impression formation task, participants completed the monetary incentive delay (MID) task. In each trial, participants saw a cue predicting the outcome of a successful response to the target. A green cue always indicated monetary reward, a blue cue always indicated no reward. Participants saw 30 reward cues and 15 no-reward cues. After a randomly jittered delay a target appeared on screen for a brief duration (determined by a 2-up-1-down staircase procedure). Participants received feedback about their performance in each trial and across the entire task. (C) In Study 3, participants rated 240 trials, including 120 stereotype-related targets and 120 person-specific trials. Each content domain (stereotypes versus person-specific) was presented in separate blocks of 15 consecutive trials per type.

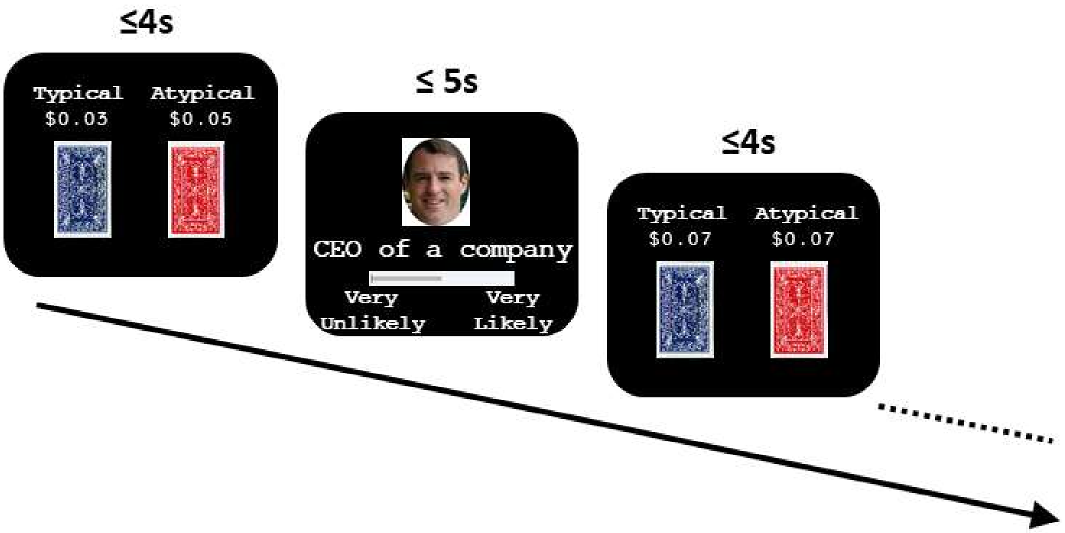


Fig S2. Experimental design of Study 2a. Participants chose which target type to rate (typical, denoting stereotype-confirming targets, versus atypical, denoting stereotype-violating targets). Each target type was associated with a variable amount of money. After deciding which target type to rate, participants then rated a target of the chosen type. The design of studies 2b, 4a and 4b followed this procedure.

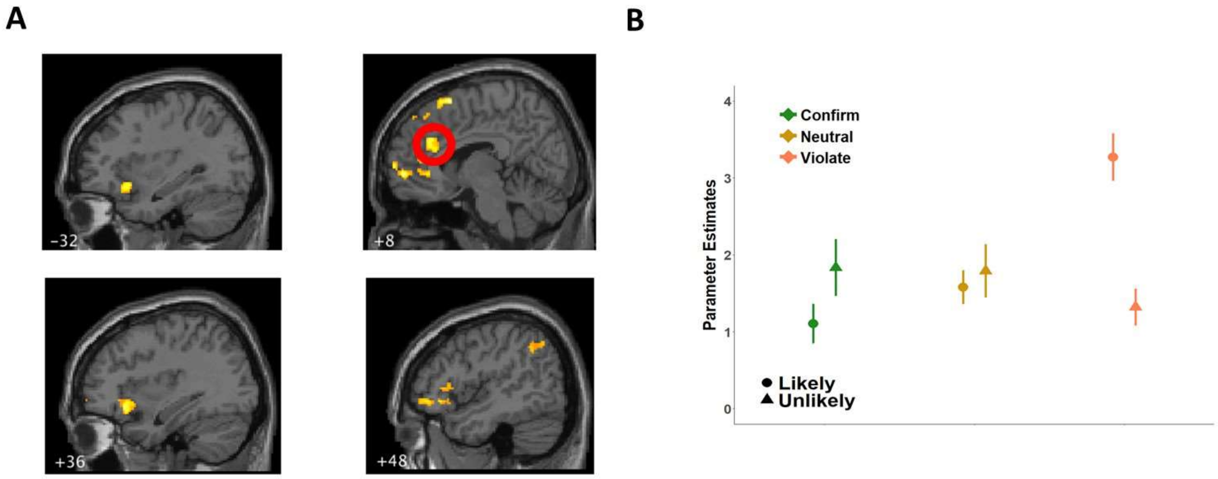


Fig S3. Neural response differences between stereotype-violating and stereotype-confirming targets. (A) A whole-brain random-effects contrast comparing *stereotype-violating* > *stereotype-confirming* targets that were associating with a *likely* response. (B) Parameter estimates extracted from the Anterior Cingulate Cortex (circled in red in A), reveal a significant interaction in which only *stereotype-violating* targets that participants indicated as likely to have the characteristic were associated with a robust response, suggesting that in order to say that a target can be associated with a counter-stereotypical attribute, one needs to engage in effortful processing. Error bars depict SE calculated for within-subject designs.

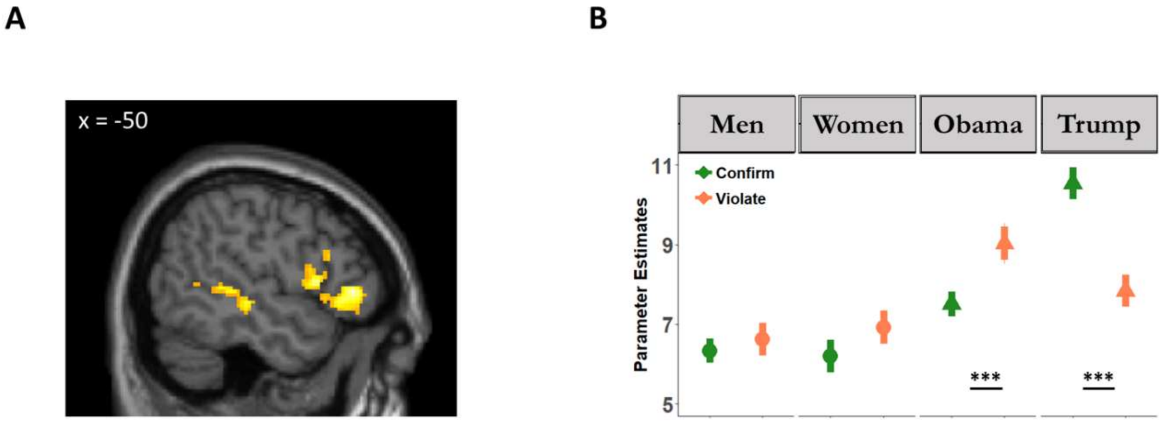
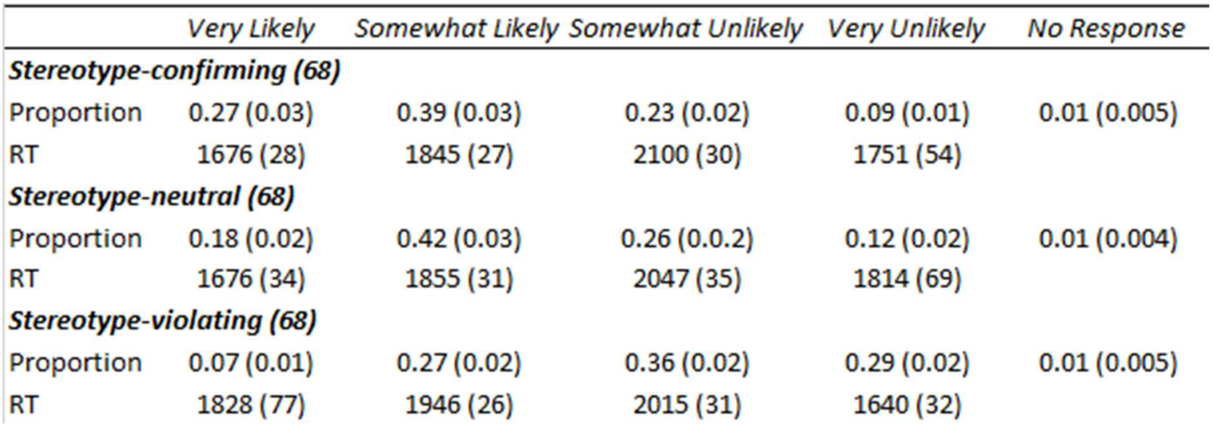


Fig S4. Neural responses as a function of expectancy-outcome and type of target. (A) results of a whole-brain analysis examining the interaction of outcome and type of target (see main text and Table S6 for details). (B) Parameter estimates drawn from the Left Inferior Frontal Gyrus ROI.

Table S1. Distribution of participants' ratings and reaction time (RT, in milliseconds) as a function of stereotypicality in the impression formation task in Study 1. Parentheses indicate standard error of the mean.



	Very Likely	*Somewhat Likely*	*Somewhat Unlikely*	*Very Unlikely*	*No Response*
Stereotype-confirming (68)					
Proportion	0.27 (0.03)	0.39 (0.03)	0.23 (0.02)	0.09 (0.01)	0.01 (0.005)
RT	1676 (28)	1845 (27)	2100 (30)	1751 (54)	
Stereotype-neutral (68)					
Proportion	0.18 (0.02)	0.42 (0.03)	0.26 (0.0.2)	0.12 (0.02)	0.01 (0.004)
RT	1676 (34)	1855 (31)	2047 (35)	1814 (69)	
Stereotype-violating (68)					
Proportion	0.07 (0.01)	0.27 (0.02)	0.36 (0.02)	0.29 (0.02)	0.01 (0.005)
RT	1828 (77)	1946 (26)	2015 (31)	1640 (32)	

Table S2. Study 1: Gray matter regions showing differences in activity between stereotypical (stereotype-confirming) and counter-stereotypical (stereotype-violating) targets in model 1 (with no specification of behavioral response; p<0.05, corrected).

Region	***MNI coordinates***			***Z value***	***# voxels***
	X	***Y***	***Z***		
(a) Stereotypical > Counter-Stereotypical					
Nucleus Accumbens	-2	4	-6	4.98	30
L Inferior Temporal	-50	-56	-12	4.03	108
(b) Counter-Stereotypical > Stereotypical					
R Motor	34	-22	58	4.38	145
R Inferior Frontal	46	32	-10	4.07	18
Dorsomedial Prefrontal	8	36	50	3.8	50
(c) Parametric Modulation (Counter-Stereotypical -> Stereotypical)					
L inferior Parietal Sulcus	-30	-82	36	4.22	17
L Inferior Temporal	-48	-58	-12	4.19	60
Extrastriate Cortex	14	-84	-12	3.72	22
Nucleus Accumbens	0	6	-6	4.15	20
R Nucleus Accumbens	-8	8	-4	3.91	13

	(d) Parametric Modulation (Stereotypical -> Counter-Stereotypical)				
R Motor	34	-20	58	4.41	128
R Medial prefrontal	12	48	12	4.33	18
R Inferior Parietal	58	-50	38	3.94	18
Dorsomedial Prefrontal	8	36	50	3.85	16

Table S3. Study 1: Gray matter regions showing differences in activity between stereotypical (stereotype-confirming) and counter-stereotypical (stereotype-violating) targets as a function of behavioral rating (likely or unlikely) in model 2 (specifying behavioral response). Sections (a) and (b) are conceptually identical in their contrasts to the corresponding sections in Table S2. (p<0.05, corrected).

Region	***MNI coordinates***			***Z value***	***# voxels***
	X	***Y***	***Z***		
	(a) Stereotypical > Counter-Stereotypical				
L Angular gyrus	-44	-72	34	4.37	90
Nuclues Accumbens	-2	4	-2	4.34	44
Extrastriate Cortex	4	-82	2	4.21	282
Medial parietal	-20	-64	58	4.19	366
R Lingual gyrus	20	-74	-4	4.14	44
Left Inferior Frontal	-46	10	26	4.01	47
L Inferior Temporal	-52	-52	-14	3.95	72

R Motor	44	-14	50	3.81	44
Nuclues Accumbens	-8	10	-2	3.78	26

(b) Counter-Stereotypical > Stereotypical

R Motor	38	-22	60	5.44	310
R Dorsolateral Prefrontal cortex	50	32	30	4.07	21
R Dorsolateral Prefrontal cortex	44	22	44	3.8	22

(c) Likely: Stereotypical > Counter-stereotypical

None

(d) Likely: Counter-stereotypical > Stereotypical

Premotor	8	16	64	4.7	475
Anterior Cingulate	14	26	26	4.6	750
R Insula	34	22	-8	4.53	443
R Anterior Prefrontal	26	54	2	4.37	71
Anterior Medial Prefrontal	8	58	-2	4.34	150
L Insula	-32	18	-12	4.21	95
R Superior Temporal	56	-26	2	4.21	134

Posterior Cingulate	-12	-48	30	4.08	40
R Tempróparietal Junction	52	-34	18	3.98	57
L Superior Temporal	-56	20	2	3.94	27
L Anterior Prefrontal	-44	48	-8	3.94	67
R Middle Frontal	28	54	26	3.92	65
R Angular gyrus	52	-62	42	3.88	92
R Insula	28	18	-18	3.83	25
Cingulate	0	-18	40	3.72	40
R Dorsolateral Prefrontal	48	22	46	3.54	29
Premotor	-4	18	58	3.54	73
Posterior Cingulate	4	-50	28	3.52	25
R Inferior frontal gyrus	64	2	6	3.49	40

(e) Unlikely: Stereotypical > Counter-stereotypical

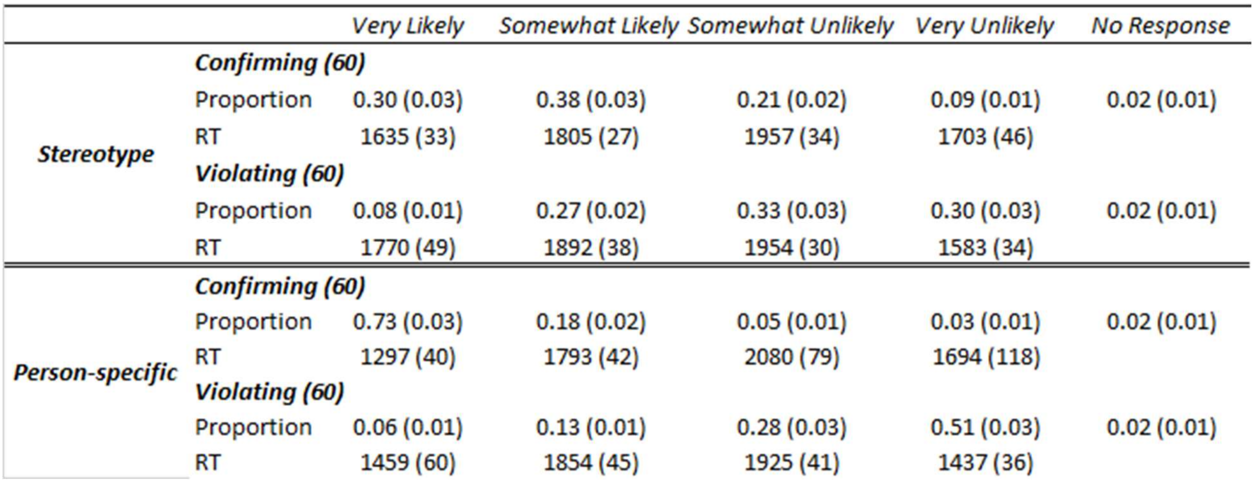
L Angular gyrus	-40	-76	44	4.51	19
L Intraparietal Sulcus	-28	-78	30	4.42	74
R Orbitofrontal	28	36	-16	4.04	23
R Middle Frontal	42	34	10	3.96	64
L Inferior Temporal	-54	-52	-16	3.81	20

R Inferior temporal	42	-52	-8	3.76	20
Medial parietal	-20	-78	50	3.52	18
Dorsomedial Prefrontal	12	46	38	3.38	20

(f) Unlikely: Counter-stereotypical > Stereotypical

None

Table S4. Distribution of participants' ratings and reaction time (RT, in milliseconds) as a function of expectation domain and expectation result (confirmed or violated) in the impression formation task in Study 3. Parentheses indicate standard error of the mean.



		Very Likely	*Somewhat Likely*	*Somewhat Unlikely*	*Very Unlikely*	*No Response*
Stereotype	***Confirming (60)***					
	Proportion	0.30 (0.03)	0.38 (0.03)	0.21 (0.02)	0.09 (0.01)	0.02 (0.01)
	RT	1635 (33)	1805 (27)	1957 (34)	1703 (46)	
	Violating (60)					
	Proportion	0.08 (0.01)	0.27 (0.02)	0.33 (0.03)	0.30 (0.03)	0.02 (0.01)
	RT	1770 (49)	1892 (38)	1954 (30)	1583 (34)	
Person-specific	***Confirming (60)***					
	Proportion	0.73 (0.03)	0.18 (0.02)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
	RT	1297 (40)	1793 (42)	2080 (79)	1694 (118)	
	Violating (60)					
	Proportion	0.06 (0.01)	0.13 (0.01)	0.28 (0.03)	0.51 (0.03)	0.02 (0.01)
	RT	1459 (60)	1854 (45)	1925 (41)	1437 (36)	

Table S1. Study 3: Gray matter regions showing differences in activity between expectation-confirming and expectation-violating targets. (a) and (b) portray the results collapsed across content domain, (c) through (f) include the results per content domain (p<0.05, corrected).

Region	***MNI coordinates***			***Z value***	***# voxels***
	X	***Y***	***Z***		
(a) Main effect: Confirming > Violating					
Extrastriate Cortex	12	-80	-4	5.63	2800
R Motor	46	-14	52	5.43	298
L Ventral Striatum	-12	20	-2	4.93	89
L Insula	-48	4	8	4.87	59
Cerebellum	10	-74	-38	4.82	56
Cerebellum	-18	-76	-44	4.58	35
L Occipitotemporal Cortex	-44	-68	0	4.57	463
R Ventral Striatum	6	6	-8	4.52	49
R Insula	36	6	10	4.51	40
L Parietal Operculum	-44	-32	20	4.34	166
Occipital Cortex	24	-96	20	4.29	266
Posterior Cingulate	-2	-34	44	4.26	240
R Fusiform Gyrus	48	-42	-16	4.25	57

R Occipitotemporal Cortex	54	-60	0	4.19	221
L Superior Temporal	-66	-36	20	4.08	62
L Insula	-36	-12	14	4.06	42
L Parahippocampus	-24	-40	-6	4.06	50
L Inferior Frontal	-42	38	12	4.01	34
Precuneus	-16	-62	16	3.98	103
Ventromedial Prefrontal	6	42	-16	3.86	28
L Motor	-50	-16	54	3.77	112
L Temporal Pole	-48	14	-14	3.72	20
R Orbitofrontal	26	38	-18	3.69	26
R Inferior Frontal	52	8	-4	3.64	39

(b) Main effect: Violating > Confirming

Dorsomedial Prefrontal	12	34	52	4.22	23
R Motor	36	-24	54	3.99	50
Dorsomedial Prefrontal	-8	42	50	3.92	62
L Middle Frontal	-38	20	46	3.88	32

(c) Gender: Confirming > Violating

Extrastriate Cortex	10	-78	-6	4.48	146
R Occipitotemporal Cortex	54	-62	2	3.80	47
Ventral Striatum	4	10	-10	3.78	21
L Motor	52	-12	54	3.66	64
R Lateral Occipital	30	-84	38	3.55	28
	(d) Gender: Violating > Confirming				
Dorsomedial Prefrontal	12	34	52	3.92	33
L Middle Frontal	-34	20	40	3.55	19
	(e) Leaders: Confirming > Violating				
Extrastriate Cortex	8	-76	-6	5.56	1854
R Motor	48	-14	52	5.42	204
Posterior Cingualte	-8	-40	54	4.34	57
L Parietal Operculum	-42	-36	26	4.31	248
R Insula	36	6	10	4.23	24
L Temporal Pole	-50	12	-12	4.19	80
Cerebellum	24	-66	-44	4.18	24
R Supramarginal gyrus	50	-36	8	4.06	50

L Inferior Frontal / Insula	-48	4	6	4.01	35
L Occipitotemporal Cortex	-44	-68	0	3.89	21
Cuneus	-2	-94	16	3.83	21
R Fusiform Gyrus	48	-44	-16	3.78	30
Cerebellum	8	-74	-38	3.76	20
L Occipitotemporal Cortex	-48	-58	-2	3.71	111
Precuneus	-14	-66	28	3.71	55
L Inferior Frontal	-60	6	10	3.68	22
Precuneus	-14	-78	42	3.62	74
Ventral Striatum	-10	20	-2	3.57	21
Anterior Cingualte	-2	14	32	3.49	31
R Inferior Frontal / Insula	50	4	-2	3.48	22
L Motor	-50	-16	54	3.45	22
(f) Leaders: Violating > Confirming					
R Motor	40	-24	60	4.17	90
Dorsomedial Prefrontal	-8	42	50	3.77	22

Table S2. Study 3: Results of a set of two whole-brain analyses examining the interactive effects of expectation-result (confirmation or violation) and specific target, done separately for gender and leaders. (p<0.05, corrected).

Region	***MNI coordinates***			***Z value***	***# voxels***
	X	***Y***	***Z***		
(a) Gender: Confirmation effect in men > Confirmation effect in women					
None					
(b) Gender: Confirmation effect in women > Confirmation effect in men					
None					
(c) Leaders: Confirmation effect in Obama > Confirmation effect in Trump					
None					
(d) Leaders: Confirmation effect in Trump > Confirmation effect in Obama					
Cuneus	12	-88	24	5.16	435
L Inferior Frontal	-50	34	2	4.83	473
L Superior Temporal	-62	-36	4	4.32	352
Extrastriate Cortex	-6	-80	6	3.67	75