nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01065-0

Biased evaluations emerge from inferring hidden causes

Yeon Soon Shin[1]✉ and Yael Niv[1,2]

How do we evaluate a group of people after a few negative experiences with some members but mostly positive experiences otherwise? How do rare experiences influence our overall impression? We show that rare events may be overweighted due to normative inference of the hidden causes that are believed to generate the observed events. We propose a Bayesian inference model that organizes environmental statistics by combining similar events and separating outlying observations. Relying on the model's inferred latent causes for group evaluation overweights rare or variable events. We tested the model's predictions in eight experiments where participants observed a sequence of social or non-social behaviours and estimated their average. As predicted, estimates were biased toward sparse events when estimating after seeing all observations, but not when tracking a summary value as observations accrued. Our results suggest that biases in evaluation may arise from inferring the hidden causes of group members' behaviours.

In impression formation, negative information tends to have a stronger impact than positive information. Thus, the overall impression formed after positive and negative information is more negative than the algebraic sum of valences of individual experiences[1]. This phenomenon, called negativity bias, has been extensively demonstrated.

One potential source of the asymmetry in impression formation is the low frequency of negative events. As they are rare, negative events are more surprising and grab more attention. Supporting this idea, participants spend more time looking at negative descriptions of a target person than they do positive descriptions, and weighting valences of descriptions by looking time predicts the resulting evaluation bias[2]. Indeed, reinforcement-learning models suggest that, when tracking a cue's value, surprising outcomes update the value more strongly, resulting in greater contributions of surprising (in this case, negative) events to the overall evaluation[3].

Diagnosticity of information is another important aspect in biases. For instance, an intelligent person can occasionally behave unintelligently, but the opposite is less likely. As intelligent behaviours are more diagnostic of intelligence than unintelligent behaviours are, these diagnostic behaviours carry more weight and dominate evaluation even when participants observe evidence of unintelligence more frequently[1,4]. In social settings, in contrast, positive behaviours are often the default[5], making negative events more diagnostic than positive ones. Therefore, negative information could weigh heavily in such situations. Work on conceptual similarities is also in line with this idea: pairwise similarity between negative words is judged to be lower than between positive words, meaning that the distribution of negative concepts is sparser than positive concepts' distribution[6,7].

If post hoc impression judgements could rely on perfect encoding and retrieval of memory, evaluations would accurately reflect experiences without distortions. However, we cannot remember every single encounter with every person we ever meet. Therefore, we might lump together different similar experiences, clustering them in memory for future use. The high variance in negative concepts could lead to less clustering of those concepts in memory, leaving each individual observation more diagnostic of the valence of its cluster. Positive concepts, on the other hand, may form a large cluster of positivity where each individual piece of information is less diagnostic of the overall 'gist' of the cluster. Supporting the idea that positive-valence concepts are clustered together more than are negative-valence concepts, processing of a positive word is faster if it follows another positive word, whereas this is not the case for negative words[6]. This could explain how the sparsity of negative events, given by rarity and variance, seems to contribute to the negativity bias[8].

'Latent causes'—hidden causal structures that are assumed to generate a set of observable events—can be a meaningful basis for summarizing experiences[9–11]. For instance, when we believe that a single underlying reason (for example, a person wanting a future favour from us) is causing ten separate events that we observe (for example, friendly encounters with the person), we can keep one summary of those ten events instead of trying to remember each one of them. However, if we believe that two separate underlying causes (for example, wanting a future favour from us and being socially anxious) generated five events each (for example, friendly encounters with the person and socially inadequate behaviours of the person), we might keep two summaries. We may also care about generalizing across people as a group, for example, for forming expectations about the norms of people in different countries we may travel to. Relying on latent causes is normative when generalizing past experiences to a new situation, as we can utilize what we learned from the past events that are caused by the latent cause we expect is active now, but not the ones caused by other causes.

This approach, although rational, can also be a source of biases because we do not know the true latent cause of each event. For example, we may infer distinctive latent causes from seemingly different events even if, in truth, the events share one cause. For instance, the sparsity of negative events may lead to inference of many distinctive latent causes, while positive events may be attributed to a small number of causes (Fig. 1a). If we then make our overall impression at the level of these latent causes, for instance by averaging over the valence of all the causes associated with the

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Department of Psychology, Princeton University, Princeton, NJ, USA. ✉e-mail: yshin@princeton.edu

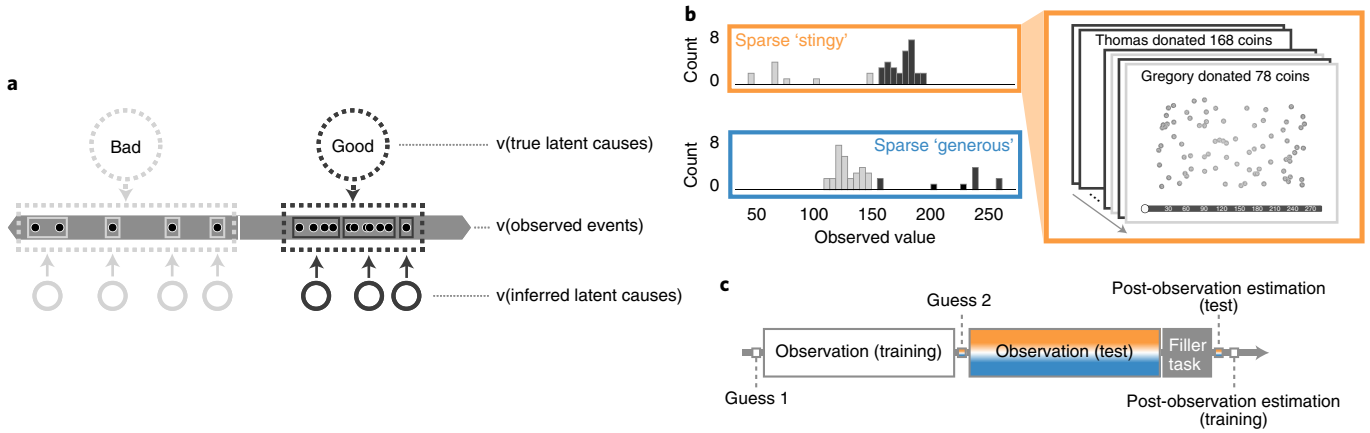

Fig. 1 | Hypothetical latent structure and experimental designs. a, Hypothetical latent structure of positive and negative events. Events (dots) are generated from two true latent causes (dashed circles). The observer infers these latent causes (solid circles) from the observed events based on the similarity of events to each other. If events generated from the 'bad' latent cause are sparse, the observer may infer many distinctive latent causes, each accounting only for a few observations. However, a small number of causes may be inferred to account for the many, similar, good events. **b**, Experimental design for the sparse 'stingy' donor and sparse 'generous' donor conditions. The donation amount was drawn from 'stingy' (below the mean, light grey) and 'generous' (above the mean, dark grey) distributions. In the sparse 'stingy' condition, most donations were drawn from a homogeneous 'generous' distribution, while few, variable, donations were drawn from a 'stingy' distribution. The distribution was flipped in the sparse 'generous' condition. In both conditions, the true mean of donation amounts was 150. The participant observed the number of coins donated on each trial, marking the amount on a slider bar to register it. **c**, Experimental procedure: after participants first guessed the general donation amount (guess 1), we showed them a symmetric training sequence (with no sparsity) to adjust their expectations. They then guessed the general donation amount again (guess 2) and observed a test sequence according to their experimental condition. After performing a filler task where they clicked a series of stimuli that appeared on random locations, participants estimated the overall donation amount for the test and training sequences (post-observation estimations).

person or group of people we are forming an impression about, the estimation will be biased toward the sparse area where there are a larger number of inferred latent causes. Here, we hypothesize that the negativity bias emerges from the combination of normative segmentation of information into causes based on similarity and incorrectly weighted averaging over latent causes.

Results

To test this hypothesis, we presented human participants with a sequence of events, described as 'donations' (experiments 1A, 1B and 2A), 'sales' (experiment 2B) or 'rewards' (experiment 2C), drawn from event distributions in which we manipulated sparsity such that either the below-average values or the above-average values were sparser (Fig. 1b). Then, in a surprise test, we asked them to estimate the average value of the observed events. We compared estimations between sparsity conditions, which all had the same true average. If biases arise from inferring underlying causes such as intentions behind behaviours, we would expect that the estimated average of a sequence consisting of sparse below-average events would be lower than the estimated average of a sparse above-average event sequence.

Experiment 1. Experiment 1A ($N=76$) examined the effect of rarity and variability of distributions in estimation biases. Participants observed two sequences of 40 'donors' making coin 'donations', where each donor made a single donation (Fig. 1c). Each sequence was presented as a group of people who attended the same college. The first sequence served as a training sequence to adjust participants' prior beliefs about donations. After reporting their general expectation for donation (guess 1) on a scale of 1 to 300 coins, participants observed a training sequence in which donation amounts were equally distributed above and below a mean donation of 150 coins. To ensure that the training sequence indeed adjusted participants' prior beliefs close to the true mean, following this sequence, we again asked participants to report their general expectation for donations (guess 2).

We then showed a test sequence for which the true mean was the same as the mean of the training sequence. Critically, the sparsity of below-average or above-average donors was manipulated between participants. Participants in the sparse 'stingy' condition ($N=34$) observed a sequence where the below-average ('stingy') donor distribution was sparser than the above-average ('generous') donor distribution. The amounts donated by the fewer 'stingy' donors were more variable than the amounts donated by the many 'generous' donors, to maintain the overall mean of 150 (Fig. 1b). Participants in the sparse 'generous' condition ($N=42$) observed a sequence whose donation values were flipped such that there were fewer and more variable 'generous' donors. After observing the test sequence, participants first estimated the average donation for the test sequence. We then asked them to estimate the average donation for the training sequence.

We predicted that the average estimate after observing the test sequence would be biased such that values from the sparse area defined by higher variance and fewer samples would weigh more heavily, leading estimates in the sparse 'stingy' condition to be lower than those in the sparse 'generous' condition. To account for individual differences in prior beliefs, we normalized estimates by subtracting individual guesses prior to the test sequence observation (guess 2) from the post-observation estimate. These normalized estimates were the main dependent variable of interest.

A two-tailed t test on the difference between the normalized estimates in the two sparsity conditions showed a significant difference ($t(74) = -2.744$, $P=0.008$, Cohen's $D=-0.633$, 95% CI [−0.077, −0.012]), with the normalized estimate (the post-observation estimate minus 'guess 2') in the sparse 'stingy' condition ($M=-0.020$) significantly below the estimate in the sparse 'generous' condition ($M=0.024$). A permutation test in which condition labels were shuffled 2,000 times further showed that the observed estimate difference between sparsity conditions lay outside the null distribution confidence interval (Fig. 2a; estimated normalized difference 0.044; 95% null distribution CI [−0.030, 0.033], $P=0.006$).

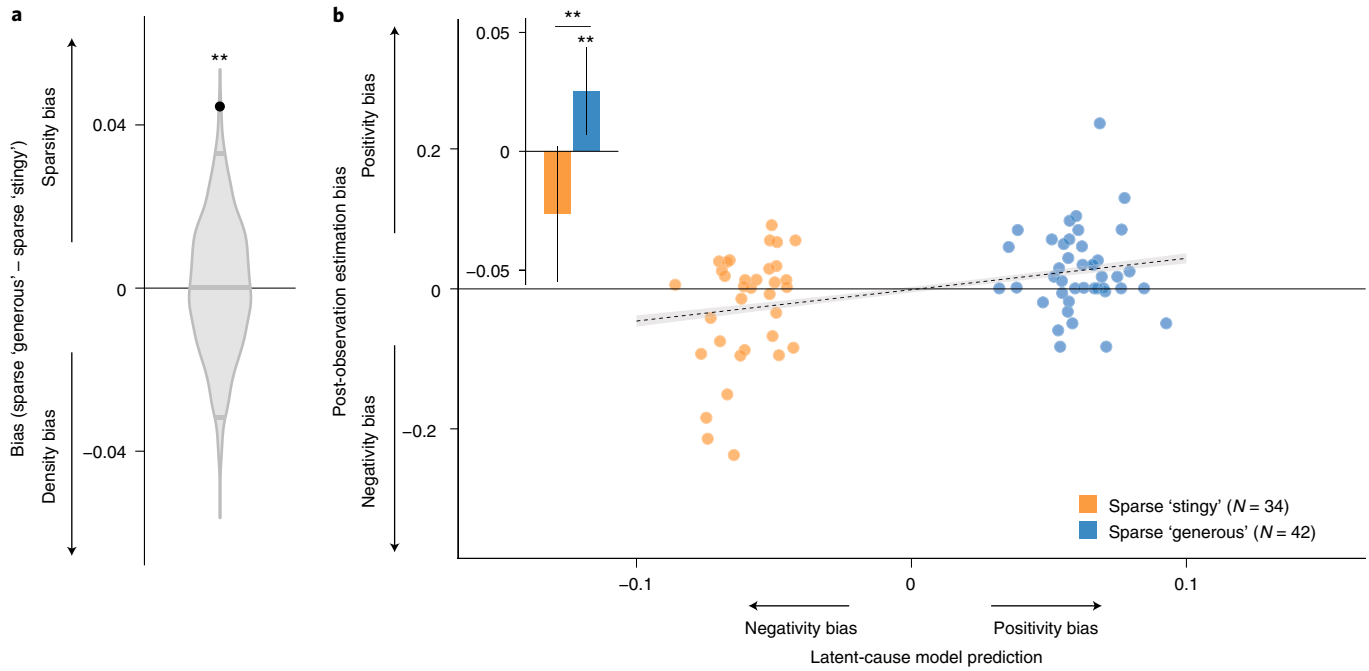

Fig. 2 | Results of experiment 1A. a, The average estimate shows a sparsity bias. A permutation test in which condition labels were shuffled 2,000 times (grey shaded) showed that the observed difference between the normalized post-observation estimate in the sparse 'generous' and the 'stingy' conditions (black dot) lay outside the null distribution confidence interval (horizontal lines in violin plot). **b**, Comparison between simulation results and behavioural results. Each dot represents the estimation bias predicted by a latent-cause model that experienced the participant-specific sequence (*x* axis) and the behavioural unnormalized post-observation estimation bias (post-estimation minus the true mean) of that participant (*y* axis). Linear regression showed that estimate biases simulated by the latent-cause inference model predicted the estimate biases observed in the experiment (illustrated by the dashed line; see text). Inset shows the post-observation estimation bias difference between the sparse 'stingy' (orange) and sparse 'generous' (blue) conditions. Error bars indicate 95% confidence intervals. $^{**}P < 0.01$.

We then asked whether the normalized post-observation estimates show a sparsity bias (that is, a positivity bias in the sparse 'generous' condition and vice versa) from the true mean in each sparsity condition (Fig. 2b, inset). One-sample *t* tests showed that there was a significant sparsity bias in the sparse 'generous' condition ($M = 0.025$; $t(41) = 2.772$, $P = 0.008$, Cohen's $D = 0.428$, 95% CI [0.007, 0.044]), whereas we found no statistically significant effect of negativity bias in the sparse 'stingy' condition ($M = -0.026$; $t(33) = -1.884$, $P = 0.068$, Cohen's $D = -0.323$, 95% CI [−0.055, 0.002]).

The raw post-observation estimates were also significantly different across conditions ($t(74) = -3.196$, $P = 0.002$, Cohen's $D = -0.737$, 95% CI [−0.084, −0.019]), with the sparse 'stingy' condition more negatively biased than the sparse 'generous' condition. To further ensure that the condition difference at the post-observation estimate was attributable to the sequence manipulation, we tested for differences between conditions in the initial, pre-observation guesses (guess 2). As expected, there was no evidence for a statistically significant difference in guess 2 between conditions (sparse 'stingy' condition $M = -0.006$; sparse 'generous' condition $M = 0.001$; $t(74) = -0.912$, $P = 0.365$, Cohen's $D = -0.210$, 95% CI [−0.023, 0.009]).

To test whether the biases can be predicted by a latent-cause inference process, we compared human participants' biases with biases of an approximate Bayesian inference algorithm that infers latent causes using an infinite capacity prior (called a Chinese Restaurant Process prior). This prior allows the assignment of each donor to a single latent cause, without predetermining the overall number of latent causes. According to the prior, if a latent cause already generated many donors, the prior probability that this popular cause would generate the next donor is higher (the 'rich-get-richer' property). However, there is always a chance that a completely new latent cause will produce the next donor, allowing flexibility in creating any number of causes. In combination with this prior probability, a likelihood was assigned to each observation (donation amount) according to its similarity to other observations inferred to be generated by the same latent cause (see Methods for details). After observing the sequence of donations and inferring the donors' latent causes, the model estimated the average donation by taking the mean of the average donation of each latent cause weighted by the log-transformed number of donors who were assigned to the cause. We used the logarithm of the number of events rather than the true number to account for participants' loss of precision over counts as more events are experienced (as in Weber's law for many perceptual estimations). As a result, causes with a smaller number of donors influenced the overall mean disproportionately, as compared with causes with a larger number of events. We therefore predicted that events from the sparse area will have higher impact on the estimated average than events from the dense area, regardless of normalized valence (that is, above or below average), because a larger number of distinctive latent causes would be inferred to account for events in the sparse area, while a small number of causes would be inferred for the dense area.

The model showed the sparsity bias (mean condition difference 0.12, $t(74) = 43.622$, $P < 0.001$), as in the behavioural results. Because in our model presentation order influences groupings of the observed values (Methods[12]), we provided the values of the donations to the model in the same order in which each participant observed them. The model could thus make specific predictions about the estimation bias per each individual participant. We used linear regression to test whether the simulated biases predict individual behavioural biases. Results showed that the estimate biases simulated by the latent-cause inference model predicted the

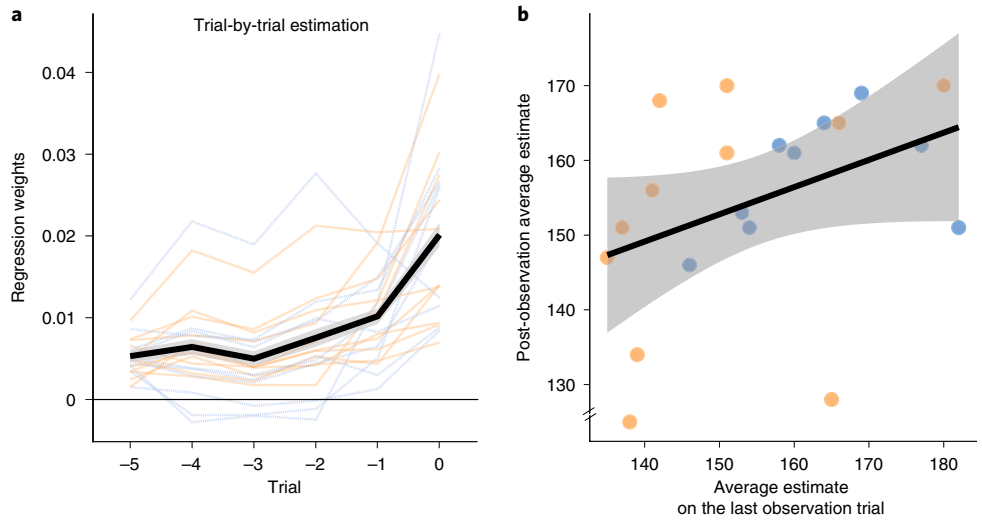

Fig. 3 | Results of experiment 1B. a, Recency biases in trial-by-trial estimates. Linear regression showed a recency bias, where more recent observations contributed more strongly to the current estimate (orange lines: sparse 'stingy' condition; blue lines: sparse 'generous' condition). **b**, Recency bias and final post-observation average estimate. The trial-by-trial estimate on the last trial (*x* axis) was marginally predictive of the post-observation (and post filler-task) estimate (*y* axis). Each dot represents one participant; grey shading shows 95% confidence intervals.

estimate biases behaviourally observed in the experiment ($\beta = 0.028$, $P < 0.001$, partial $\eta^2 = 0.145$; Fig. 2b, dashed line). This relationship was marginally significant even when controlling for the sparsity condition ($\beta = 0.080$, $P = 0.061$, partial $\eta^2 = 0.047$).

The core assumption of our latent-cause model is that estimation of the average donation is made by averaging over summary values of latent causes rather than the observations themselves. However, if participants knew in advance that they would only need to keep track of one summary value (that is, the mean of all observations) they might update a running average rather than (or in addition to) grouping donors into latent causes[13]. To test this, in experiment 1B ($N = 22$), we asked participants to report their average estimate on every trial (with all other procedures being identical to experiment 1A). We predicted that this requirement would eliminate biases toward the sparse area, and may generate biases towards the most recently experienced donations ('recency bias').

Results showed that, when participants were required to estimate the mean donation after every observation, the normalized estimates were no longer significantly different across conditions (sparse 'stingy' $N = 10$, $M = 0.032$; sparse 'generous' $N = 12$, $M = 0.012$; $t(20) = 0.984$, $P = 0.337$, Cohen's $D = 0.421$, 95% CI [−0.023, 0.063]), with the numerical difference in the direction opposite to the latent-cause model prediction. A two-way ANOVA on data pooled from both experiments 1A and 1B showed a significant interaction between the sparsity and tracking conditions ($F(1,94) = 4.056$, $P = 0.047$, partial $\eta^2 = 0.04$), suggesting that the sparsity in the distribution induces biases only when the average values are not tracked on a trial-by-trial basis.

Further, we explored the relationship between observation, trial-by-trial estimates and final post-observation estimates in experiment 1B. Linear regression showed that estimates were influenced more strongly by more recent observations (Fig. 3a). The trial-by-trial estimate on the last trial, however, was only marginally predictive of the post-observation (and post filler-task) estimate of the total donation mean (Fig. 3b; $\beta = 5.362$, $P = 0.082$, partial $\eta^2 = 0.158$). Together, these results suggest that the overall estimate of the average may be derived via a different strategy when there is a clear goal of tracking the average value throughout the task, and support our hypothesis that latent-cause inference could be the mechanism by which sparse events become overweighted in the overall estimate.

Experiment 2. In experiment 1, the sparsity manipulations induced biases in average estimation when there was no explicit goal of tracking the average value, which we interpreted as resulting from a process of latent-cause inference. However, there are at least two alternative explanations for the bias we observed. First, even if participants did not infer multiple causes for the sparse events but rather perfectly inferred that there are two latent causes ('stingy' and 'generous'), a log-weighted average of the mean values of these two causes would have resulted in an estimate that is biased toward the cause that has fewer events, due to the uneven number of events in the two causes. To address this, in experiment 2, we equated the frequency of events that are generated by the 'stingy' and 'generous' causes and only manipulated the variance of the event distributions (Fig. 4a).

Second, Pearce and Hall (1980) suggest that more surprising events will update values of an entity to a greater degree. Because events in the sparse area elicit higher prediction errors (surprises), they may have a greater impact on the learned averages (formally, these surprising events will have a higher learning rate). To adjudicate between the latent-cause inference model and the Pearce–Hall dynamic learning rate model, in experiment 2, we chose a specific presentation order where the distributions from which observed values were drawn quasi-alternated between the dense and sparse distributions, and the end of the sequence was predominantly populated with values from the dense distribution (Fig. 4a). Alternating between the dense and sparse distributions made the trials from both distributions similarly surprising on average, eliciting similar levels of prediction error and therefore equating attention to both distributions in the Pearce–Hall model. In addition, as values from the dense distribution were observed just prior to average estimation, error-driven learning would show a density bias due to the enhanced effect of recent experiences in such models (Supplementary Fig. 2). Given these properties of the chosen presentation order, the Pearce–Hall model with dynamic learning rates predicted a density bias, while latent-cause inference still predicted a sparsity bias (Fig. 4b).

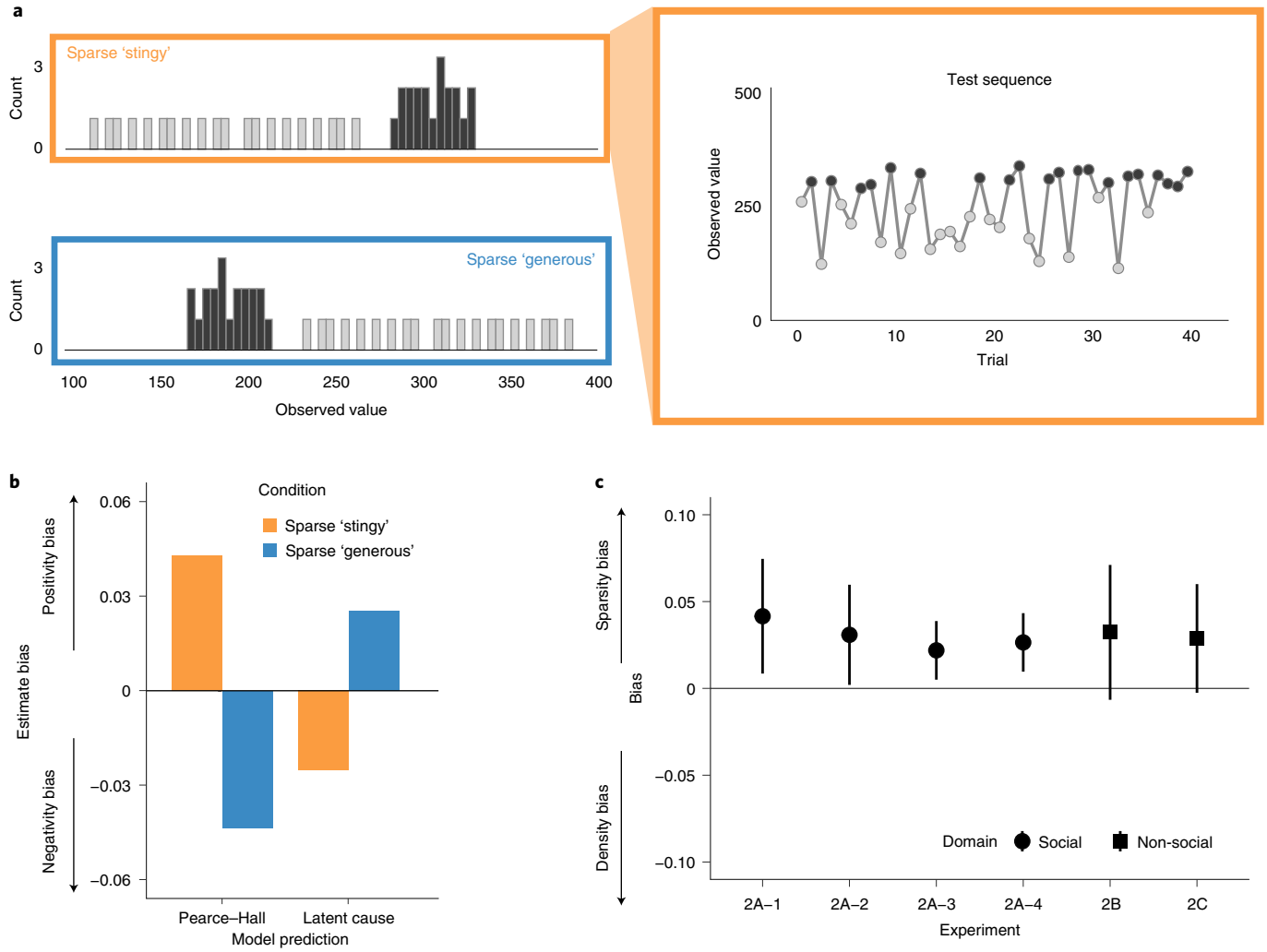

Fig. 4 | Experimental design and results of experiment 2. Sparsity biases when rarity was matched between the sparse and dense causes. **a**, Distributions and test sequence. In experiment 2, the frequency of 'generous' (dark grey) and 'stingy' (light grey) donations was matched (left). For each condition, we also chose a specific sequence such that 'generous' and 'stingy' donations would elicit similar levels of prediction error (right). In this sequence, the two distributions mostly alternated throughout the sequence, making both donation amounts locally surprising. Finally, the sequence ended with values drawn from the dense distribution (see text). **b**, Model predictions and empirical results. For the sequences in experiment 2, the Pearce–Hall model (left) and latent-cause model (right) predicted biases in opposite directions. **c**, The empirical results were in line with the latent-cause model prediction, showing a sparsity bias. Error bars indicate 95% confidence intervals.

Experiment 2A's cover story was that each observation was the amount that a community member was willing to pay for a charity event. All other procedures were identical to experiment 1A. In the first study (experiment 2A-1, $N=70$), a two-sample t test showed that the normalized estimate in the sparse 'stingy' condition ($M=-0.023$, $N=26$) was significantly lower than that in the sparse 'generous' condition ($M=0.018$, $N=44$; $t(68) = -2.511$, $P=0.014$, Cohen's $D=-0.621$, 95% CI [−0.075, −0.009]). A permutation test in which condition labels were shuffled 2,000 times further supported that the observed sparsity bias effect lay above the null distribution confidence interval (normalized condition difference 0.044; 95% null distribution CI [−0.035, 0.033], $P<0.05$). Critically, these results were in the direction predicted by the latent-cause inference model and opposite to those of the Pearce–Hall model (Fig. 4c).

To strengthen this finding, we ran three independent sets of replications (experiment 2A-2, $N=67$; experiment 2A-3, $N=260$; experiment 2A-4, $N=229$; pre-registered https://aspredicted.org/99em9.pdf). All three experiments replicated the main finding that the sparse 'stingy' condition's normalized estimates (experiment 2A-2, $N=28$, $M=-0.014$; experiment 2A-3, $N=133$, $M=-0.009$; experiment 2A-4, $N=118$, $M=-0.005$) were below the sparse 'generous' condition's normalize estimates (experiment 2A-2, $N=39$, $M=0.017$; experiment 2A-3, $N=127$, $M=0.013$; experiment 2A-4, $N=111$, $M=0.021$), showing sparsity bias (Fig. 4c). The difference between conditions was significant in all three replications (experiment 2A-2, $t(65) = -2.137$, $P=0.036$, Cohen's $D=-0.529$, 95% CI [−0.060, −0.002]; experiment 2A-3, $t(258) = -2.556$, $P=0.011$, Cohen's $D=-0.317$, 95% CI [−0.039, −0.005]; experiment 2A-4, $t(198.54) = -3.098$, $P=0.002$, Cohen's $D=-0.407$, 95% CI [−0.043, −0.010]). An additional meta-analytic Bayes factor analysis across all four instances of experiments 2A using the 'BayesFactor' R package[14,15] showed a Bayes factor of $BF_{+0}=17{,}732$, indicating that the data are 17,732 times more likely under the hypothesis that the normalized estimates are different across sparsity manipulations than under the null hypothesis.

We then tested whether post-observation estimates in each condition show a sparsity bias from the true mean across these four replication data sets, using meta-analytic Bayes factor analyses on the

two-tailed t tests (Supplementary Fig. 3). Negativity biases observed in the sparse 'stingy' condition (experiment 2A-1, $M=-0.027$; 2A-2, $M=-0.023$; 2A-3, $M=-0.012$; 2A-4, $M=-0.017$) were 636.6 times more likely under the bias hypothesis as compared with the null hypothesis ($BF_{+0}=636.584$), providing extremely strong evidence for a bias. In the sparse 'generous' condition, positivity biases (experiment 2A-1, $M=0.006$; 2A-2, $M=0.008$; 2A-3, $M=0.006$; 2A-4, $M=0.014$) were 4.6 times more likely under the hypothesis that there is a bias ($BF_{+0}=4.556$), providing moderate evidence for a bias.

We further sought to investigate whether the sparsity bias in the social domain emerges from a domain-general inference process, or is specific to the type of social evaluation cover story. We thus ran two more experiments with different cover stories. In experiment 2B, participants were asked to observe the weight of coffee beans that customers buy from different towns, and then estimate the average coffee bean purchase. In experiment 2C, participants observed slot machine earnings from different casinos, and then estimated the average win. All other procedures were identical to experiment 2A, and the sparse 'below-average' and 'above-average' conditions here respectively corresponded to the sparse 'stingy' and 'generous' conditions in experiment 2A. If summary statistics of a number of observations are estimated differently in more social versus less social domains, we would expect to observe an interaction in the bias between the sparsity conditions and the domains (social and non-social). Specifically, in the non-social domain, we would expect no difference between the sparsity conditions or a difference in the opposite direction (that is, the density bias), following the Pearce–Hall model's prediction and a traditional error-correcting learning process. However, if the bias arises from a latent-cause inference process that is shared across domains, the sparsity in the observed values should lead to biases regardless of domain.

Experiment 2B ($N=81$) and 2C ($N=101$) showed that the sparse 'below-average' condition's normalized estimates (experiment 2B, $N=38$, $M=0.015$; experiment 2C, $N=51$, $M=0.001$) were numerically below those of the sparse 'above-average' condition (experiment 2B, $N=43$, $M=0.048$; experiment 2C, $N=50$, $M=0.030$), although neither experiment showed a statistically significant bias (two-tailed two-sample t tests: experiment 2B, $t(79) = -1.654$, $P=0.102$, Cohen's $D=-0.368$, 95% CI [−0.071, 0.007]; experiment 2C, $t(99) = -1.822$, $P=0.071$, Cohen's $D=-0.363$, 95% CI [−0.060, 0.003]). Nevertheless, differences in significance are not an indication of significant differences between the domains. To test whether the difference in post-observation estimates across conditions interacted with the domain, we ran a mixed-effects linear regression model predicting normalized estimates with the sparsity and social ('social' or 'non-social') conditions as fixed effects and experiments (experiments 2A-1, 2A-2, 2A-3, 2A-4, 2B and 2C) as random effects, pooling data across all six experiments that used the same event sequences ($N=808$). Tests of significance using Satterthwaite's approximation showed no significant interaction between the sparsity and social conditions ($\beta = 0.005$, s.e. 0.012, $t(804) = 0.393$, $P=0.695$). A Bayes factor analysis provided moderate evidence that there is no interaction between event domain (social versus non-social) and the sparsity conditions ($BF_{0+}=7.637$).

Examining each condition's sparsity bias in the non-social domain (Supplementary Fig. 3) showed strong evidence for a positivity bias in the sparse 'above-average' condition ($BF_{+0}=16.153$; null hypothesis, Cohen's $D=0$), and moderate evidence for lack of bias in the sparse 'below-average' condition ($BF_{0+}=4.902$). This absence of a negativity bias in the sparse 'below-average' condition in the non-social domain may be due to prior expectations for the observations, given the cover stories used. That is, if prior beliefs caused participants to expect smaller values in the non-social domain, above-average events that deviate strongly from expectations would be more diagnostic and weigh more heavily in overall estimation of observed events. To this end, we ran a mixed-effects linear regression model predicting a participant's a priori estimates (guess 2) with social conditions as fixed effects and experiments as random effects. A test of significance using Satterthwaite's approximation showed that prior beliefs in the non-social domain ($M=-0.015$) were significantly below priors in the social domain ($M=-0.007$; $\beta = 0.007$, s.e. 0.003, $t(7.308) = 2.539$, $P=0.037$).

Discussion

Together, these experiments demonstrated that overall estimation of a quantity is biased toward the value of events that are rare and/or more variable. Comparing human participants' biases to simulated estimation biases from a semi-rational latent-cause inference model suggested that the behavioural results can be attributed to a process of inferring latent causes for observations and estimating the overall average by averaging over these causes.

Our results are in line with empirical findings in social cognition research showing that, given rarity[2] and variability[6,16] of negative events, negative information can have a higher impact in impression formation and updating[1]. To test our hypothesis that the distributional sparsity is driving such bias, we used donation events in a positive monetary domain and manipulated sparsity of below-average (relatively negative) or above-average (relatively positive) events. This was a strong test of our hypothesis, as we avoided events in the negative domain altogether so as to not confound our findings with subjective value differences for monetary wins and losses (that is, the fact that losses loom larger than gains of the same amount[17]). We expect that the effects we observed would occur even more strongly when the valence of the stimuli varies across the full spectrum of negativity and positivity, since there are features of negative events that make biased processing of negative stimuli adaptive. For instance, an untrustworthy person can impose a risk on our wellbeing, and thus it would be wise to avoid such risk. Indeed, if we choose to avoid a presumably untrustworthy (or otherwise negatively assessed) person, we deprive ourselves of opportunities to observe their behaviour and update our impression. Effectively, the interaction between first impressions and our behavioural choices makes our samples of that person's behaviours sparse, leading to a negativity bias[18]. On the other hand, we may update our impressions differently when we observe bad behaviours, leaving more chance to forgive potentially bad targets[19]. Moreover, valence may not change monotonically with magnitude for some behaviours. For example, talking too much could be evaluated as negatively as talking too little[16], while donating more money is generally positive, and thus avoids this potential confound. Future studies could explore the potential interaction between sparsity and the valence of events in social and non-social domains.

We explored whether the inference processes that give rise to the sparsity bias are domain general or uniquely social, by using various social and non-social scenarios. There was moderate evidence that the sparsity bias was not different across domains, suggesting domain-general inference processes. However, our social and non-social cover stories may diverge in other important ways, especially with regard to prior beliefs. Although we presented participants with a training sequence to neutralize their prior beliefs, the training may not be enough to adjust people's expectations for various situations, as these were built through a lifetime of experiences. This may be the reason that there was a difference in 'guess 2' between the social and non-social domains. In a donation scenario, the donation amounts observed during the first training sequence may be a more informative ground for processing the next set of observations, as they form a social norm for generosity that the members of the community should follow. This would make the prior beliefs sharper, and any donations outside the normal range could be perceived as good or bad. On the other hand, our non-social scenarios (how many grams of coffee beans are purchased, or how

much slot machines return) may have evoked priors whose values are lower than the donation scenario, and thus every win in a casino or sale of coffee beans, even if below average, may have continued to be seen as a positive event. This would be a potential explanation for the absence of a negativity bias when below-average events were sparse in the non-social scenarios. Future research should investigate how prior beliefs differ across situations, and whether priors under social situations have unique characteristics that alter the inference process or allow it more flexibility. Quantifying individuals' prior beliefs can also allow the model to predict individual differences in negativity biases, which indeed tend to be stable within an individual across time[20]. For instance, an individual who has experienced largely negative life experiences may have a prior expectation that positive events are sparse, and therefore will weigh positive events more heavily in their overall estimation.

Our study also differs from previous impression-formation studies in that we assessed the impression of a group rather than an individual. When forming impressions about an individual, we tend to assume more unity and coherence in their behaviours than we do about groups, drawing inferences about dispositional properties[21–23]. Group impression formation may rely on a different process from person impression formation[24], which involves representing multiple individual experiences, or exemplars[25], depending on perceived entitativity (that is, the degree of having the properties of an entity)[26,27] of a group. The level of judgement, thus, can vary between lumping all individuals into one category and representing each individual as its own entity. This is reminiscent of the tension between representing only prototypes of a category (prototype models of categorization[28]) and preserving all exemplars (exemplar models[29])[30]. The latent-cause inference model can be seen as an intermediate between the two alternatives, achieving either end of the spectrum, and the range between them, by varying a single parameter that governs the probability of creating a new cluster[31]. This model could therefore be suitable for group impression formation. That is, the model can provide a framework to further explore how entitativity influences group impression formation. For example, in experiments 2B and 2C, where the causal link between observed behaviour and the groups was weak, the sparsity bias predicted by the latent-cause inference model was less pronounced. A town might not be a coherent entity to predict coffee sales as much as a school is for predicting social norms such as generosity. That is, when evaluating a heterogeneous social group that we assume has a common latent causal property (such as intention) that generates individual observations, sparse experiences with the group can drive our overall impression of the group. In any case, if these biases are the results of fundamental inferential processes that partition our experiences into meaningful causal chunks, the model should hold true in individual impression formation as well. Given that there is a closer and more immediate causal link between an individual and their actions than between a group and the group members' actions, the sparsity bias effect would potentially even be stronger for individual impressions.

Experiment 1 suggested a way to reduce the sparsity bias. We showed that the requirement to evaluate the overall mean after every observation promoted unbiased estimation, as evidenced by the interaction between the task requirements and the sparsity conditions. This suggests that we may be less affected by rare and variable interactions if we try to track a particular quality of another person every time we interact with them, rather than leaving the judgement until later. This could be desirable in a situation where we want an unbiased evaluation, for instance during a hiring process. Nevertheless, placing people on a positive–negative scale is usually not the only goal in our rich day-to-day interactions, and we often need more flexible representation of our social counterparts.

Our model estimates the overall average by taking the log-weighted mean of latent causes' mean values. That is, the model assumes that low-frequency events are relatively overweighted and high-frequency ones are underweighted when the overall mean is estimated based on the latent causes' mean donation, as each cause is weighted by the log-transformed number of donors that were assigned to the cause. This loss of precision is based on numerous studies showing this exact pattern of distortions in frequency or probability space[32]. Although this type of precision loss assumed by our model is repeatedly found in literature[33,34], it is worth noting that the degree of sparsity bias depends on the degree to which the latent causes' frequency information is distorted. If latent-cause frequency were perfectly kept, there would be neither a sparsity nor a density bias. At the other extreme, if frequency information were completely lost such that the overall mean would be taken as the mean of the latent causes, with equal weights for all causes, the sparsity bias would be much stronger. This could also explain the stronger sparsity bias in social scenarios as compared with non-social situations, as we represent the groupings of people by relying on existing schema that we already have from previous experiences with other people, thereby further losing precision on frequency of encounters in this particular setting.

Another possibility is that the frequency is distorted in the inference process as well as in the averaging of inferred causes. The Chinese Restaurant Process prior is a rich-gets-richer process where a new event will more likely be assigned to a cause with a larger number of events already assigned to it, than to unpopular causes. Inference in this model therefore requires counting how many events already belong to the cause. Including the aforementioned frequency distortion in this counting step does not change the direction of the bias, although the degree of the bias decreases. Similar distortions can occur even when there is no inference involved. That is, if the frequency of values observed multiple times is lost (that is, only unique observations are maintained in memory), the dense distribution may contribute less to the overall mean, due to the higher chance of repeated observations. However, this type of frequency distortion cannot account for the biases seen in experiment 2A, as the values of donations were unique per trial, with no repetition. Furthermore, in experiment 1A, where outcomes were repeated, if we took into account only the first presentation of each value, the mean of unique donations in the sparse 'stingy' condition would be above the true mean and vice versa, which is the opposite direction to the sparsity bias observed in the data.

Finally, logarithmic transformation can occur in representing the donation amount as well. If the donation amount is perceived on a logarithmic scale, the final average estimation would show a negativity bias in our experiments. To test whether this logarithmic representation of values is the source of negativity bias, we conducted a control experiment matching the log-transformed means, not the linear-scale means of the 'sparse stingy' and 'sparse generous' conditions. There, the estimation biases were positive in both conditions, contrary to the log-scale representation prediction (Supplementary Fig. 1), suggesting that logarithmic representation cannot account for the empirical biases.

Given the possible hypotheses about precision loss when accounting for the number of observations in each latent cause, an ideal approach would be to fit these models to empirical data and compare which type of precision loss predicts our results the best. This is not possible in the current study, as we collected only one estimate per participant. We chose the current design to prevent continuous reporting from altering the cognitive processes by which latent causes are inferred and the final evaluation is made. Of course, this choice came at the expense of model fitting; future work could characterize the online inference process, for instance, by probing groupings on a trial-by-trial basis.

In conclusion, we have shown evidence that supports an account of evaluation as consisting of a stage in which sparse experiences are segmented into a large number of latent causes, which in turn

bias the overall impression such that rarer and/or more variable experiences are overweighted. Here, we showed the sparsity bias in a mildly social domain. We would expect that, in more realistic social scenarios that involve evaluations of others with real stakes at hand, the biases may manifest even more strongly. This cognitive bias could indeed be the core mechanism underlying the negativity bias in social evaluations.

Methods

Experiment 1A. *Participants.* Seventy-six participants (34 in the sparse 'stingy' condition, 42 in the sparse 'generous' condition) were recruited using Amazon Mechanical Turk (MTurk). The Princeton University Institutional Review Board approved the experiment, and we obtained informed consent online before participants began the task.

Participants were excluded when they did not pass the following criteria: (1) not completing the task to the end, (2) failing to answer correctly on attentional checks, (3) responding too slowly (> 60 s) on any of the observation and (4) not adjusting general expectation properly after the training sequence and making a guess with a value that they never observed during the training trials. This filtering was done to ensure that participants attended to every observation, as our prediction is based on the particular set of values and order of the events.

Materials and procedures. Participants were told that they were visiting different schools for fundraiser events and their job was to log the donation amounts. They observed 40 donors making donations in each school. Each donor made a single donation with coins, ranging from 1 to 300 coins.

Participants first guessed how much people would donate in general (guess 1). Then they observed the training sequence ('Brookview University') and logged the donation amounts. On each trial, coins were dropped on the screen with a prompt indicating the donor and the amount (for example, 'Bradley donated 148 coins.'). Participants made a response either on a slider bar or in a text box next to the slider. The two response methods were yoked such that moving the slider would show the number in the text box and putting a number in the text box would move the slider to the number. The trial could proceed only when the response exactly matched the prompted amount.

The purpose of the training sequence was to adjust participants' overall expectations, and reduce individual differences in prior estimation. At the end of the training sequence, we therefore asked participants again to guess the general donation amount (guess 2), to ensure that participants adequately adjusted their prior beliefs.

The mean donation amount of the test sequence was matched to the mean donation amount of the training sequence. They repeated the logging task with the test sequence ('Cedar Springs University'). Here, we manipulated the sparsity of distributions of 'stingy' and 'generous' donors between participants. A 'stingy' donor was operationalized as a donor who made a below-average donation, and a 'generous' donor was one who made an above-average donation. Note that participants already learned the average donation amount in the training sequence, and thus they have an anchor to judge below- and above-average donations. The sparsity was defined by rarity and variance.

After observing all sequences and finishing a filler task, a surprise test asked them to estimate the average donation of the test sequence. This was followed by a test on the average donation of the training sequence.

In the sparse 'stingy' condition, there were fewer stingy donors with higher variance in donation amount (10 'stingy' donors; $M = 79.7$, s.d. 35.86) than generous donors (30 'generous' donors; $M = 173.73$, s.d. 10.28). In the sparse 'generous' condition, we flipped the donor distributions such that the overall mean stays the same with fewer generous payers and more variable generous donations (10 'generous' donors, $M = 220.3$, s.d. 35.86; 30 'stingy' donors, $M = 126.26$, s.d. 10.28).

Experiment 1B. *Participants.* Twenty-two participants (10 in the sparse 'stingy' condition, 12 in the sparse 'generous' condition) were recruited using Amazon Mechanical Turk (MTurk). Exclusion criteria were identical to experiment 1A.

Materials and procedures. We added an average estimation task upon each observation. After observing and logging each donation, participants were asked to estimate the average thus far. All other procedures and materials were identical to those used in experiment 1A.

Experiment 2A. *Participants.* A total of 626 participants were included in experiment 2A (experiment 2A-1: $N = 70$, sparse 'stingy' $N = 26$, sparse 'generous' $N = 44$; experiment 2A-2: $N = 67$, sparse 'stingy' $N = 28$, sparse 'generous' $N = 39$; experiment 2A-3: $N = 260$, sparse 'stingy' $N = 133$, sparse 'generous' $N = 127$; experiment 2A-4: $N = 229$, sparse 'stingy' $N = 118$, sparse 'generous' $N = 127$). For experiment 2A-3, the sample size was chosen from a power analysis based on experiment 2A-1 and 2A-2. For experiment 2A-4, we took a Bayesian approach[35] and collected a minimum of 50 usable participants in each condition and continued data collection until we reached one of three criteria: (1) a Bayes Factor of 10 in favour of H+ (normalized estimate in sparse 'stingy' condition < normalized estimate in sparse 'generous' condition) and against H0 (no difference in normalized estimates between sparsity conditions), (2) a Bayes factor of 10 in favour of H0 and against H+ or (3) we reached the maximum number of participants (500 usable participants in each condition). This procedure was pre-registered (https://aspredicted.org/99em9.pdf).

Exclusion criteria were identical to experiments 1A and 1B, except that participants who missed the 5-s response window for logging the amount were excluded.

Materials and procedures. We used a different cover story to generalize our results. In experiment 2A, participants were told that they were selling coffee for charity events at community fairs in different towns ('Lambtonville' and 'Brookfield') and taking coffee orders. The customers could pay in tokens as they wish, and participants' task was to log the payment amount for each customer. Customer names were shown in the prompts (for example, 'Brennan: 218 tokens for Cappuccino'). We instructed participants to pay attention to both the names and the payment amount, as some pairs of name and payment amount would be tested at the end. This was to orient them to pay attention to the task (for experiments 2A-1 and 2A-2, we asked participants to report the payment amount for given customers at the end of the experiment; for experiment 2A-3 and 2A-4, we did not test participants' memory. In all cases, we did not analyse these data, as they were outside the scope of our interest).

Tokens did not appear on the screen as visual cues (as they did in experiment 1), and the response was made either by moving a slider ranging from 1 to 500 tokens (experiment 2A-1 and 2A-2; to help participants make a response within the response window, the slider snapped to the correct number when the distance between the marker and the target was less than five tokens) or by typing in the number (experiment 2A-3 and 2A-4), with a 5-s time limit. Participants earned a 50-cent bonus if they did not miss any orders.

Critically, the sparsity of 'stingy' and 'generous' observations was manipulated by variance alone. In both conditions, the number of customers generated from stingy and generous causes were matched to 20. In the sparse 'stingy' customer condition, the 'stingy' customers' payment amounts were more variable than the 'generous' donors (20 'stingy' donors $M = 188$, s.d. 47.29; 20 'generous' donors $M = 308$, s.d. 13.73), and vice versa in the sparse generous customer condition (20 'generous' donors $M = 312$, s.d. 47.29; 20 'stingy' donors $M = 192$, s.d. 13.73).

The order of payment values was chosen such that the latent-cause inference model and the Pearce–Hall model predict the opposite biases.

Experiment 2B and 2C. *Participants.* The total of 182 participants participated in experiment 2B ($N = 81$, sparse 'stingy' $N = 38$, sparse 'generous' $N = 43$) and 2C ($N = 101$, sparse 'stingy' $N = 51$, sparse 'generous' $N = 50$). Exclusion criteria were identical to experiment 2A.

Materials and procedures. To investigate the sparsity bias in non-social domains, we changed the cover story such that participants were logging weights of coffee beans for customers in supermarkets in different towns (experiment 2B) or logging slot machine earnings in different casinos (experiment 2C; participants' compensation did not depend on observed earnings to avoid those amounts from playing the role of personally relevant rewards). Critically, the stimuli sequences were identical to experiment 2A, where the sparsity was manipulated by variance. All procedures were identical to experiment 2A-3 and 2A-4, where responses were made in a text box.

Latent-cause inference model. Each event sample was sequentially introduced into a Bayesian inference model with an infinite-capacity Chinese Restaurant Process (CRP) prior[36]. In this model, before observing any behaviour, an observer has prior beliefs about the target group's stable latent causes:

$$p(Z = k) = \begin{cases} \frac{n_k}{\sum n_k + \alpha} & \text{if } k \text{ is an old cause} \\ \frac{\alpha}{\sum n_k + \alpha} & \text{if } k \text{ is a new cause} \end{cases},$$

where Z is a variable denoting the latent cause of the next observation, k indexes latent causes, n_k is the number of observations already assigned to latent cause k and the concentration parameter α determines the prior tendency to assume new latent causes. This prior formalizes the idea that a prolific latent cause is more likely to generate future events (top case), and the total number of latent causes is unbounded and can grow with the number of observations (bottom case).

After observing an event, the likelihood that the current event x_t was generated from latent cause k is estimated by marginalizing over all 'consequential regions'[37] h' that encompass the past events $\{x_i\}_k$ generated by cause k (n_k in total):

$$p(x_t \in k | \{x_i\}_k) = \sum_{h' \in H} p(x_t \in k | \{x_i\}_k, h') p(h' | \{x_i\}_k)$$

The posterior probability $p(h'|\{x_i\}_k)$ in the right-hand side is calculated as

$$p(h' | \{x_i\}_k) = \frac{p(\{x_i\}_k | h') p(h')}{\sum_{h \in H} p(\{x_i\}_k | h) p(h)},$$

where the prior $p(h')$ follows an Erlang distribution[34] with a size prior set to the range of training sequence events, and the likelihood $p(\{x_i\}_k|h')$ is the product of the likelihood of events that are sampled from consequential region h. Under the 'strong sampling'[38] assumption that each event is independently sampled from the cause,

$$p(\{x_i\}_k|h') = \prod_{i:x_i \in \text{cause } k} p(x_i|h').$$

Assuming uniform sampling from the consequential region, the likelihood that event x_i is sampled from consequential region h' is inversely proportional to the width of the region $|h'|$ if the event is within the consequential region, and zero otherwise:

$$p(x_i|h') = \begin{cases} \frac{1}{|h'|} & \text{if } x_i \in h' \\ 0 & \text{otherwise} \end{cases}.$$

This gives, for the likelihood of the current observation under latent cause k,

$$p(x_t \in k|\{x_i\}_k) = \frac{\sum_{h':\{x_i\}_k, x_t \in h'} \frac{1}{|h'|^{n_k}} p(h')}{\sum_{h:\{x_i\}_k \in h} \frac{1}{|h|^{n_k}} p(h)}.$$

The posterior probability of latent cause k is then updated using Bayes rule:

$$p(Z = k|x_t) = \frac{p(x_t|Z = k)p(Z = k)}{\sum_{k'=1}^{t} p(x_t|Z = k')}.$$

Because the Bayesian inference process becomes intractable as the number of observations grows, we approximated the process using a particle filter[10,39] in which each particle maintains a single maximum a posteriori estimate of the assignment of observations to latent causes, rather than maintaining the full posterior distribution. We ran eight simulations (four with $\alpha = 0.25$ and four with $\alpha = 0.5$) using 50 particles each. As the number of true clusters was one for the training sequence and two for the test sequence, the concentration parameters were chosen such that the prior would produce one ($\alpha = 0.25$) or two ($\alpha = 0.5$) clusters after 40 trials.

Finally, the evaluation of a group was made by taking the mean of the latent-cause values weighted by the log number of events assigned to each latent cause. The latent-cause value was taken to be the mean value of events assigned to that latent cause. The estimation bias was calculated by subtracting the true mean value of the events from the mean value estimated from the latent causes.

Pearce–Hall model. Donation amounts were normalized to the maximum potential amount (that is, the maximum amount on the response slider bar) and then introduced into the Pearce–Hall model[3]. The value estimate v was updated according to

$$v_{t+1} = v_t + a_{t+1} \times S \times x_t,$$

where a is the associability parameter, S denotes salience of the cue and x represents the observed amount. The key component of the Pearce–Hall model is that the associability a is updated according to the absolute prediction error (the difference between the observed and expected values), with a learning rate η:

$$a_{t+1} = (1 - \eta) \times a_t + \eta \times |x_t - v_t|.$$

This means that more surprising events have greater impact on the overall value estimates. We ran simulations with salience parameter S ranging from 0.1 to 1, and learning rate η ranging from 0.1 to 1. The evaluation of a group v was made using each combination of the two parameters.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available at https://osf.io/fdcvw.

Code availability

Custom code that supports the findings of this study is available from the corresponding author upon request.

Received: 16 October 2018; Accepted: 2 February 2021;
Published online: 08 March 2021

References

1. Rozin, P. & Royzman, E. B. Negativity bias, negativity dominance, and contagion. *Personal. Soc. Psychol. Rev.* **5**, 296–320 (2001).
2. Fiske, S. T. Attention and weight in person perception: the impact of negative and extreme behavior. *J. Pers. Soc. Psychol.* **38**, 889–906 (1980).
3. Pearce, J. M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
4. Mende-Siedlecki, P., Cai, Y. & Todorov, A. The neural dynamics of updating person impressions. *Soc. Cogn. Affect. Neurosci.* **8**, 623–631 (2013).
5. Alves, H., Koch, A. S. & Unkelbach, C. The 'common good' phenomenon: why similarities are positive and differences are negative. *J. Exp. Psychol. Gen.* https://doi.org/10.1037/xge0000276 (2017).
6. Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M. & Danner, D. Why positive information is processed faster: the density hypothesis. *J. Pers. Soc. Psychol.* **95**, 36–49 (2008).
7. Alves, H. et al. A density explanation of valence asymmetries in recognition memory. *Mem. Cogn.* **43**, 896–909 (2015).
8. Alves, H., Koch, A. S. & Unkelbach, C. A cognitive–ecological explanation of intergroup biases. *Psychol. Sci.* **29**, 1126–1133 (2018).
9. Courville, A. C., Daw, N. D. & Touretzky, D. S. in *Advances in Neural Information Processing Systems 17* (eds Saul, L., Weiss, Y. & Bottou, L.) 313–320 (MIT Press, 2005).
10. Gershman, S. J., Blei, D. M. & Niv, Y. Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
11. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
12. Austerweil, J. L. & Griffiths, T. L. A nonparametric Bayesian framework for constructing flexible feature representations. *Psychol. Rev.* **120**, 817–851 (2013).
13. Eyal, T., Hoover, G. M., Fujita, K. & Nussbaum, S. The effect of distance-dependent construals on schema-driven impression formation. *J. Exp. Soc. Psychol.* **47**, 278–281 (2011).
14. Morey, R. D., Rouder, J. N. and Jamil, T. *BayesFactor package* https://richarddmorey.github.io/BayesFactor/ (2015).
15. Rouder, J. N. & Morey, R. D. A Bayes factor meta-analysis of Bem's ESP claim. *Psychon. Bull. Rev.* **18**, 682–689 (2011).
16. Alves, H., Koch, A. S. & Unkelbach, C. Why good is more alike than bad: processing implications. *Trends Cogn. Sci.* **21**, 69–79 (2017).
17. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–291 (1979).
18. Denrell, J. Why most people disapprove of me: experience sampling in impression formation. *Psychol. Rev.* **112**, 951–978 (2005).
19. Siegel, J. Z., Crockett, M. J. & Dolan, R. J. Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition* **167**, 201–211 (2017).
20. Ito, T. A. & Cacioppo, J. T. Variations on a human universal: individual differences in positivity offset and negativity bias. *Cogn. Emot.* **19**, 1–26 (2010).
21. Hamilton, D. L. & Sherman, S. J. Perceiving persons and groups. *Psychol. Rev.* **103**, 336–355 (1996).
22. Asch, S. E. Forming impressions of personality. *J. Abnorm. Soc. Psychol.* **41**, 258–290 (1946).
23. Jones, E. E. & Davis, K. E. in *Advances in Experimental Social Psychology* Vol. 2 (ed. Berkowitz, L.) 219–266 (Academic Press, 1965).
24. Fiske, S. T. & Neuberg, S. L. in *Advances in Experimental Social Psychology* Vol. 23 (ed. Zanna, M. P.) 1–74 (Academic Press, 1990).
25. Smith, E. R. & Zárate, M. A. Exemplar-based model of social judgment. *Psychol. Rev.* **99**, 3–21 (1992).
26. Campbell, D. T. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behav. Sci.* **3**, 14–25 (1958).
27. Lickel, B. et al. Varieties of groups and the perception of group entitativity. *J. Pers. Soc. Psychol.* **78**, 223–246 (2000).
28. Reed, S. K. Pattern recognition and categorization. *Cogn. Psychol.* **3**, 382–407 (1972).
29. Nosofsky, R. M. Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* **115**, 39–57 (1986).
30. Hilton, J. L. & von Hippel, W. Stereotypes. *Annu. Rev. Psychol.* **47**, 237–271 (1996).
31. Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* **117**, 1144–1167 (2010).
32. Zhang, H. & Maloney, L. T. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front. Neurosci.* **6**, 1 (2012).
33. Merten, K. & Nieder, A. Compressed scaling of abstract numerosity representations in adult humans and monkeys. *J. Cogn. Neurosci.* **21**, 333–346 (2008).
34. Dehaene, S., Dehaene-Lambertz, G. & Cohen, L. Abstract representations of numbers in the animal and human brain. *Trends Neurosci.* **21**, 355–361 (1998).
35. Wagenmakers, E. J. et al. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* **25**, 35–57 (2018).
36. Aldous, D. J. in *Exchangeability and Related Topics* 1117 (ed. Hennequin, P. L.) 1–198 (Springer, 1985).
37. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).

38. Tenenbaum, J. B. & Griffiths, T. L. Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* **24**, 629–640 (2001).
39. Fearnhead, P. Particle filters for mixture models with an unknown number of components. *Stat. Comput.* **14**, 11–21 (2004).

Acknowledgements

This work is supported by grant number W911NF-14-1-0101 from the Army Research Office and grant R01DA042065 from the National Institute of Drug Abuse. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors are grateful to S. DuBrow and A. Radulescu for comments on an earlier draft.

Author contributions

Y.S.S. and Y.N. designed the study. Y.S.S. ran the experiment. Y.S.S. and Y.N. analysed the data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01065-0.

Correspondence and requests for materials should be addressed to Y.S.S.

Peer review information Primary Handling Editors: Marike Schiffer; Mary-Elizabeth Sutherland.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Yeon Soon Shin

Last updated by author(s): Jan 5, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Experiments 1A, 1B, 2A-1, 2A-2 were coded in Unity (https://unity3d.com) and conducted online using psiTurk (McDonnell et al., 2012; https://github.com/NYUCCL/psiTurk). Experiments 2A-3, 2A-4 and Experiment 2B and 2C were coded and conducted online using Inquisit (https://www.millisecond.com).
Data analysis	All data analysis was done in R 3.3.2 and R 3.6.1 (https://www.R-project.org).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data generated during the study is available in a public repository at https://osf.io/fdcvw/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	All experiments were quantitative studies with between-subject designs.
Research sample	The research participants included male and female adults over 18 years old who resided in the United States.
Sampling strategy	For Experiments 1A, 1B, 2A-1, 2A-2, 2B, 2C, the sample size was based on pilot data. For Experiment 2A-2 (Experiment 2A-3 replication) , we ran power analysis based on Experiment 2A-1. For Experiment 2A-4, we took a Bayesian approach and collected a minimum of 50 usable participants in each condition and continued data collection until we reached one of three criteria : (1) a Bayes Factor of 10 in favor of H+ (normalized estimate in sparse "stingy" condition < normalized estimate in sparse "generous" condition) and against H0 (no difference in normalized estimates between sparsity conditions), (2) a Bayes Factor of 10 in favor of H0 and against H+, or (3) we reached the maximum number of participants (500 usable participants in each condition). This procedure was pre-registered (https://aspredicted.org/99em9.pdf).
Data collection	Data was collected via Amazon Mechanical Turk, and there was no experimenter present during the data collection.
Timing	Experiment 1A was conducted in January 2017; Experiment 1B was conducted in February 2017; Experiment 2A-1 was conducted in October 2017; Experiemtn 2A-2 was conducted in December 2017; Experiments 2A-3, 2B, 2C were conducted between May and June 2019;.Experiment 2A-4 was conducted between February and April 2020.
Data exclusions	Participants who did not pass the criteria described in Methods were excluded from analyses.
Non-participation	None
Randomization	Participants were assigned to one of the conditions via the randomization code within Unity (Experiments 1A, 1B, 2A-1, 2A-2) and Inquisit (Experiments 2A-3, 2A-4, 2B, 2C), and this randomization was not based on any particular features of the participants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Participants for all studies were recruited on Amazon Mechanical Turk (https://www.mturk.com) using psiTurk ad server (https://psiturk.org; Experiments 1A, 1B, 2A-1, 2A-2) and TurkPrime (https://www.turkprime.com; Experiments 2A-3, 2A-4, 2B, 2C). Participants were compensated for their participation.
Ethics oversight	The Princeton University Institutional Review Board approved the experiment protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.