**Shame orchestrates its outputs rationally**

**Method**

The experiment was preregistered before data collection began: http://aspredicted.org/blind.php?x=kq2xq4

***Participants and procedure***

Standard power analyses to determine sample size of participants could not be conducted, because the correlations are computed over the sample of items (fixed in quantity), not over participants. However, pilot data suggested that 25 participants per scale per country yield adequate power. This number was supplemented to compensate for likely exclusions due to participant inattention; we assumed 30% of data exclusions due to inattention. Thus, we set the total number of participants to be recruited per country to 245—35 participants per scale.

We collected data with Amazon Mechanical Turk from 246 participants (134 females) in the United States and 249 participants (67 females) in India. As per the preregistration protocol, participants were excluded from analyses if they failed to pass an attention check. Seven American participants and 50 Indian participants were excluded from analyses due to failure to pass an attention check, leaving an effective sample of 239 American participants (132 females) (age: M = 39, SD = 12) and 199 Indian participants (56 females) (age: M = 28, SD = 5). The stimuli consist of 27 brief hypothetical scenarios, developed by (Sznycer et al., 2016), in which someone’s acts, traits, or circumstances might lead them to be viewed negatively. The scenarios were designed to elicit reactions in a wide variety of evolutionarily relevant domains, such as mating, parenting, social exchange, aggressive contests, status, skills, and the violation of coordinative norms, and were phrased at a relatively high level of abstraction (e.g., “You stole goods from a shop owned by your neighbor”, rather than, e.g., “You stole jackets and shoes from a sports shop owned by your neighbor”).

Participants were randomly assigned to one of seven between-subjects scales: one *devaluation* scale and six scales relevant to the shame system: *shame feeling*, *hide*, *lie*, *destroy evidence*, *threaten witness*, and *communicate event*. In all seven scales participants rated the same basic set of 27 scenarios. The main difference across scales—the experimental manipulation—was a prompt, displayed immediately before the scenarios, instructing participants to interpret the scenarios in a way that would elicit either devaluation of a target individual or one of the six outputs relevant to the shame system.

In the devaluation scale, the prompt asked participants to imagine that the acts and traits described in the 27 scenarios (e.g., “She is not generous with others,” “She dropped out of school much earlier than others,” “She stole goods from a shop owned by her neighbor”) are true of a target individual: an individual other than the participant who is of the same sex and age as the participant. Then, participants were asked to “indicate [for each scenario] how you would view this person,” with scales ranging from 1 (I'd view her not negatively at all if this were true of her) to 7 (I'd view her very negatively if this were true of her). These ratings provide situation-specific measures of the degree to which members of a given population would socially devalue the individual described in the scenarios.

In the six other scales (shame feeling, hide, lie, destroy evidence, threaten witness, and communicate event), the prompts asked participants to imagine that the acts and traits described in the 27 scenarios are true of the participant herself (e.g., “You are not generous with others,” “You dropped out of school much earlier than others,” “You stole goods from a shop owned by your neighbor”), and to indicate the degree to which they would experience feelings or motivations relevant to the shame system, on scales ranging from 1 (not at all…) to 7 (a lot… / very much…). The prompts asked participants to indicate the following. In the shame feeling scale: how much shame they would feel if those events were true of them. In the hide scale: how much they would want to hide if those events were true of them. In the lie scale: how willing they would be to lie to others by denying that those things are true of them. In the destroy evidence scale: how willing they would be to destroy evidence or clues that might tell others that those things are true of them. In the threaten witness scale: how willing they would be to threaten direct witnesses to prevent them from telling others that those things are true of them. In the communicate event scale: how willing they would be to communicate to others that those things are true of them.

In sum, participants rated, for each of 27 scenarios describing negative acts and traits: (*i*) their devaluation of another individual, if those acts and traits were true of that individual, as well as: (*ii*) their shame feelings about those acts and traits, (*iii*) their willingness to hide, (*iv*) their willingness to lie about those acts and traits, (*v*) their willingness to destroy evidence about those acts and traits, (*vi*) their willingness to threaten witnesses to prevent the spread of information about those acts and traits, and (*vii*) their willingness to communicate those acts and traits to others, if those acts and traits were true of the participants themselves. Each participant rated only one of the seven sets of 27 scenarios.

The scenarios were presented in randomized order within scales. The stimuli were presented in English in the United States and India. Full text of the scale prompts and scenarios used in the United States and India are provided in the Appendix, Tables S1, S2 & S3.

**Results**

***Within-Country Results***. First, we report the results for each country. Descriptive statistics are provided in Tables S2 & S3.

*Do participants within countries agree on how negatively they would view the target individual in each of these scenarios?* Yes. To measure agreement among raters on how socially discrediting the 27 acts and traits are relative to one another, we computed intra-class correlations (ICC) in each country. There was agreement about how socially negative these acts and traits are relative to one another: United States: ICC (2,35) = .98, *P* << .05; India: ICC (2,21) = .38, *P* < .05.

*Emotions: Do participants within countries agree on the degree to which they would experience one of the five outputs of shame if the acts and traits described in the scenarios were true of them ?* In the United States there was widespread agreement about the relative intensity of outputs of shame the 27 situations would elicit: shame feeling: ICC (2,34) = .95; hide: ICC (2,34) = .95; lie: ICC (2,33) = .88; destroy evidence: ICC (2,35) = .86; threaten witness: ICC (2,33) = .88, *Ps* << .05; all of the intra-class correlations in the United remain significant after applying, as per the preregistration, a false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) of P < 0.05. In India there was agreement about the relative intensity of shame feeling: ICC (2,33) = .78, and hide: ICC (2,31) = .65 (*Ps* << .05), but there was no agreement for lie: ICC (2,30) = −.01, destroy evidence: ICC (2,29) = −.10, or threaten witness: ICC (2,27) = .04. In India, the intra-class correlations of shame feeling and hide remain significant at FDR P < 0.05 but the intra-class correlation of devaluation does not; all the other intra-class correlations in India (lie, destroy evidence, threaten witness) are not significant at FDR P < 0.05.

*Does* *the intensity of audience devaluation correlate positively with the intensities of the five outputs of the shame system?* Yes. The intensity of social devaluation participants express (as audiences) if 27 negative acts and traits were true of someone else correlates positively with the intensities of: shame feeling, hide, lie, destroy evidence, and threaten witness, if those 27 negative acts and traits were true of the participant themselves. For each of the 27 scenarios, we calculated the mean ratings of each of the five outputs of shame provided by participants in the shame-relevant scales, and the mean devaluation ratings provided by participants in the devaluation scale. In the United States, for a given scenario, ratings of devaluation predicted ratings of the five outputs of the shame system: shame feeling, hide, lie, destroy evidence, and threaten witness (*r*s = .73–.83, *p*s = .000001–.00002). In India too, for a given scenario, ratings of devaluation predicted ratings of shame outputs: shame feeling, hide, lie, destroy evidence, and threaten witness (*r*s = .39–.74, *p*s = .000009–.045) (see Figs. 1, 2 [panels A, B], and Table 1). Recall that the ratings of devaluation, shame feeling, hide, lie, destroy evidence, and threaten witness originated from different participants. Consequently, these correlations cannot be attributed to participants matching their devaluation ratings to their shame-related ratings. Bayes factors (BFs) were computed to quantify the odds favoring each experimental hypothesis relative to a corresponding null hypothesis. Bayesian correlation analyses were performed using the default priors (stretched beta prior width = 1; JASP 0.10.2). The data files and R scripts for all experiments reported here are available through OSF (link here). In the US, the experimental hypotheses that the intensity of audience devaluation correlates with the intensities of the five outputs of the shame system was more likely than the null that there was no correlation (all BFs10 > 1000). In India, the experimental hypotheses that the intensity of audience devaluation correlates with the intensities of the five shame outputs was more likely than the null that there was no correlation for some but not all of the shame outputs (shame feeling: BF10 = 2718; hide: BF10 = 72.6; lie: BF10 = 1.60; destroy: BF10 = 2.79; threaten: BF10 = 11.4; see Fig #)

*Does the intensity of audience devaluation fail to correlate positively with participants’* *willingness to communicate negative personal characteristics to others?* Yes. In fact, ratings of devaluation correlated *negatively* with ratings of communicate event; in the United States: *r* = −.84, *p* = 10-7; in India: *r* = −.52, *p* = .006. The experimental hypothesis that the intensity of audience devaluation negatively correlates or is uncorrelated with willingness to communicate was more likely than the null that there was a positive correlation (In the US: BF10 = 21.9; in India: BF10 = 14.4).

*Do the intensities of the five outputs of the shame system correlate positively with each other?* Yes. In the United States, ratings of shame feeling, hide, lie, destroy evidence, and threaten witness are positively correlated with each other, with a mean *r* = .80 (SD = .09; minimum *r* = .70; maximum *r* = .97; N *r* values = 10); *P* values = 10−15–.00004; all of these correlations remain significant at FDR P < 0.05. The experimental hypotheses that each pair of the five outputs of the shame system correlate positively with each other were all more likely than the null that there was no correlation (BFs10 range from 628 to 1.44e+13). In India, ratings of shame feeling, hide, lie, destroy evidence, and threaten witness are positively correlated with each other, with a mean *r* = .53 (SD = .12; minimum *r* = .39; maximum *r* = .80; N *r* values = 10); *P* values = 10−6–.045; all of these correlations remain significant at FDR P < 0.05. The experimental hypotheses that pairs of the five outputs of the shame system correlate positively with each other was more likely than the null that there was no correlation for some of the pairs (shame feeling-hide: BF10 = 3.91e+4; shame feeling-destroy: BF10 = 5.16; shame feeling-threaten: BF10 = 13.8; hide-lie: BF10 = 14.7; hide-destroy: BF10 = 5.04; hide-threaten: BF10 = 8.45; lie-threaten: BF10 = 68.4; destroy-threaten: BF10 = 3.21) and indeterminate for others (shame-lie: BF10 = 1.60; lie-destroy: BF10 = 2.74).

-----------------------------------

Insert Figure 1 about here

-----------------------------------

-----------------------------------

Insert Table 1 about here

-----------------------------------

-----------------------------------

Insert Table 2 about here

-----------------------------------

***Between-Country Results***. To test for between-country agreement in devaluation, in the outputs of the shame system, and in the devaluation–shame-outputs links, we computed the extent to which the mean devaluation ratings and the mean shame outputs ratings are correlated across countries.

*Devaluation: Do American and Indian participants agree on how negatively they would view the target individual in each of these scenarios?* Yes. There was between-country agreement on the degree to which a given negative act or trait would elicit devaluation: *r* = .82, *p* = 10-6. The more American participants devalued a target individual for taking a given act or displaying a given trait, the more Indian participants devalued a target individual for taking those acts or displaying those trait. The experimental hypothesis that the intensity of audience devaluation would correlate between the two countries was more likely than the null that there was no correlation (BF10 = 1.08e+5).

*Outputs of shame: Do American and Indian participants agree on the degree to which they would experience feelings and motivations that the shame system mobilizes?* Yes. American and Indian participants agreed about the relative extent to which a given negative act or trait would elicit shame feeling (*r* = .73, *p* = .00002), as well as the motivations to hide (*r* = .62, *p* = .0005), lie (*r* = .42, *p* = .032), and threaten witness (*r* = .54, *p* = .004). There was no cross-country agreement in ratings of destroy evidence, however (*r* = .14, *p* = .48). Further, in 10 of 20 cases, a shame output in one country correlated positively with a different shame output in the other country (e.g., hide in India vs. threaten witness in the United States); mean *r* = .43 (SD = .22; minimum *r* = .15; maximum *r* = .72; N *r* values = 20); *P* values = .00003–.46. The experimental hypotheses that ratings of the same shame output would correlate across the two countries was more likely than the null that there was no correlation for some outputs (shame feeling: BF10 = 1.49e+3; hide: BF10 = 72.7; threaten: BF10 = 13.4), less likely in one case (destroy: BF10 = .303), and indeterminate in another (lie: BF10 = 2.10). Among the 20 comparisons between different shame outputs, the evidence favored the alternative of a correlation relative to the null of no correlation for ten pairs (BFs10 between 3.3 and 1.03e3), was indeterminate for seven pairs (BFs10 between 0.38 and 1.19), and favored the null relative to the alternative for three pairs (BFs10 between 0.310 and 0.326).

*Does intensity of devaluation in one country correlate positively with intensities of outputs of shame in the other country?* In general, yes. Indian participants’ ratings of devaluation correlated positively with American participants’ ratings of shame feeling, hide, lie, destroy evidence, and threaten witness; mean *r* = .71 (SD = .05; minimum *r* = .65; maximum *r* = .76; N *r* values = 5); *P* values = .000004–.0003. In addition, American participants’ ratings of devaluation correlated positively with Indian participants’ ratings of shame feeling (*r* = .64, *p* = .0003), hide (*r* = .54, *p* = .003), and threaten witness (*r* = .44, *p* = .02), but not with Indian participants’ ratings of lie (*r* = .27, *p* = .18) or destroy evidence (*r* = .31, *p* = .11). To put some of this more vividly: One can accurately predict Indians’ willingness to threaten a witness of a negative personal characteristic to prevent the spread of that information from Americans’ devaluation of that characteristic, or Americans’ willingness to lie and deny that a negative characteristic is true of them from Indians’ devaluation of that characteristic. All the cross-country correlations (devaluation vs. devaluation, output vs. same output, output vs. different output) that are significant at P < .05 also remain significant at FDR P < 0.05, except for the US lie vs. India lie correlation. The experimental hypotheses that the intensity of audience devaluation in the US correlates with the intensities of the five outputs of the shame system in India was more likely than the null that there was no correlation for some of the shame outputs (shame feeling: BF10 = 102; hide: BF10 = 14.5) and indeterminate for others (lie: BF10 = 0.55; destroy: BF10 = 0.79; threaten: BF10 = 2.69). The experimental hypotheses that the intensity of audience devaluation in India correlates with the intensities of the five shame outputs in the US was more likely than the null that there was no correlation (all BFs10 > 137).

**RVC**