LETTERS
https://doi.org/10.1038/s41562-018-0425-1
nature human behaviour

Beliefs about bad people are volatile

Jenifer Z. Siegel[1], Christoph Mathys[2,3,4], Robb B. Rutledge[3,5] and Molly J. Crockett[1,6*]

People form moral impressions rapidly, effortlessly and from a remarkably young age[1–5]. Putatively 'bad' agents command more attention and are identified more quickly and accurately than benign or friendly agents[5–12]. Such vigilance is adaptive, but can also be costly in environments where people sometimes make mistakes, because incorrectly attributing bad character to good people damages existing relationships and discourages forming new relationships[13–16]. The ability to accurately infer the moral character of others is critical for healthy social functioning, but the computational processes that support this ability are not well understood. Here, we show that moral inference is explained by an asymmetric Bayesian updating mechanism in which beliefs about the morality of bad agents are more uncertain (and therefore more volatile) than beliefs about the morality of good agents. This asymmetry seems to be a property of learning about immoral agents in general, as we also find greater uncertainty for beliefs about the non-moral traits of bad agents. Our model and data reveal a cognitive mechanism that permits flexible updating of beliefs about potentially threatening others, a mechanism that could facilitate forgiveness when initial bad impressions turn out to be inaccurate. Our findings suggest that negative moral impressions destabilize beliefs about others, promoting cognitive flexibility in the service of cooperative but cautious behaviour.

Signs of bad character capture attention[9–12] because people are strongly motivated to avoid being exploited by others[16,17]. However, erroneously inferring bad character can lead people to prematurely terminate valuable relationships and thereby miss out on the potential benefits of future cooperative interactions[13–16]. Thus, successfully navigating social life requires strategies for maintaining social relationships even when others behave inconsistently and sometimes commit immoral acts.

One possible strategy is to respond to defection with probabilistic cooperation[18]. Evolutionary models show that such 'generous' strategies outcompete strategies that summarily end cooperative relationships in the face of a single betrayal[19,20]. Generous strategies are also observed in humans who play repeated prisoner's dilemmas in which the intended actions of others are implemented with noise[20]. Although evolutionary and economic models provide descriptive accounts of these behaviours, the cognitive mechanisms that enable them are not well understood. In particular, the computational processes that support adaptive moral inference in humans are unknown.

We propose that when people form beliefs about the moral character of others, their impressions about bad agents are more uncertain than their impressions about good agents. This makes impressions about bad agents more amenable to Bayesian updating, by which belief updates are proportional to the uncertainty of beliefs in accordance with Bayes' rule[21]. Our hypothesis is based on evidence that threatening social stimuli are arousing[22] and that arousal increases belief uncertainty in non-social perceptual learning[23]. This evidence suggests that threatening social stimuli (such as agents with inferred bad character) might induce belief uncertainty. Our proposal provides a possible solution for maintaining social relationships when others sometimes act immorally by enabling negative impressions to be more easily revised: if beliefs about putatively 'bad' agents are volatile, such beliefs could be readily updated if the initial impression turned out to be mistaken.

At first glance, our hypothesis may seem inconsistent with decades of research in social psychology, much of which has examined impression formation from narrative descriptions of extreme and rare behaviours, such as theft or violence. This work provides evidence for a negativity bias in impression formation, in which people update their moral impressions to a greater degree from negative relative to positive information[9,12,24]. The primary explanation for this valence asymmetry is that it reflects a differential diagnosticity of immoral versus moral behaviours: bad people often behave morally, but good people rarely behave immorally[9]. Indeed, recent work has suggested that valence asymmetries in impression updating can be explained by the perceptions of how rare immoral behaviours are, relative to moral behaviours[25]. This leaves open the question of whether people actually learn differently about agents who are inferred to be more versus less moral when their actions are equally diagnostic of their underlying character. This is the central question that we addressed in the current studies. We focused on moral inference from behaviours that are not extreme or definitive of character. Such behaviours comprise the vast majority of our daily social interactions: we most often judge others based on behaviours that are nasty or nice, not evil or saintly. Inferring character from minor slights or small favours is considerably more difficult than doing so from criminal deeds or heroic actions, but our success as a social species suggests that we are nevertheless able to do this effectively.

We developed an approach to investigate the computational basis of moral inference and its temporal dynamics. Participants predicted and observed the choices of two 'agents' who repeatedly decided whether to inflict painful electric shocks on another person in a different room in exchange for money (Fig. 1a). We generated agent behaviour using a model that accurately captures typical preferences in this choice setting[26,27]. The model includes a 'harm aversion' parameter, κ, which quantifies the subjective cost of harming the victim as an exchange rate between money and pain and ranges from 0 (profit maximizing) to 1 (pain minimizing) (Supplementary Fig. 1). Because ethical systems universally judge harming others for personal gain as morally wrong[28], we operationalized moral

[1]Department of Experimental Psychology, University of Oxford, Oxford, UK. [2]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy. [3]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK. [4]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland. [5]Wellcome Trust Centre for Neuroimaging, University College London, London, UK. [6]Department of Psychology, Yale University, New Haven, CT, USA. *e-mail: mj.crockett@yale.edu

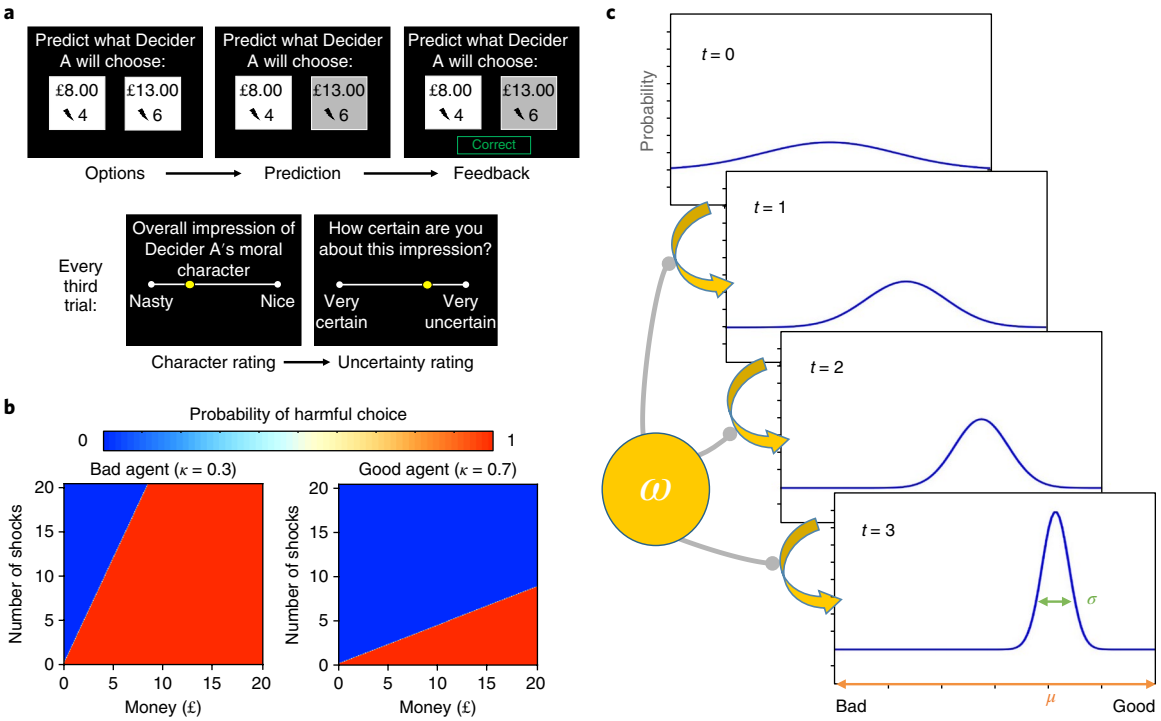

Fig. 1 | Learning task and model. a, Participants predicted sequences of choices for two agents: 'Decider A' and 'Decider B'. On each trial, the agent chose between a more harmful (more shocks inflicted on another person for more money) and a less harmful (fewer shocks and money) option. After every third trial, participants rated their impression of the moral character of the agent. In studies 2–5, participants also rated the uncertainty of their impression. For each study, the learning task used local currency (GBP for study 1, USD for studies 2–8). **b**, Heat maps summarize the bad agent's ($\kappa = 0.3$) or the good agent's ($\kappa = 0.7$) probability of choosing the more harmful option as a function of money gained and shocks delivered. **c**, Model schematic for learning about a good agent. Beliefs about moral character are represented by probability distributions. The mean of the distribution (μ) describes the current belief about the agent after trial t, and the variance of the distribution (σ) describes the current uncertainty on that belief. Beliefs evolve over time as a Gaussian random walk whose step-size is governed by ω, a participant-specific parameter that captures individual differences in belief volatility.

character as harm aversion in our paradigm. The two agents differed substantially in their harm aversion, with the 'good' agent requiring more compensation per shock to inflict pain on others than the 'bad' agent (bad: $\kappa = 0.3$ or £0.43 per shock; good: $\kappa = 0.7$ or £2.40 per shock; Fig. 1b). The preferences of the good and bad agents were symmetric around the participants' expectations of 'average' behaviour, which was not significantly different from $\kappa = 0.5$ (see Supplementary Results, study 8 for details).

On each trial, participants predicted the choice made by the agent and received immediate feedback on their accuracy. After every third trial, participants rated their subjective impressions of the agent's morality on a scale ranging from 'nasty' to 'nice' and rated how uncertain they were about their impression on a scale ranging from 'very certain' to 'very uncertain'.

We modelled participants' predictions for each agent separately with a Bayesian learning model[21] that generated a trial-wise sequence of belief estimates about each agent's character (that is, the exchange rate between money and pain, latent variable, μ); a trial-wise sequence of uncertainties on those beliefs (latent variable, σ); and a global estimate of belief volatility (parameter, ω) that describes the rate at which beliefs evolve over time (Fig. 1c). We use the term volatility here to be consistent with previous work using a similar model[21,29,30] and because the volatility parameter in our model captures how rapidly beliefs change. Belief volatility is set in log space and is monotonically related to belief uncertainty (that is, more uncertain beliefs are more volatile[21]; for example, a change in ω from −3.5 to −4.0 corresponds to a 20% decrease in the average variance of posterior beliefs, σ). We report ω here; see Supplementary Information and Supplementary Table 1 for results for trial-wise uncertainty σ.

Formal model comparisons indicated that our model outperformed simpler Rescorla–Wagner models that do not account for uncertainty in beliefs (see Supplementary Methods, study 1, and Supplementary Table 2 for details). To test our hypothesis that character ratings and model parameter estimates μ and ω will differ between good and bad agents, we compared them using two-tailed non-parametric statistical tests that do not make assumptions about underlying distributions of the character ratings and parameter estimates. We report means and the standard error of the mean (s.e.m.) as mean ± s.e.m.

Our approach extends previous methods that were used to probe impression formation in several ways. First, because our paradigm used a computational model of moral preferences rather than narrative descriptions of behaviours (as in past social psychology research), we were able to very tightly control how informative the behaviours of agents were with regard to their underlying preferences. We precisely matched the trial sequences with respect to how much information was provided about each agent's character over the course of learning (see Supplementary Methods, study 1 for details). In this way, we ensured that the statistics of the environment did not advantage learning about either the good or the bad agent, and this symmetry was confirmed by the fact that an ideal Bayesian observer learned identically about the good and the bad agents (Supplementary Table 3). Because of this design feature, we can confidently infer that the belief asymmetries that we observed in our studies were not due to asymmetries in the information that we provided to participants (in contrast to past studies using narrative descriptions of behaviours, in which moral information was evaluated as less diagnostic than immoral information[25]).

Second, in contrast to past work, which focused on descriptive measures over relatively few trials, our methods allowed us to measure the dynamics of impression formation over time. Finally, our paradigm allowed us to measure the uncertainty and volatility of people's impressions in addition to the valence of those impressions, which has been the primary focus of past work. By doing so, we are able to bridge our investigation of moral inference with foundational work on perceptual and reinforcement learning[21,23] and show that similar computational principles underlie learning across these diverse domains[29,31,32].

In an initial study (study 1), we measured moral inference in 38 participants in the laboratory. Our model fit the predictions of participants well, explaining behaviour with 87% accuracy on average (Supplementary Table 4). Participants accurately inferred that the bad agent was less moral than the good agent, as evident in subjective character ratings (Wilcoxon signed-rank test, final character rating: bad = 42.663 ± 4.021; good = 78.831 ± 2.869; $P < 0.001$; Supplementary Table 5) and the model's estimates of beliefs (final μ: bad = 0.332 ± 0.004; good = 0.681 ± 0.004; $P < 0.001$; Supplementary Table 1). Notably, subjective character ratings and modelled beliefs followed different dynamics: although subjective character ratings rapidly distinguished between the agents, beliefs in the model integrated over more information and updated gradually over a longer timescale. Participants formed subjective impressions of the agents' character well before they developed precise beliefs about the agents' moral preferences.

As predicted, beliefs about the morality of bad agents were more volatile than beliefs about good agents (ω: bad = −3.779 ± 0.102; good = −4.212 ± 0.104; $P = 0.001$; Supplementary Table 1). Participants were consciously aware of this asymmetry, as they rated their impressions of the bad agent as more uncertain than their impressions of the good agent (mean uncertainty rating: bad = 28.623 ± 2.428; good = 20.612 ± 2.367; $P < 0.001$; Supplementary Tables 5 and 6). We found that the difference in the volatility of beliefs about the moral character of the good and the bad agents was significantly larger for participants than an ideal Bayesian observer ($\Delta\omega$: participants = 0.433 ± 0.121; Bayesian = 0.015 ± 0.011; $P < 0.001$; Supplementary Table 3). Thus, the asymmetry that we observe in moral learning cannot be due to the statistics of the environment.

In a second study ($N = 163$), we sought to replicate our findings in a larger and more diverse sample and to test whether the participants' moral impressions of the two agents affected their social behaviour by inviting them to entrust money to each agent in a one-shot trust game after learning about both agents (see Supplementary Methods, study 2 for details). Replicating our previous results, participants accurately inferred that the bad agent was less moral than the good agent (final μ: bad = 0.301 ± 0.004; good = 0.707 ± 0.003, $P < 0.001$; character rating: bad = 42.227 ± 1.962; good = 80.706 ± 1.444, $P < 0.001$; Fig. 2a and Supplementary Tables 1 and 5). Participants also entrusted the good agent with twice as much money as the bad agent, demonstrating that these moral impressions are relevant to social economic decisions (amount entrusted: bad = 3.36 ± 0.30; good = 7.15 ± 0.29, $P < 0.001$; Fig. 2b and Supplementary Table 7). As in the first study, beliefs about the moral character of the bad agent were more uncertain and volatile than beliefs about the good agent (mean uncertainty rating: bad = 33.078 ± 1.330; good = 24.078 ± 1.371, $P < 0.001$; ω: bad = −3.411 ± 0.051; good = −3.877 ± 0.051, $P < 0.001$; Fig. 2c,d and Supplementary Tables 1 and 5). Our model predicts that there would be a larger trial-wise updating of character ratings for the bad agent than for the good agent. This was confirmed in a model-free analysis in which we compared the magnitude of changes in trial-to-trial ratings between the good and the bad agents (see Supplementary Results, study 2, and Supplementary Table 8).

In a third study ($N = 135$), we increased the stochasticity of agent choices to test whether the differences that we observed for learning about bad compared to good agents are robust to noisy environments. We replicated all of the findings from studies 1 and 2 (see Supplementary Results, study 3, and Supplementary Tables 1 and 5), including the key result that beliefs about the moral character of bad agents are more volatile than those about good agents (ω: bad = −3.468 ± 0.042; good = −3.974 ± 0.043, $P < 0.001$). Furthermore, to ensure that our findings in studies 1–3 were not an artefact of the scale that participants used to rate the morality of agents (ranging from nasty to nice), we replicated all findings in an additional study using an alternative scale (ranging from bad to good; see Supplementary Methods and Supplementary Results, study 7, and Supplementary Tables 1 and 5).

One possible explanation for why people form more uncertain beliefs about the moral character of bad than of good agents is a strong prior expectation that people will behave morally[33,34], thus rendering the behaviour of the bad agent more surprising. To investigate this possibility, we asked a separate group of participants to predict, in the context of decisions to profit from the pain of others, how 'most people' would choose (see Supplementary Methods, study 8). This allowed us to estimate participants' expected level of harm aversion (κ) within the context of our task. No feedback was provided to participants during the task, but to motivate accurate predictions, participants received a financial bonus for each trial in which they successfully predicted the majority response. We found no evidence that people expected others to behave more like the good agent. In fact, we cannot reject the hypothesis that the expected κ came from a distribution with a median ($\kappa = 0.5$) equidistant from that of the good and bad agents (mean expected $\kappa = 0.445$, one sample signed-rank test, $P = 0.178$). These results suggest that our observation of more uncertain beliefs about the morality of bad agents cannot be fully attributed to prior beliefs about the morality of others.

In addition, we measured the beliefs of participants about the character of the agents before starting the learning task. If the asymmetry in learning is explained by prior expectations that people will behave morally, then this asymmetry should be larger when people expect others to be nicer. However, we found no relationship between prior expectations about the agents' moral character and between-agent differences in our key dependent measures, such as ω and subjective uncertainty ratings (see Supplementary Results, study 2, Supplementary Table 9 and Supplementary Fig. 2). We also did not find a relationship between learning asymmetries and self-reports of generalized trust in others (see Supplementary Results, study 8, and Supplementary Fig. 3).

In a fourth study ($N = 220$), we examined whether the asymmetry in learning about bad compared to good agents extends to learning about a trait unrelated to morality. If the asymmetry is specific to moral impressions, then it should be larger when learning about moral character than when learning about a non-moral trait, such as competence. To test this, we randomized participants into either a morality condition ($N = 109$; Fig. 3a) or a competence condition ($N = 111$; Fig. 3b). In the morality condition, participants predicted the moral choices of a bad agent and a good agent as before. In the competence condition, participants predicted the basketball performance (the number of points scored per minute) of a low-skill agent and a high-skill agent. Crucially, task parameters were precisely matched across conditions so that an ideal Bayesian observer would learn identically in all cases, permitting direct comparison of model estimates and subjective ratings. We chose to examine learning about basketball ability rather than other traits related to competence, such as intelligence or social ability, because previous work has shown that the latter are not independent of impressions of moral character[35]. By contrast, we expected inferences about basketball ability to be independent from inferences about moral character. Pilot testing supported this claim (see Supplementary Methods, study 4). Thus, our design allowed us to directly test the specificity

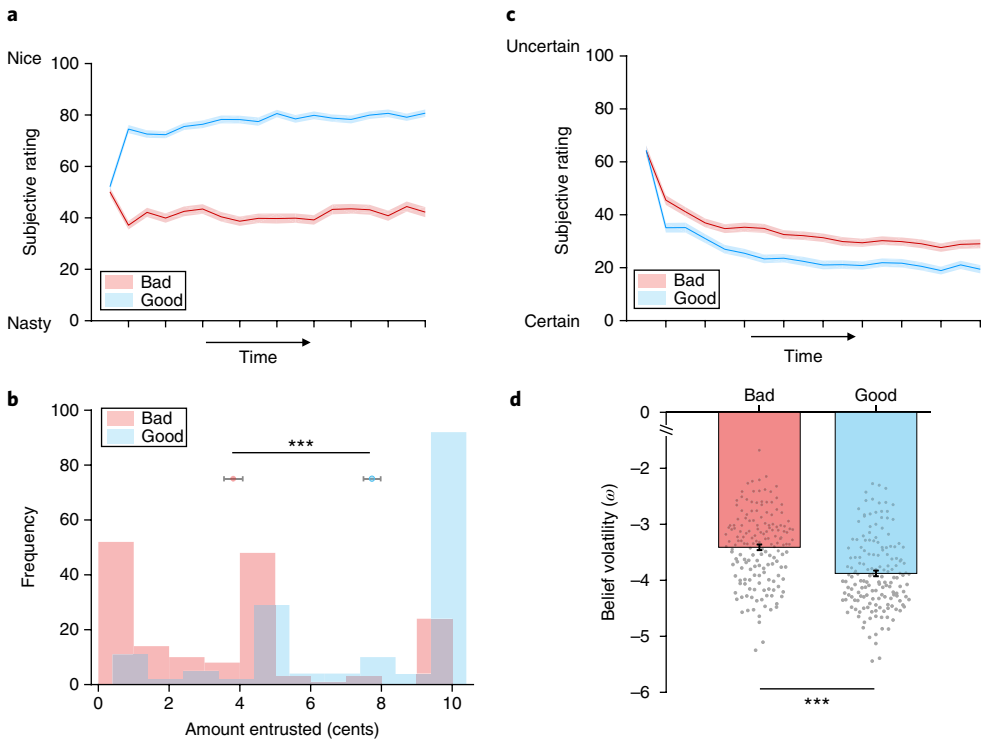

Fig. 2 | Asymmetry in moral impression formation, study 2. a, The trajectory of subjective character ratings over time in study 2, averaged across participants. **b**, In a one-shot trust game, participants entrusted the good agent with twice as much money as they did the bad agent ($Z = 6.831$, $P < 0.001$). **c**, The trajectory of subjective uncertainty ratings over time, averaged across participants. Participants reported greater uncertainty about bad agents. **d**, The volatility of beliefs (ω in the model) was higher for the moral character of the bad agent than for the good agent ($Z = -3.411$, $P < 0.001$). In all panels, the error bars and the shaded bounds in the trajectories represent the s.e.m and $N = 163$. ***$P < 0.001$.

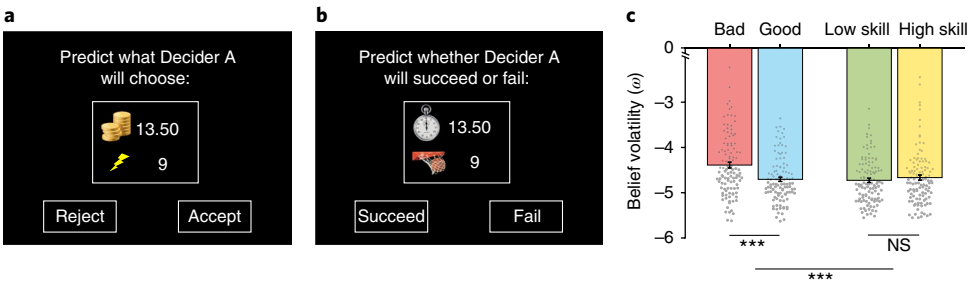

Fig. 3 | Forming impressions of morality versus competence, study 3. a, In the morality condition, participants ($N = 109$) predicted whether the agent would deliver a certain number of shocks for a specified profit. **b**, In the competence condition, participants ($N = 111$) predicted whether the agent would succeed in scoring a certain number of basketball points within a specified amount of time. **c**, The interaction between the agent (bad/low skill versus good/high skill) and the condition (morality versus competence) for the volatility of beliefs (ω in the model; $Z = 4.219$, $P < 0.001$). The error bars represent the s.e.m. ***$P < 0.001$; NS, not significant.

of our observed effect for moral inference because it is unlikely that participants would form moral impressions from observations of a basketball performance alone.

As predicted, between-agent differences in uncertainty ratings and belief volatility were significantly larger in the morality condition than in the competence condition (rank sum test, difference in the mean uncertainty rating: morality $= 4.870 \pm 1.467$, competence $= -2.778 \pm 1.161$, $P < 0.001$; difference in ω: morality $= 0.316 \pm 0.069$, competence $= -0.060 \pm 0.069$, $P < 0.001$). Participants' beliefs about bad agents were more uncertain and volatile than beliefs about good agents (mean uncertainty rating: bad $= 29.335 \pm 1.598$; good $= 24.166 \pm 1.607$, $P < 0.001$; ω: bad $= -4.390 \pm 0.064$; good $= -4.714 \pm 0.048$, $P < 0.001$; Supplementary Tables 1 and 5), but there was no difference in the volatility of beliefs about low-skill and high-skill agents (mean uncertainty rating: low skill $= 18.457 \pm 1.227$; high skill $= 20.653 \pm 1.274$, $P = 0.076$; ω: low skill $= -4.726 \pm 0.047$; high skill $= -4.655 \pm 0.057$, $P = 0.566$; Fig. 3c).

Previous work has shown that bad behaviours carry more weight than good behaviours in moral impression formation[9,10,12,25]. In our studies, the bad agent by definition makes more immoral choices than the good agent, and so we cannot be sure that the observed asymmetry in learning is driven by inferences about the moral character of the good and bad agents rather than responses to the choices that the good and bad agents make. We predicted that the threatening nature of bad agents would increase the uncertainty

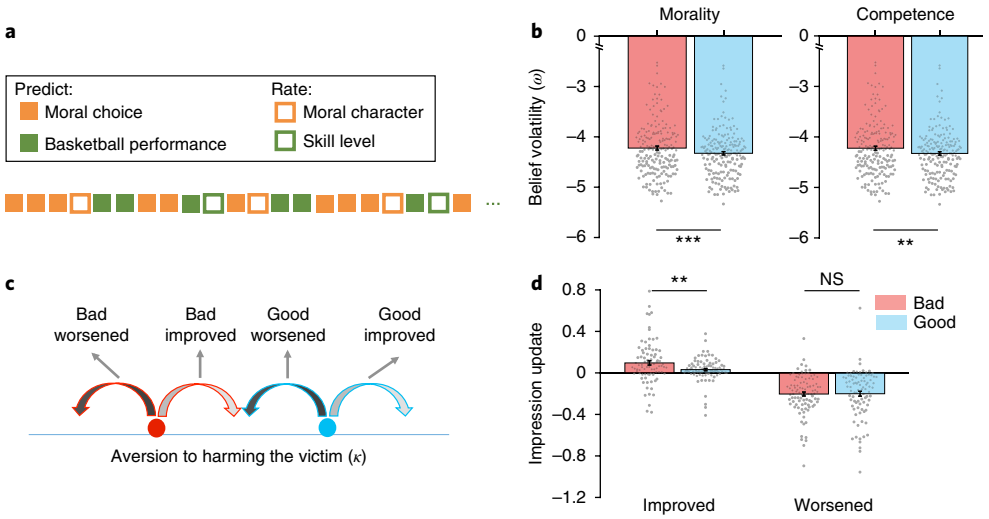

Fig. 4 | Inferences about moral character affect learning about non-moral traits and impression updating. a, In study 5, participants experienced trial sequences with interleaved morality (Fig. 3a) and competence trials (Fig. 3b). Participants rated their impressions of and uncertainty about the moral character and skill level of agents after every third morality and competence trial, respectively. **b**, Comparison of the volatility of beliefs about the morality (left) and competence (right) of the good and bad agents in study 5; $N = 189$ (morality: $Z = 5.079$, $P < 0.001$; competence: $Z = 3.030$, $P = 0.002$). **c**, In study 6, participants were randomized to learn about a bad agent ($\kappa = 0.3$) or a good agent ($\kappa = 0.7$) whose moral character either improved ($\kappa + 0.2$) or worsened ($\kappa - 0.2$). **d**, In study 6, participants more strongly updated their impressions of bad agents than of good agents when moral character improved but not when it worsened ($N = 364$; $F(1,360) = 6.803$, $P = 0.009$; chi-squared = 57.227, $P < 0.001$). In **b** and **d**, the error bars represent the s.e.m. $^{**}P < 0.01$; $^{***}P < 0.001$; NS, not significant.

and volatility of beliefs, thereby destabilizing beliefs in a non-specific manner. This prediction is consistent with past literature showing that task-irrelevant threatening cues increase attention and information processing[36]. If inferring bad moral character exerts a global effect on social impression formation, then beliefs about other traits, such as competence, should also be more volatile for agents that are believed to be immoral. We tested this hypothesis in a fifth study where participants ($N = 189$) simultaneously inferred the morality and competence of a good agent and a bad agent with similar levels of competence (Fig. 4a). Supporting our hypothesis, participants formed more volatile beliefs about the bad agent's morality and competence relative to the good agent (Fig. 4b; moral ω: bad $= -4.116 \pm 0.046$; good $= -4.428 \pm 0.039$, $P < 0.001$; competence ω: bad $= -4.224 \pm 0.039$; good $= -4.327 \pm 0.034$, $P = 0.002$; Supplementary Table 1). Moral impressions also affected participants' own conscious awareness of the uncertainty of their beliefs: participants expressed greater uncertainty in their impressions of the bad agent's morality and competence (moral uncertainty rating: bad $= 27.880 \pm 1.019$; good $= 24.209 \pm 1.027$, $P < 0.001$; competence uncertainty rating: bad $= 28.875 \pm 1.995$; good $= 27.277 \pm 1.992$, $P = 0.020$; Supplementary Table 5).

Our results suggest that impressions of bad agents are more rapidly updated in the face of new evidence than impressions of good agents. We hypothesized that this mechanism would enable people to rapidly revise an initially bad impression of another person if their behaviour subsequently improves. To test this, in a final preregistered study, we examined how people update their impressions of bad and good agents following a shift in their behaviour (https://osf.io/5s23d/). Participants ($N = 364$) were randomized to learn about an agent who was initially either bad or good, but then began to make choices that were consistently either more or less moral than previously (Fig. 4c and Supplementary Methods, study 6). In this study, we explicitly set participants' prior beliefs at $\kappa = 0.5$ by instructing them that "on average, people require \$1 per additional shock to the 'victim'". Because beliefs about bad agents are more volatile, we predicted that participants would more strongly update their impressions of bad agents than of good agents. We tested our hypothesis by comparing, for bad versus good agents, the extent to which participants updated their impressions, defined as the difference between character ratings before versus after the agents' preferences shifted. Because this study investigated how people update character impressions in response to contradictory information, the design most closely resembled those implemented in past social psychology studies[24,25].

As predicted, we observed a main effect of agent on impression updating, in which participants updated their character ratings more for bad agents than for good agents (rank sum, update: bad $= 18.951 \pm 1.245$, good $= 14.928 \pm 1.316$, $P < 0.001$). There was also a main effect of shift direction: updating was greater when morality worsened than when it improved (rank sum, update: worsen $= 22.083 \pm 1.389$; improve $= 11.468 \pm 1.010$, $P < 0.001$). This is consistent with past reports of negativity bias in impression formation[9–11], in which people show stronger impression updating in response to inconsistent immoral behaviours relative to moral behaviours. The main effects were qualified by an interaction between agent and shift direction (Kruskal–Wallis, $P < 0.001$), in which asymmetric updating was more pronounced when morality improved than when morality worsened (Fig. 4d). At first glance, this interaction may seem surprising because our model only predicts a main effect of agent and does not differentiate between positive and negative updating. However, our theoretical framework proposes that people form more volatile beliefs about putatively bad agents due to an adaptive mechanism whereby potentially threatening cues increase attention and learning. Thus, when a 'good' agent's behaviour suddenly worsens, participants may infer a potential threat, prompting their beliefs about the agent to become more uncertain and amenable to rapid updating. Consistent with these predictions, the degree of impression updating tracked with participants' change in subjective ratings of uncertainty before versus after the agents' behaviour shifted (Spearman's ρ, $P = 0.006$; Supplementary Results, study 6).

We have demonstrated in six studies that bad moral impressions are more volatile than good moral impressions. Furthermore, inferring bad character destabilized overall social impression formation, spilling over into learning about a non-moral trait. When moral behaviour improved, impressions were updated faster for putatively

bad agents than for good agents. Thus, the volatility of bad moral impressions may facilitate forgiveness by enabling initially bad impressions to be rapidly updated if behaviour improves.

Despite the robustness of our findings, our paradigm has an important limitation: accepting money in exchange for shocks that are painful but not dangerous is a relatively mild moral transgression. Mild transgressions represent the vast majority of transgressions that will be personally experienced by most individuals, and thus, the mechanisms that we identify here may explain everyday changes in beliefs about the moral character of others. However, it is unclear how these results will generalize to learning about more extreme transgressions, such as assault, rape or murder.

By simultaneously measuring implicit beliefs about moral preferences that guided behavioural predictions, as well as explicit subjective impressions of moral character, our paradigm revealed that beliefs and subjective impressions followed different dynamics. Consistent with previous work[3], participants rapidly formed subjective impressions about moral character after just a few trials. Meanwhile, beliefs in the model integrated over more information and updated gradually over a longer timescale, reflecting the fact that the model estimates the precise exchange rate between money and pain, which cannot be inferred from a single trial. These different dynamics highlight how subjective moral impressions are often based on highly impoverished information; in our studies, participants were readily willing to judge the character of others well before they formed precise beliefs about their moral preferences. Why and how people jump to conclusions about the character of others despite lacking sufficient information to accurately predict their behaviour remain important questions for further study.

Although theoretical models of person perception have claimed the independence of trait dimensions (namely, warmth and competence)[1], other evidence suggests that judgements across trait dimensions may share a positive relationship[35,37]. Our work lends further support to the possibility that the cognitive processing of different traits belonging to the same individual are related and offers tools for addressing this question. By considering uncertainty of beliefs in addition to valence, future work may shed new light on how the mechanisms supporting different dimensions of person perception relate to one another.

Overall, our findings are consistent with research identifying a negativity bias in impression formation, in which bad behaviours command more attention than good behaviours[9–12] and research showing that uncertain attitudes are susceptible to change[38]. Taken together, our results extend this literature to show that, when considered within a Bayesian learning framework, a negativity bias naturally makes impressions more volatile, in which impressions about bad agents are more rapidly updated than impressions about good agents. We suggest that, by destabilizing the overall impressions of others, the learning mechanism described here promotes cognitive flexibility in the service of building richer models of potentially threatening others. This mechanism provides an algorithmic solution to the problem of moral inference in a world where people sometimes make mistakes and helps to resolve the paradox of how people can forgive despite the potency of negative information in judging the moral character of others.

Methods

The research was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford, UK (study 1, MSD-IDREC-C1-2015-001; studies 2–8, MSD-IDREC-C1-2015-098). All participants provided informed consent and were compensated for their time. For each study, the learning task used local currency (GBP for study 1, USD for studies 2–8).

For study 1, 39 participants were recruited from the University of Oxford participant pool. One participant was excluded from the analysis as their performance was below chance in the learning task (<50% accuracy). For study 2, 253 participants were recruited from Amazon Mechanical Turk (AMT) and 87 were excluded for below-chance performance. For study 3, 162 participants were recruited from AMT and 27 were excluded for below-chance performance. All participants from studies 1–3 completed a learning task that involved predicting sequences of moral decisions made by two agents who differed in their moral character (Fig. 1a). Throughout the task, participants indicated their impression of the agents' moral characters (on a scale from nasty to nice) and how certain they were about this impression. To motivate accurate performance, participants in studies 2 and 3 were instructed that they would later decide whether to trust each of the agents in a one-shot trust game that could earn them additional money.

For study 4, 280 participants were recruited from AMT and randomly assigned to complete either a moral learning task or a competence learning task (Fig. 3a,b). In the morality condition, participants predicted the moral choices of two agents who differed in moral character. In the competence condition, participants predicted the basketball performance of two agents who differed in skill level. For study 4, 31 participants from the morality condition and 29 participants from the competence condition were excluded for below-chance performance. To motivate accurate predictions, participants received a monetary bonus for high accuracy.

For study 5, 259 participants were recruited from AMT and 70 were excluded for below-chance performance. Participants completed a learning task in which they simultaneously predicted and observed the moral choices and basketball performance of two agents who substantially differed in their moral character (one bad agent and one good agent) but were equally competent at basketball (Fig. 4a). As in study 4, participants received a monetary bonus for high accuracy.

For study 6, 408 participants were recruited from AMT and 44 were excluded for below-chance performance. Participants were randomized to learn about an agent who was initially either bad or good, but then began to make choices that were consistently either more or less moral than previously. Together, this resulted in four conditions, manipulated between participants: (1) bad agent becomes more moral, (2) bad agent becomes less moral, (3) good agent becomes more moral and (4) good agent becomes less moral. Prior to observing any of the agents' choices, participants were explicitly instructed on how the average person behaved in the task. As in studies 4 and 5, participants received a monetary bonus for high accuracy.

In an additional study (referred to as study 7 in Supplementary Methods and Supplementary Results), 125 participants were recruited from AMT and 9 were excluded for below-chance performance. Study 7 was identical to studies 1 and 2; however, instead of rating the moral character of the agents on a scale ranging from nasty to nice, participants rated the moral character of the agents on a scale ranging from bad to good. To motivate accurate predictions, participants received a monetary bonus for high accuracy.

In a second additional study (referred to as study 8 in Supplementary Methods and Supplementary Results), 30 participants were recruited from AMT to predict, in the context of decisions to profit from the pain of others, how 'most people' choose. No feedback was provided to participants during the task, but each trial that participants correctly predicted the majority response was awarded as a bonus payment upon completion of the study.

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

Code availability. All relevant MATLAB codes are available from the corresponding author upon request.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Received: 13 September 2017; Accepted: 30 July 2018;
Published online: 17 September 2018

References

1. Fiske, S. T., Cuddy, A. J. C. & Glick, P. Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
2. Uleman, J. S. & Kressel, L. M. in *Oxford Handbook of Social Cognition* (ed. Calston, D. E.) 53–73 (Oxford Univ. Press, Oxford, 2013).
3. Todorov, A., Pakrashi, M. & Oosterhof, N. N. Evaluating faces on trustworthiness after minimal time exposure. *Soc. Cogn.* **27**, 813–833 (2009).
4. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* **19**, 1508–1519 (2007).
5. Kiley Hamlin, J., Wynn, K. & Bloom, P. Three-month-olds show a negativity bias in their social evaluations. *Dev. Sci.* **13**, 923–929 (2010).
6. Schupp, H. T. et al. The facilitated processing of threatening faces: an ERP analysis. *Emotion* **4**, 189–200 (2004).
7. Öhman, A., Lundqvist, D. & Esteves, F. The face in the crowd revisited: a threat advantage with schematic stimuli. *J. Pers. Soc. Psychol.* **80**, 381–396 (2001).
8. Vanneste, S., Verplaetse, J., Hiel, A. V. & Braeckman, J. Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evol. Hum. Behav.* **28**, 272–276 (2007).

9. Skowronski, J. J. & Carlston, D. E. Negativity and extremity biases in impression formation: a review of explanations. *Psychol. Bull.* **105**, 131–142 (1989).
10. Fiske, S. T. Attention and weight in person perception: the impact of negative and extreme behavior. *J. Pers. Soc. Psychol.* **38**, 889–906 (1980).
11. Pratto, F. & John, O. P. Automatic vigilance: the attention-grabbing power of approach- and avoidance-related social information. *J. Pers. Soc. Psychol.* **61**, 380–391 (1991).
12. Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D. Bad is stronger than good. *Rev. Gen. Psychol.* **5**, 323–370 (2001).
13. McCullough, M. E. *Beyond Revenge: The Evolution of the Forgiveness Instinct* (John Wiley & Sons, San Francisco, CA, 2008).
14. Axelrod, R. M. *The Evolution of Cooperation* (Basic Books, New York, NY, 2006).
15. Molander, P. The optimal level of generosity in a selfish, uncertain environment. *J. Conflict Resolut.* **29**, 611–618 (1985).
16. Johnson, D. D. P., Blumstein, D. T., Fowler, J. H. & Haselton, M. G. The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. *Trends Ecol. Evol.* **28**, 474–481 (2013).
17. Cosmides, L. & Tooby, J. in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (eds Barkow, J. H. et al.) 163–228 (Oxford Univ. Press, Oxford, 1992).
18. Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).
19. Wu, J. & Axelrod, R. How to cope with noise in the iterated prisoner's dilemma. *J. Conflict Resolut.* **39**, 183–189 (1995).
20. Fudenberg, D., Rand, D. G. & Dreber, A. Slow to anger and fast to forgive: cooperation in an uncertain world. *Am. Econ. Rev.* **102**, 720–749 (2012).
21. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5**, 39 (2011).
22. Ohman, A. Face the beast and fear the face: animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology* **23**, 123–145 (1986).
23. Nassar, M. R. et al. Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.* **15**, 1040–1046 (2012).
24. Reeder, G. D. & Coovert, M. D. Revising an impression of morality. *Soc. Cogn.* **4**, 1–17 (1986).
25. Mende-Siedlecki, P., Baron, S. G. & Todorov, A. Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *J. Neurosci.* **33**, 19406–19415 (2013).
26. Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P. & Dolan, R. J. Harm to others outweighs harm to self in moral decision making. *Proc. Natl Acad. Sci. USA* **111**, 17320–17325 (2014).
27. Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P. & Dolan, R. J. Moral transgressions corrupt neural representations of value. *Nat. Neurosci.* **20**, 879–885 (2017).
28. Gert, B. *Common Morality: Deciding What to Do* (Oxford Univ. Press, Oxford, 2004).
29. Diaconescu, A. O. et al. Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput. Biol.* **10**, e1003810 (2014).
30. Vossel, S. et al. Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. *Cereb. Cortex* **24**, 1436–1450 (2014).
31. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. Associative learning of social value. *Nature* **456**, 245–249 (2008).
32. Hackel, L. M., Doll, B. B. & Amodio, D. M. Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nat. Neurosci.* **18**, 1233–1235 (2015).
33. Brañas-Garza, P., Rodríguez-Lara, I. & Sánchez, A. Humans expect generosity. *Sci. Rep.* **7**, 42446 (2017).
34. Rand, D. G. Cooperation, fast and slow: meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychol. Sci.* **27**, 1192–1206 (2016).
35. Rosenberg, S., Nelson, C. & Vivekananthan, P. S. A multidimensional approach to the structure of personality impressions. *J. Pers. Soc. Psychol.* **9**, 283–294 (1968).
36. Robinson, O. J., Vytal, K., Cornwell, B. R. & Grillon, C. The impact of anxiety upon cognition: perspectives from human threat of shock studies. *Front. Hum. Neurosci.* **7**, 203 (2013).
37. Judd, C. M., James-Hawkins, L., Yzerbyt, V. & Kashima, Y. Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *J. Pers. Soc. Psychol.* **89**, 899–913 (2005).
38. Tormala, Z. L. & Rucker, D. D. Attitude certainty: a review of past findings and emerging perspectives. *Soc. Pers. Psychol. Compass* **1**, 469–492 (2007).

Acknowledgements

We thank D. Carlston, E. Boorman, C. Summerfield and T. Behrens for helpful feedback. We thank T. Tyurkina and L. Caviola for developing the web applications utilized in studies 2–7 for data collection. J.Z.S. was supported by a Clarendon and Wellcome Trust Society and Ethics award (104980/Z/14/Z). R.B.R. was supported by a MRC Career Development award (MR/N02401X/1). This work was supported by a Wellcome Trust ISSF award (204826/Z/16/Z), the John Fell Fund and the Academy of Medical Sciences (SBF001/1008). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.J.C. and J.Z.S. conceived the studies. J.Z.S., C.M., R.B.R. and M.J.C. designed the studies. J.Z.S. collected the data. J.Z.S., C.M. and M.J.C. analysed the data. J.Z.S. and M.J.C. wrote the manuscript with edits from R.B.R. and C.M.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0425-1.

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

Correspondence and requests for materials should be addressed to M.J.C.

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

nature research

Corresponding author(s): Molly Crockett

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 Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Data collection for study 1 was conducted using Matlab (Mathworks) R2016a, and presented using the Cogent Toolbox. All subsequent studies were conducted online and utilized the web application framework Ruby on Rails (studies 2 through 7), or the Qualtrics research platform (study 8).
Data analysis	All data analysis was completed in Matlab (Mathworks). Model parameter estimates were estimated from trial-wise predictions using the Broyden Fletcher Goldfarb Shanno optimization algorithm as implicated in the HGF toolbox, which is available at https://tnu.ethz.ch/tapas

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/reviewers upon request. 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

The data that support the findings of this study are available from the corresponding author upon request. All relevant Matlab code are available from the corresponding author upon request.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

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

Study description	Our manuscript consists of 8 quantitative studies. Studies 1, 2, 3, 5, 7, and 8 are fully within-subject designs. Study 4 is a mixed within and between-subject design, while study 6 is a fully between-subject design.
Research sample	Study 1 included male and female participants, ages 18 to 35, with no history of systemic or neurological disorders, psychiatric disorders, medication/drug use, pregnant women, or more than a years' study of psychology. All subsequent studies included male and female American adults (>18 years of age), residing within the United States. This allowed us to engage a large number of diverse respondents outside of a University subject pool, allowing for a more representative sample.
Sampling strategy	The sample size for the initial study was calculated to detect a moderate effect. All subsequent sample sizes were based directly on power calculations based on the observed effect sizes from previous studies. Supplemental materials, section 4.1.1 and section 6.1.1
Data collection	Data collection for study 1 was conducted in the laboratory using Matlab (Mathworks) R2016a and utilized the Cogent graphics toolbox for stimulus presentation. Responses where made using a standard computer mouse and keyboard. A researcher was present in the laboratory only during task instructions, however no researcher was present while participants completed subsequent experimental procedures/tasks. Data collection for all subsequent studies were conducted online, with participants recruited from Amazon's Mechanical Turk. Investigators were blinded to group allocation during data collection.
Timing	Study 1 was conducted between May 1 2016 and October 11 2016. Study 2 was conducted in October 2015 Study 3 was conducted in February 2016 Study 4 was conducted in March 2016 Study 5 was conducted in May 2016 Study 6 was conducted in July 2017 Study 7 was conducted in May 2017 Study 8 was conducted in October 2016
Data exclusions	Participants were excluded from the analysis if their performance in the learning task (described in Supplementary materials section 1.1.2) was below chance (i.e., less than 50% accuracy). This criteria was pre-established. However, we confirmed that the pattern of the results holds when all participants are included (see Additional Data Table S1 and S2). Methods, page 9 of manuscript. Supplementary materials, sections 1.1.1, 2.1.1, 3.1.1, 4.1.1, 5.1.1, 6.1.1, 7.1.1
Non-participation	No participants dropped out or declined.
Randomization	Participants were randomized within our web application framework, Heroku. Thus, experimenters were unaware of which participants were allocated to which experimental condition. Randomization was not associated with any features of the participant, such as demographic or other individual differences variables.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

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	For study 1, participants were recruited from the University of Oxford Psychology Research recruitment scheme. For all subsequent studies, participants were recruited from Amazon's Mechanical Turk. All participants were compensated for their participation.