Beliefs about bad people are volatile

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

[1]Department of Experimental Psychology, University of Oxford
[2]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[3]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London
[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
[6]Department of Psychology, Yale University
*Correspondence to: mj.crockett@yale.edu

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 ones[13–16]. The ability to accurately infer others' moral character 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 where beliefs about the morality of bad agents are more uncertain (and thus more volatile) than beliefs about the morality of good agents. This asymmetry appears to be a property of learning about immoral agents in general, as we also find greater uncertainty for beliefs about bad agents' non-moral traits. 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 behavior.

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 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 playing repeated prisoner's dilemmas where others' intended actions are implemented with noise[20]. Although evolutionary and economic models provide descriptive accounts of these behaviors, 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 others' moral character, 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 blush, our hypothesis may appear inconsistent with decades of research in social psychology, much of which has examined impression formation from narrative descriptions of extreme and rare behaviors, such as theft or violence. This work provides evidence for a negativity bias in impression formation, where 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 vs. moral behaviors: 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 perceptions of how rare immoral behaviors are, relative to moral ones[25]. This leaves open the question of whether people actually learn differently about agents inferred to be more vs. less moral when their actions are equally diagnostic of their underlying character. This is the central question we addressed in the current studies. We focused on moral inference from behaviors that are not extreme or definitive of character. Such behaviors comprise the vast majority of our daily social interactions: we most often judge others based on behaviors that are nasty or nice, not evil or saintly. Inferring character from minor slights or small favors is

considerably more difficult than doing so from criminal deeds or heroic actions, but our success as a social species suggests 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 behavior 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 Figure 1**). Because ethical systems universally judge harming others for personal gain as morally wrong[28], we operationalized moral 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 participants' expectations of "average" behavior, which was not significantly different from $\kappa = 0.5$ (see Supplementary Materials, Study 8 for details).

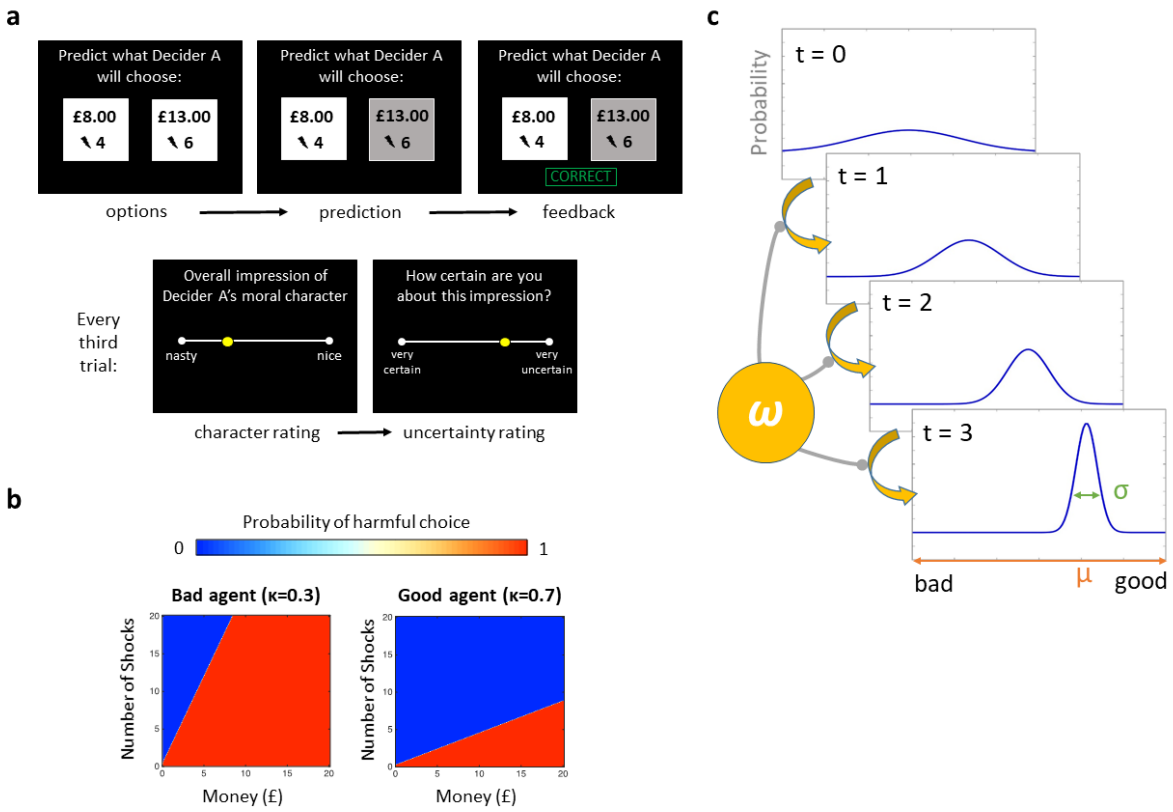


Figure 1. Learning task and model. (**A**) Participants predicted sequences of choices for two agents, "Decider A" and "Decider B". On each trial, the agent chose between a more harmful (more shocks inflicted on another person for more money) and a less harmful (fewer shocks/money) option. After every third trial, participants rated their impression of the agent's moral character. In Studies 2 to 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 and good agent's 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.

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 modeled 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 (i.e., the exchange rate between money and pain, μ); a trial-wise sequence of uncertainties on those beliefs (σ); and a global estimate of belief volatility (ω) that describes the rate at which beliefs evolve over time (**Fig. 1c**).. Belief volatility is set in log space and is monotonically related to belief uncertainty (i.e., 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 Materials, 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 standard error of the mean (sem) as mean±sem.

Our approach extends previous methods employed to probe impression formation in several ways. First, because our paradigm used a computational model of moral preferences rather than narrative descriptions of behaviors (as in past social psychology research), we were able to very tightly control how informative agents' behaviors 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 Materials, study 1 for details). In this way, we ensured that the statistics of the environment did not advantage learning about either the good or bad agent, and this symmetry was confirmed by the fact that an ideal Bayesian observer learned identically about the good and bad agents (**Supplementary Table 3**). Because of this design feature, we can confidently infer that the belief asymmetries we observed in our studies were not due to asymmetries in the information we provided to participants (in contrast to past studies using narrative descriptions of behaviors, where 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].

In an initial study (Study 1) we measured moral inference in 38 participants in the lab. Our model fit participants' predictions well, explaining behavior with 87% accuracy on average (**Supplementary Table 4**). Participants accurately inferred 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**).

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 Table 5 and 6**) We found that the difference in the volatility of beliefs about the good and bad agents' moral character was significantly larger for participants compared to 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 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 participants' moral impressions of the two agents affected their social behavior by inviting them to entrust money to each agent in a one-shot trust game after learning about both agents (see **Supplementary Materials**, 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 Table 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, Supplementary Table 1 and 5**). Our model predicts that there would be larger trial-wise updating of character ratings for the bad agent than the good agent. This was confirmed in a model-free analysis where we compared the magnitude of changes in trial-to-trial ratings between good and bad agents (see Supplementary Materials, study 2 and **Supplementary Table 8**).

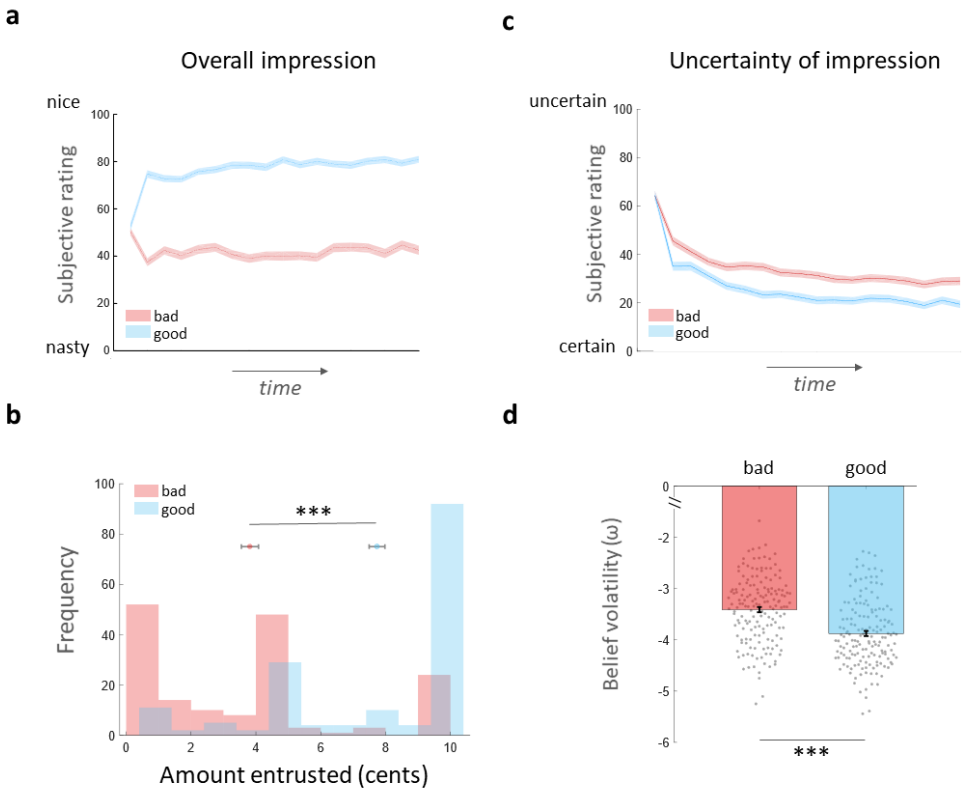


Figure 2. Asymmetry in moral impression formation, Study 2. (**A**) Trajectory of subjective character ratings over time in Study 2, averaged across participants (N=163 for all panels). (**B**) In a one-shot trust game, participants entrusted the good agent with twice as much money as they did the bad agent. (**C**) Trajectory of subjective uncertainty ratings over time, averaged across participants. Subjects reported greater uncertainty about bad agents. (**D**) Volatility of beliefs (ω in the model) was higher for the moral character of the bad compared to the good agent. Error bars and shaded bounds in trajectories represent SEM. ***P < 0.001

In a third study (N = 135), we increased the stochasticity of agent choices to test whether the differences we observed for learning about bad compared to good agents are robust to noisy environments. We replicated all the findings from Studies 1 and 2 (see Supplementary Materials, Study 3, **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 artifact of the scale participants used to rate the agents' morality (ranging from *nasty* to *nice*), we replicated all findings in a supplementary study using an alternative scale (ranging from *bad* to *good*; see Supplementary Materials, 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 good agents is a strong prior expectation that people will behave morally[32,33], thus rendering the bad agent's behavior more surprising. To investigate this possibility, we asked a separate group of participants to predict, in the context of decisions to profit from others' pain, how "most people" would choose (see Supplementary Materials, 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 where they successfully predicted the majority response. We found no evidence that people expect 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 (κ = 0.5) equidistant from that of the good and bad agents (mean expected κ = 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 participants' beliefs about the agents' character prior to 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 Materials, Study 2 and **Supplementary Table 9** and **Supplementary Figure 2**). We also did not find a relationship between learning asymmetries and self-reports of generalized trust in others (see Supplementary Materials, Study 8 and **Supplementary Figure 3**).

In a fourth study (N=220), we examined whether the asymmetry in learning about bad compared to good agents extend 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 and a good agent as before. In the competence condition, participants predicted the basketball performance (number of points scored per minute) of a low-skill 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[34]. In contrast, we expected inferences about basketball ability to be independent from inferences about moral character. Pilot testing supported this claim (See Supplementary Materials, study 4). Thus, our design allowed us to directly test the specificity of our observed effect for moral inference because it is unlikely participants would form moral impressions from observations of basketball performance alone.

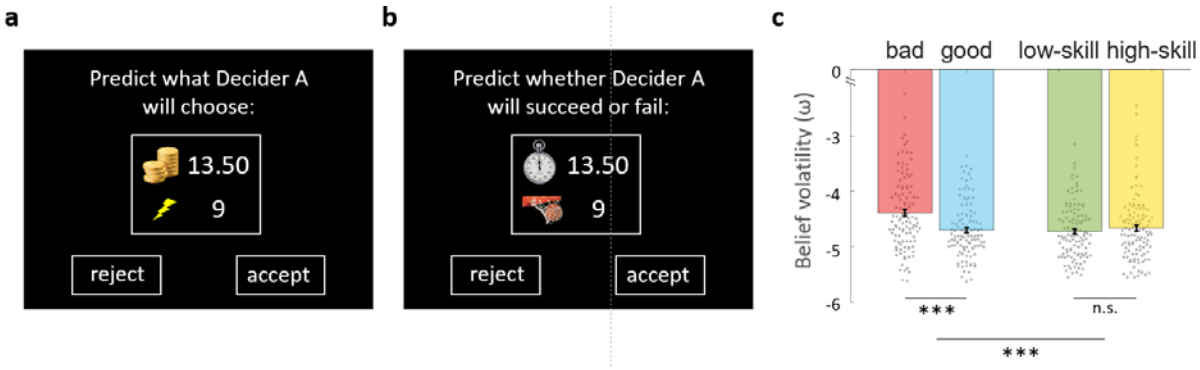


Figure 3. Forming impressions of morality vs. competence, Study 3. (**A**) In the morality condition, participants (N=109) predicted whether the agent would deliver a certain number of shocks for a specified profit. (**B**) In the competence condition, participants (N=111) predicted whether the agent would succeed in scoring a certain number of points within a specified amount of time. (**C**) Interaction between agent (bad/low-skill vs. good/high-skill) and condition (morality vs. competence) for the volatility of beliefs (ω in the model). Error bars represent SEM. ***P < 0.001; n.s. = not significant

As predicted, between-agent differences in uncertainty ratings and belief volatility were significantly larger in the morality condition than the competence condition (rank sum test,

difference in mean uncertainty rating: morality = 4.870±1.467, competence = -2.778±1.161, P<0.001; difference in ω: morality = 0.316±0.069, competence = -0.060±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±1.598; good = 24.166±1.607, P<0.001; ω: bad = -4.390±0.064; good = -4.714±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±1.227; high-skill = 20.653±1.274, P=0.076; ω: low-skill = -4.726±0.047; high-skill = -4.655±0.057, P=0.566; **Fig. 3c**).

Previous work has shown bad behaviors carry more weight than good behaviors 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 good and bad agents make. We predicted that the threatening nature of bad agents would increase the uncertainty 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[35]. 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 and 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±0.046; good = -4.428±0.039, *P*<0.001; competence ω: bad = -4.224±0.039; good = -4.327±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±1.019; good = 24.209±1.027, *P*<0.001; competence uncertainty rating: bad = 28.875±1.995; good = 27.277±1.992, *P*=0.020; **Supplementary Table 5**).

Figure 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 agents' moral character and skill level after every third morality and competence trial, respectively. (**B**) Comparison of volatility of beliefs about the good and bad agent's morality (left) and competence (right) in Study 5, N=189. (**C**) In Study 6, participants were randomized to learn about a bad agent (κ=0.3) or a good agent (κ=0.7) whose moral

character either improved (κ+0.2) or worsened (κ–0.2). (**D**) In Study 6, participants more strongly updated their impressions of bad than good agents when moral character improved but not when it worsened (N=364). Error bars represent SEM. **P < 0.01, ***P < 0.001

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 behavior 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 behavior (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 Materials, study 6). In this study we explicitly set participants' prior beliefs at k = 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 good agents. We tested our hypothesis by comparing, for bad vs. good agents, the extent to which participants updated their impressions, defined as the difference between character ratings before vs. 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, where participants updated their character ratings more for bad good than agents (rank sum, update: bad = 18.951±1.245, good = 14.928±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 ±1.389; improve = 11.468 ±1.010, P<0.001). This is consistent with past reports of negativity bias in impression formation[9–11], where people show stronger impression updating in response to inconsistent immoral behaviors relative to moral behaviors. Main effects were qualified by an interaction between agent and shift direction (Kruskal-Wallis, P<0.001), where asymmetric updating was more pronounced when morality improved than when morality worsened (**Fig. 4d**). At first glance, this interaction may appear 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 behavior 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 vs. after the agents' behavior shifted (Spearman's ρ, P = 0.006; Supplementary Materials, 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 behavior improved, impressions were updated faster for putatively bad agents than good agents. Thus, the volatility of bad moral impressions may facilitate forgiveness by enabling initially bad impressions to be rapidly updated if behavior 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 we identify 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.

Although theoretical models of person perception have claimed the independence of trait dimensions (namely warmth and competence)[1], other evidence suggests that judgments across trait dimensions may share a positive relationship[34,36]. 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, where bad behaviors command more attention than good behaviors[9–12], and research showing that uncertain attitudes are susceptible to change[37]. 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, where impressions about bad agents are more rapidly updated than impressions about good agents. We suggest that by destabilizing 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 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 (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 subject 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 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 where 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 subjects: 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 how the average person behaved in the task. As in Studies 4 and 5, participants received a monetary bonus for high accuracy.

In a supplementary study (referred to as Study 7 in Supplementary Materials), 125 participants were recruited from AMT, and nine were excluded for below-chance performance. Study 7 was identical to Studies 1 and 2, however instead of rating the agents' moral character on a scale ranging from *nasty* to *nice*, participants rated the agents' moral character on a scale ranging

from *bad* to *good*. To motivate accurate predictions, participants received a monetary bonus for high accuracy.

In a second supplementary study (Supplementary Materials, study 8), 30 participants were recruited from AMT to predict, in the context of decisions to profit from others' pain, 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 the completion of the study.

References and notes

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. A brief history of theory and research on impression formation. in *The Oxford handbook of social cognition* (ed. Carlson, D.E.) 53–73 (Oxford University Press, 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. *Emot. Wash. DC* **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. Confl. 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. Cognitive adaptations for social exchange. in *The adapted mind: Evolutionary psychology and the generation of culture* (eds. Barkow, J. H., Cosmides, L. & Tooby, J.) 163–228 (Oxford University Press, 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. Confl. 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.* **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 University Press, 2004).

29. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. Associative learning of social value. *Nature* **456,** 245–249 (2008).

30. Diaconescu, A. O. *et al.* Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLoS Comput Biol* **10,** e1003810 (2014).

31. 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 (2015).

32. Brañas-Garza, P., Rodríguez-Lara, I. & Sánchez, A. Humans expect generosity. *Sci. Rep.* **7,** 42446 (2017).

33. Rand, D. G. *Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation*. (Social Science Research Network, 2016).

34. Rosenberg, S., Nelson, C. & S, P. A multidimensional approach to the structure of personality impressions. *J. Pers. Soc. Psychol.* **9,** 283–294 (1968).

35. 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,** (2013).

36. 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 (2005).

37. Tormala, Z. L. & Rucker, D. D. Attitude Certainty: A Review of Past Findings and Emerging Perspectives. *Soc. Personal. Psychol. Compass* **1,** 469–492 (2007).

Data availability

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

Code availability

All relevant Matlab code are available from the corresponding author upon request.

Correspondence

All correspondence regarding this article should be addressed to Dr. Molly Crockett, molly.crockett@yale.edu.

Acknowledgments

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).

Author contributions

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

Financial and non-financial competing interests

The authors declare no financial or non-financial competing interests.

Supplementary Materials

Table of Contents

Study 1: Beliefs about bad people are volatile

Methods

Participants

Study 1 took place at the Department of Experimental Psychology, University of Oxford and was approved by the Oxford research ethics committee (MS-IDREC-C1-2015-001). Thirty-nine participants were recruited from the Oxford Psychology Research recruitment scheme. Participants with a history of systemic or neurological disorders, psychiatric disorders, medication/drug use, pregnant women, and more than a years' study of psychology were excluded from participation. All participants provided informed consent prior to initiation of the study and were compensated for their time. One participant was excluded from the analysis as their performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 38 participants. We confirm the pattern of results is similar when we include all participants in **Supplementary Table 1** and **5**.

Experimental procedure

Learning task

We developed a novel learning task to examine the computational mechanisms of moral inference. In the task, participants predicted a sequence of 50 choices made by each of two agents and on each trial received feedback about their accuracy. On each trial, an agent chose between two options: more money for themselves plus more shocks for a third-party victim (C_{harm}), or less money for themselves plus fewer shocks for the victim (C_{help}; see **Figure 1a**). No *a priori* information was given about the agents. Instead, participants were required to learn about the agents' moral preferences through trial and error. Participants completed the whole sequence for one agent before beginning with the second, and the order of agents was randomized across participants. Additionally, on every third trial participants rated their subjective impression of the agent's character on a continuous visual analogue scale ranging from 0 (*nasty*) to 100 (*nice*). After making subjective character ratings, participants indicated how uncertain they were about this characterization on a scale ranging from 0 (*very certain*) to 100 (*very uncertain*) (see **Figure 1a**). These measures were additionally collected before participants observed either of the agents' choices, in order to obtain an estimate of participants' prior beliefs about how moral agents will behave in this setting. Together, this provided us with a trajectory of participants' explicit subjective ratings of each agent's moral character, and how uncertain participants were about their characterization, which was later used to validate findings from our computational model.

To manipulate moral character, we created agents with different preferences towards harming the victim. This was operationalized as their exchange rate between money for themselves and pain for the victim, described with a single *harm aversion* parameter, κ. We previously showed that this parameter accurately captures individual differences in moral decision-making and correlates with

several traits related to prosocial behavior, including empathy and psychopathy[1–3]. When $\kappa = 0$, agents are minimally harm averse and will accept any number of shocks to the victim to increase their profits; as κ approaches 1, agents become maximally harm averse and will forgo infinitely increasing amounts of money to avoid delivering a single shock. For the learning task, we created one agent who was characteristically 'bad' ($\kappa = 0.3$), and another who was characteristically 'good' ($\kappa = 0.7$; **Supplementary Figure 1**). Effectively, this meant that the bad agent was less averse to harming the victim and would therefore require less money to inflict pain than the good agent. Participants observed the two agents make choices for identical trial sequences. On every trial, the agents faced the same two options, but because the agents had different preferences towards harming the victim, they often chose differently.

Each trial contained a pair of choices [s-, m-] and [s+, m+] that matched the indifference point of a specific κ value. We first created a set of 24 trials where values of κ were randomly drawn from a normal distribution around the good agent's indifference point (M = 0.7, s.d. = 0.15), and constrained such that $\kappa < 0.95$. Next, we created a set of 24 matched trials around the bad agent's indifference point by subtracting each κ value from 1. We wanted participants to observe identical trial sequences for the two agents, but also minimize any potential differences in learning about the agents that could be explained by discrepancies in the informational value of the trial sequence. Note that a trial with high informational value for the bad agent will have relatively low informational value for the good agent, and vice versa. Consequently, we created pairs of trials [κ, 1- κ] where the members of each pair were matched in informational value for the good and bad agent. Effectively, this meant that a trial that was highly informative about one agent's indifference point was paired with a trial that was equally informative about the other agent's indifference point (**Supplementary Figure 4**). We then randomized the order of presentation of each member of the pair. The pairs comprised trials 2-49 of the sequence, while the initial and final trials were fixed to $\kappa = 0.5$.

Given a sequence of κ values, we then generated shock and money options for each κ value by generating 10,000 random pairs of positive shock movements Δs ($1 < \Delta s < 20$), and positive money movements Δm ($0.10 < \Delta m < 19.90$), and selected the pair closest to the indifference point of that κ value [Δs, Δm]. Next, these pairs were transformed into choices containing smaller amounts of shocks and money (s- and m-) and greater amounts of shocks and money (s+ and m+) as follows: s- was a positive integer between 0 and 20, randomly drawn from a uniform discrete distribution with the constraint that $0 < s- + \Delta s < 20$. Similarly, m- was a positive number between 0 and 20, randomly drawn from a uniform discrete distribution, rounded to the nearest 10th and constrained such that $0 < m- + \Delta m < 20$. s+ and m+ were then set by adding Δs and Δm to s- and m-, respectively.

We simulated the agents' decisions by computing the utility for choosing the more harmful option (V_{harm}) as a function of the agent's κ ($\kappa_{bad} = 0.3$, $\kappa_{good} = 0.7$). This model is identical to the model that best predicts human choices in the same setting[1,2].

$$V_{harm} = (1 - \kappa_n)\Delta m - \kappa_n \Delta s \quad \textbf{(1)}$$

Where κ_n is the κ for agent n. A softmax function was used to transform V_{harm} into a probability of choosing the more harmful option, P_{harm} :

$$P_{harm} = \frac{1}{1+e^{-\beta \times V_{harm}}} \quad \textbf{(2)}$$

Where β defines the steepness of the slope in the sigmoid function. As β approaches 0 the slope become increasingly horizontal, signifying a large amount of noise in the agent's choices. As β approaches infinity the sigmoid approximates a step function, and indicates increasingly deterministic choice preferences. In Study 1 β was fixed to 100 to simulate agents that were completely deterministic in their choices.

$$u = [x_{rand} < P_{harm}] \quad \textbf{(3)}$$

Eq. (3) converts the probability of choosing the more harmful option into a binary choice, u. x_{rand} is a random number between 0 and 1.

Computational modelling

Perceptual Model

Our main goal was to assess how participants updated their beliefs about an agent's moral character when that agent was either bad or good. For this purpose we applied one primary perceptual model to participants' trial-by-trial responses: a reduced version of the Hierarchical Gaussian Filter (HGF), a Bayesian model for learning hidden states with informational uncertainty (due to a lack of knowledge) and without environmental volatility (See Mathys and colleagues[4] for a theoretical background on the full model). The HGF draws on the belief that the brain has evolved to process information in a manner that approximates statistical optimality given individually varying priors about the nature of the process being predicted; effectively maintaining and updating a generative model of its inputs (u) to infer on hierarchically organized hidden states (see **Fig.1c**). For the purpose of this study, our model comprises only two hidden states x_1^i and x_2^i, where i signifies the trial index. The first state, x_1, is time-varying and denotes the agent's upcoming choice. x_1 is binary because there are only two options that the agent can choose: the more harmful option (greater profit for the self and more shocks for the victim) or the less harmful option (less profit for the self and fewer shocks for the victim). The probability that an agent will chose the more harmful option ($x_1^i = 1$) versus the less harmful option ($x_1^i = 0$) is governed by the next state in the hierarchy, x_2. x_2 is a continuous state evolving over time as a Gaussian random walk, and signifies the agent's (logit-transformed) κ. The hierarchical coupling between x_1^i and x_2^i explains that a participant's prediction about an agent's choice on trial i is dependent on their current belief about that agent's κ, defined as a probability density. The conditional probability of x_1 given x_2 is described in Eq. (4).

$$p(x_1|x_2) = s(x_2)^{x_1}\left(1 - s(x_2)\right)^{1-x_1} = \text{Bernoulli}(x_1; s(x_2)) \quad \textbf{(4)}$$

Where $s(\cdot)$ is a logistic sigmoid (softmax) function:

$$s(x) \stackrel{\text{def}}{=} \frac{1}{1+\exp(-x)} \quad \textbf{(5)}$$

The temporal evolution of x_2 is governed by a participant-specific parameter ω, which allows for inter-individual differences in belief updating. Thus, ω represents a measure of the tonic volatility – i.e., the extent of trial-wise changes in x_2. In other words, ω captures inter-individual variability in the rate at which beliefs evolve over time, and consequently how rapidly people update their beliefs about the agent's harm aversion. As ω approaches ∞ beliefs become increasingly unstable and new information is favored over prior beliefs. Conversely, as ω approaches -∞ beliefs become increasingly stable, so greater weight is instead placed on prior beliefs. Given ω and the previous value (with time index $i-1$) of x_2, we now have the generative model for the current values (with time index i) of x_1 and x_2 in Eq. (6) (graphically represented in **Fig.1c** of the main text; for details see reference 3).

$$p(x_1^i, x_2^i, |\omega, x_2^{i-1}) = p(x_1^i|x_2^i)p(x_2^i|x_2^{i-1}, \omega) \quad \textbf{(6a)}$$

with

$$p(x_2^i|x_2^{i-1}, \omega) = \mathcal{N}(x_2^i; x_2^{i-1}, \exp(\omega)) \quad \textbf{(6b)}$$

Model inversion was used to optimize the posterior densities over hidden states, x_1 and x_2, and parameter ω. Participants' posterior beliefs were represented by probability distributions with mean μ and variance σ. Variational Bayesian inversion yields a simple update equation under a mean-field approximation, where beliefs are updated as a function of precision-weighted prediction errors. For the present study we focus on the update at level 2 of the hierarchy[5].

$$\Delta\mu \propto \sigma_2 \delta_1^i \quad \textbf{(7a)}$$

with

$$\delta_1^i = \mu_1^i - \hat{\mu}_1^i \quad \textbf{(7b)}$$

and

$$\sigma_2 = \frac{\hat{\pi}_1^i}{\hat{\pi}_2^i \hat{\pi}_1^i + 1} \quad \textbf{(7c)}$$

Where π is the precision (i.e., the inverse variance) in participants' posterior belief $\frac{1}{\sigma}$, and δ_1^i is the prediction error on the trial outcome. Caret symbols (^) are used to denote predictions *prior* to observing the outcome at trial *i*. Thus, $\hat{\pi}_1^i$is the precision of the prediction at the first hierarchical level and $\hat{\pi}_2^i$ is the precision of the prediction of the posterior belief. It can be shown[3] from Eq. (7c) that prediction errors are given a larger weight when the precision of the prediction of the agent's choice is high, or when the precision of the belief about the agent's κ is low. In summary, these equations describe trial-wise updating of beliefs about an agent's preference towards harming the victim, which approximates Bayes optimality (in an individualized sense given differences in ω) and determines the participant's estimate of the probability that an agent will harm. Crucially, our model provides a trial-by-trial estimate of the subject's uncertainty about the agent's preference towards harming the victim.

Decision Model

The decision model describes how the participant's posterior belief about the agent's κ maps onto their predictions of the agent's decisions (y). In the HGF, this belief $\hat{\mu}_1^i$ corresponds to the logistic sigmoid transformation of the predicted preference μ_2^{i-1} of the agent towards harming the victim.

$$\hat{\mu}_1^i = s(\mu_2^{i-1}) \quad \textbf{(8)}$$

For the present study, we assumed that participants would predict others' decisions using a similar rationale to how they make decisions themselves. In other words, we assumed that people's preferences are described by a utility model, and that people think others' preferences are described by the same model. Consequently, we applied a decision model that accurately describes human choices in the same choice setting[1–3] (and that was also used to simulate the agent's actual choices).

$$V_{\text{harm}}^i = \left(1 - \hat{\mu}_1^i\right) \Delta \text{m}^i - \hat{\mu}_1^i \Delta \text{s}^i \quad \textbf{(9)}$$

This model replaces a participant-specific parameter κ with the predicted belief derived from the perceptual model $\hat{\mu}_1^i$ to compute the value that the agent will choose C_{harm} on trial i. The probability that the participant predicts C_{harm} (y = 1) as opposed to C_{help} (y = 0) is described by the softmax function in Eq. (10).

$$P_{\text{harm}}^i = s(\beta V_{\text{harm}}^i) \quad \textbf{(10)}$$

Where β is a free parameter (individually estimated like ω) that describes how sensitive predictions are to the relative utility of different outcomes, or the prediction noise.

Estimation of Model Parameters

A crucial aspect of Bayesian inference is the specification of a prior distribution for the belief (listed in **Supplementary Table 10**). We defined the priors based on our experimental design. Specifically, for the current study we wanted to compare learning parameters between the two agents. In keeping with our experimental design, which did not give participants any basis for assumptions about the agent's tendency to harm, we chose to fix the prior mean over μ_2 and σ_2. such that it amounted to a neutral prior belief about κ which was equidistant from the true value of the agents' preferences. For the free parameters ω and β, we chose a prior mean that was relatively uninformative (with large variance) to allow for substantial individual differences in learning both between participants and within participants (i.e. between agents). Notably, the prior means on ω and β were equally unconstrained with a variance of 1. This ensured that adjustments in parameter estimates were not biased towards favoring one parameter over the other.

The perceptual model parameter ω and decision model parameter β were estimated from the trial-wise predictions using the Broyden Fletcher Goldfarb Shanno optimization algorithm as implemented in the HGF Toolbox (https://tnu.ethz.ch/tapas). This allowed us to obtain the maximum-a-posteriori estimates of the model parameters and provided us with state trajectories

and parameters representing an ideal Bayesian observer given the individually estimated parameter ω.

For the present study we were interested in how the computational processes during learning differ when observing agents with different moral preferences. Specifically, we predicted that the inferred morality of agents would affect how uncertain participants were in their beliefs about the agent's κ and consequently the volatility of their beliefs, ω. The variance in the posterior belief (σ_2) reflects a measure of uncertainty, however because it exists in logit-probability space it is not directly interpretable in its current form. Thus, σ_2 is transformed into prior uncertainty (σ_T) about κ by appealing to the standard variable transformation rule for probability distributions $p(y) = \frac{dy}{dx}p(x)$. With $\frac{ds(x)}{dx} = s(x)(1 - s(x))$, this gives us Eq. (11)

$$\sigma_T^i = s(\mu_2^i)\left(1 - s(\mu_2^i)\right)\sigma_2^i \quad \textbf{(11)}$$

We fit the model separately for participant's predictions of the bad and good agent. This produced for each agent a sequence of trial-wise beliefs about the agent's κ ($\hat{\mu}_1^i$), as well as prior uncertainties σ_T^i, and two participant-specific parameters, ω and β. Where possible, we validated findings from the HGF model with raw behavioral data not derived from a model.

Matching optimal Bayesian trajectories

A main goal was to investigate whether observed differences in learning about good and bad agents reflected systematic deviations from the performance of a task-local definition of optimal Bayesian learning. Thus, the crucial test is whether human learning differs for good and bad agents in a setting where an ideal Bayesian observer learns identically about these agents. Although we took great efforts to minimize differences in learning that might stem from discrepancies in informational value of trials, it was not possible to eliminate such discrepancies completely. Indeed, even for an optimal observer, the only way two learning trajectories (one for each of two actors) could be identical is if both the trials and the choices made were identical. Thus, small residual discrepancies in informational value about good and bad agents across trials could potentially create learning differences that do not reflect a true asymmetry in learning between agents. Consequently, we generated 100 permutations of trial sequences and simulated behavior for an ideal Bayesian observer for each (see tapas_fitModel.m and tapas_bayes_optimal_binary_config.m in the HGF toolbox). Two sequences were selected that best minimized differences in our main dependent variables (ω and $\overline{\sigma_T}$) for an ideal observer. Each participant in the study was randomly assigned to complete 1 of the 2 trial sequences. With this process in place, we minimized the possibility that any differences between agents observed could be explained by the order of observations, and instead reflect a systematic deviation from optimal learning.

Model comparison

To demonstrate that the HGF model offers a reasonable description of behavior above simpler models, we compared our HGF model to two alternative models: (a) a Rescorla Wagner (RW) model, in which beliefs are updated by prediction errors with a single fixed learning rate (1 learning

rate RW), and (b) a Rescorla Wagner model, in which beliefs are updated by prediction errors with separate fixed learning rates for positive and negative outcomes (2 learning rate RW). For details about the alternative models, see **Supplementary Table 2**. We verified that the log-model evidence (LME) indicated that our model outperforms both a simple single learning rate RW model and a RW model with separate learning rates for positive and negative outcomes. See **Supplementary Table 11** for details. We validated these findings using formal Bayesian Model Selection, which is a random-effects procedure that takes into account inter-subject heterogeneity[6,7]. To this end, we used LME data across all studies (N = 1419) to compare between the HGF and our two RW models. This analysis yielded a protected exceedance probability indistinguishable from 1 for the HGF model for both agents, indicating effectively a 100% probability that the HGF model better explains the data than the other models included in the comparison.

Statistical Analyses

All data analysis was completed in Matlab (Mathworks). To test whether group mean parameter estimates and mean ratings differed significantly between good and bad agents, we used nonparametric statistical tests that do not make any assumptions about their underlying distributions. Effect sizes were computed for significant results using Rosenthal's formula: $r = Z/\sqrt{n}$, which has been proposed as a viable alternative calculation when the general assumptions of Cohen's formula have been violated[8].

Although we primarily focus on the difference in *average* uncertainty (subjective rating uncertainty and model parameter estimate σ_T) between good and bad agents, we confirmed that all our results are robust to controlling for the effects of time, using linear regression of the time-series data with agent and time as separate regressors (for results, see **Supplementary Table 6**).

Results

We first investigated whether we were able to recover participant's choices using the estimates derived from the model. To this end, we simulated 100 sequences of choices using each participant's parameter estimates and compared the simulated choices to participants' actual predictions. The model fit participants' predictions well, explaining behavior with a mean 87.0% accuracy for the bad agent and a mean 86.3% accuracy for the good agent in Study 1 (see **Supplementary Table 4** for details of model goodness-of-fit for all studies).

Next, we investigated whether participants indeed learned through trial-and-error about the agents' moral preferences in the task. We analyzed the model's final estimates of participants' beliefs about each agent's κ ($\hat{\mu}_1^{107}$), and verified that participants formed beliefs that closely resembled the agent's true κ and significantly differed from one another (mean±SEM bad: 0.322±0.004; good: 0.681±0.004; $Z = -5.373$, $p < 0.001$; **Supplementary Table 1**). Subjective character ratings also confirmed sufficient learning by our participants; final ratings indicated the good agent was generally characterized as nice and the bad agent as nasty (bad: 0.427±0.040; good: 0.789±0.029; $Z = -5.303$, $p < 0.001$; **Supplementary Table 5**), where ratings above 0.5 are classified as 'nice' and ratings below 0.5 are classified as 'nasty'.

In line with our predictions, participants were significantly more uncertain in their beliefs about the bad agent on average ($\overline{\sigma_T}$) relative to the good agent (bad: 0.060±0.002; good: 0.054±0.002; *Z* = 2.302, p = 0.021; see **Supplementary Table 1** for results of all studies and effect sizes). Subjective uncertainty ratings validated the findings from the model, as participants reported greater uncertainty about their characterizations for the bad agent on average (bad: 28.62 ±2.428; good: 20.612±1.371; Z = 3.444, p < 0.001; **Supplementary Table 5**). This translated into faster updating for the bad agent than the good agent, as demonstrated by a larger ω (bad: -3.779±0.102; good: -4.212±0.104; *Z* = 3.212, p = 0.001; **Supplementary Table 1**). Taken together Study 1 suggests that moral inferences modulate the computational and cognitive mechanisms for updating beliefs, which may stem from a reduced reliance on priors for more threatening agents.

In a supplementary analysis, we investigated whether observed differences in ω could be explained by differences in β or an overlap between β and ω. Across studies, we found no consistent relationship between ω and β (see **Supplementary Table 12**). In addition, we found no consistent differences in β between good and bad agents across studies (see **Supplementary Table 1**).

Study 2: Replication, subjective uncertainty, and comparison with ideal Bayesian observer

Methods

Participants

Two-hundred and fifty-three U.S. residents were recruited from Amazon's Mechanical Turk (AMT[9]). The study utilized the web application framework Ruby on Rails, and was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Conducting the study online had the advantage of engaging a large number of diverse respondents outside of the University's limited subject pool. All participants provided informed consent and were compensated for their time. Eighty-seven participants were excluded from the analysis as their behavioral performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 163 participants. We confirm the pattern of results is similar when we include all participants in **Supplementary Table 1** and **5**.

Experimental Procedure

We constructed the learning task for Study 2 using identical procedures to those outlined in Supplementary Methods, Study 1. In order to motivate accurate predictions, participants in Study 2 were explicitly instructed to pay attention and learn about the behavior of the agents, as they would later have to decide whether to trust the agents in a one-shot investment game[10] that could earn them additional money.

Trust Game: After completing the learning task, participants engaged in a trust game with each of the agents. Participants were given ten cents that they could entrust with each agent. Any amount that they entrusted with the agent would be tripled, and the agent can choose how much to return of the tripled amount. Thus, if participants do not entrust any amount they could keep the initial ten cents. However, if they choose to entrust some amount of money then they might receive a higher amount in the end, depending on how much the agent gives back to them. We instructed participants that the percent returned by each agent has been predetermined, and thus the agents are not playing actively. We set the returned amount to correspond to the agents' actual moral characters, such that the bad agent behaved more selfishly than the good agent (the bad agent returned 20% and the good agent returned 50% of the tripled amount). The final amount was paid out to participants as a bonus.

$$\text{Bad agent bonus} = (10 - \text{amount entrusted}) + (\text{amount entrusted} * 3 * 0.2)$$

$$\text{Good agent bonus} = (10 - \text{amount entrusted}) + (\text{amount entrusted} * 3 * 0.5)$$

Results

As a manipulation check, we examined whether participants trusted the good agent to a greater extent than the bad agent by comparing the amount participants entrusted to each agent in the trust game. As expected, participants entrusted significantly more to the good agent (good: 7.74±0.24) than the bad agent (bad: 3.82±0.27; Z = -8.522, p < 0.001; **Figure 2b** and **Supplementary Table 7**).

Next, we investigated whether participants indeed learned through trial-and-error about the agents' moral preferences in the task. We analyzed the model's final estimates about each agent's κ ($\hat{\mu}_1^{50}$), and verified that participants formed beliefs that closely resembled the agent's true κ (bad: 0.301±0.004; good: 0.707±0.003; **Supplementary Table 1**; See **Supplementary Figure 5** for a graphical depiction of the temporal evolution of $\hat{\mu}_1$). Subjective ratings also confirmed sufficient learning by our participants; final ratings indicated that the good agent was generally characterized as nice and the bad agent as nasty (bad: 0.422±0.020; good: 0.807±0.014; **Supplementary Table 5**).

Participants were significantly more uncertain in their beliefs about the bad agent on average ($\overline{\sigma_T}$, bad: 0.080±0.002; good: 0.067±0.001; Z = 6.583, p < 0.001; **Supplementary Table 1**; see **Supplementary Figure 5** for a graphical depiction of the temporal evolution of σ_T). Subjective uncertainty ratings validated the findings from the model, as participants reported greater uncertainty about their characterizations for the bad agent on average (bad: 33.078±1.329; good: 24.087±1.371; Z = 7.213, p < 0.001; **Supplementary Table 5**). A generalized linear regression was performed to investigate whether trial-wise belief uncertainty extracted from the model (σ_T) predicted participants' subjective ratings of uncertainty. We found that the relationship was significantly positive as indicated by a statistical difference in the slope from zero (bad: Z = 5.778, p < 0.001, r = 0.433; good: Z = 8.469, p < 0.001).

In line with Bayesian theory, greater uncertainty about the bad agent effectively translated into faster updating as demonstrated by a larger tonic volatility ω (bad: -3.411±0.050; good: -3.877±0.051; Z = 6.830, p < 0.001; **Supplementary Table 1**). To investigate whether participants similarly exhibited larger updating in their subjective character ratings for the bad agent, we computed for each participant the absolute change in rating from trial to trial (Δrating), and compared the average Δrating between the two agents. As expected, the Δrating was larger for the bad agent (bad: 9.780±0.602; good: 7.909±0.452; Z = 2.787, p = 0.005; **Supplementary Table 8**), indicating a greater tendency to adjust moral impressions in response to new information.

Our data showed that beliefs were more uncertain when observing a bad agent relative to a good agent, and this was accompanied by a faster learning rate. An additional goal of Study 2 was to investigate whether participants' behavior deviated from that of an ideal Bayesian observer. Although we took great efforts to minimize differences in optimal learning parameters between agents in our trial sequences, we additionally pursued post-hoc testing as validation. To this end, we computed for each participant the difference in parameter estimates between good and bad agents for each of our main dependent measures ($\Delta\overline{\sigma_T}$, Δω) and compared this to the difference in parameter estimates for an optimal agent observing the same trials. We found that human learning in this setting significantly differed from Bayes-optimal learning, as the effective difference

between agents was significantly greater in our sample for each of the dependent measures extracted from the model ($\Delta\omega$: Z = -7.383, $p < 0.001$; $\Delta\overline{\sigma_r}$: Z = -7.504, $p < 0.001$).

Because, in theory, belief uncertainty is directly related to prior expectations, one possible explanation for why people show more uncertain beliefs about the bad agent than the good is a strong prior expectation that people will behave morally. However, we found no consistent relationship between explicitly stated prior beliefs and the difference in parameter estimate, $\Delta\omega$, across studies (see **Supplementary Table 9** and **Supplementary Figure 2** for a graphical depiction of the relationship between prior beliefs and $\Delta\omega$, aggregated across all studies). This suggests that the observed differences in uncertainty and volatility between good and bad agents are unlikely to be explained by prior expectations.

In a final analysis, we asked whether behavior using a simpler reinforcement learning mechanism would lead to a similar pattern of results as identified in our HGF model. That is, we aimed to determine whether larger ω for bad agents could be recovered from behavior based on simpler Rescorla Wagner models. To investigate, we first extracted the average parameter estimates from the two Rescorla Wagner models outlined in our model comparison, Supplementary Methods, Study 1 (**Supplementary Table 2**). The first model included a single fixed learning rate and a noise parameter, while the second model included separate learning rates for positive and negative outcomes and a noise parameter.

We next simulated behavior for 1000 fake participants on the set of trials from Study 2, whose parameter estimates were drawn from distributions with means equal to the extracted average parameter estimates and standard deviations equal to the standard deviation of those estimates. Next we fit our HGF model to the simulated data and checked whether our observed effect (larger ω for bad relative to good agents), could be recovered by fitting the behavior of 'participants' who were actually behaving in a manner consistent with either of these Rescorla-Wagner models.

When we simulated behavior based on either Rescorla-Wagner updating process, this did not lead to the same parameter differences observed in our data. While our data showed larger volatility estimates for the bad agents than the good agents, the data simulated using the Rescorla Wagner models led to larger volatility estimates for the good agent than the bad agent (Single learning rate RW ω: bad = -4.250±0.005; good = -4.152±0.006; 2 learning rate RW ω: bad = -4.284±0.001; good = -4.173±0.004). Thus, in addition to the fact that the Rescorla-Wagner models do not fit participants' behavior as well as the HGF model, these alternative learning models do not lead to the same volatility parameter differences we observed in our data.

Study 3: Adding noise to agents' choice behavior

Methods

For Studies 1 and 2, agents were simulated to behave deterministically, never deviating from their preference towards harming the victim. In other words, agents deterministically chose the more harmful option when V_{harm} in Eq. 1 was greater than zero, given the agent's harm aversion, □, and deterministically chose the less harmful option when V_{harm} was smaller than zero. However, human behavior is not always consistent, especially when choices become increasingly difficult (i.e., when V_{harm} is close to zero). Consequently, Study 3 investigated whether our effects would hold when behavior is not deterministic.

Participants

One-hundred and sixty-two U.S. residents were recruited from AMT. All participants provided informed consent and were compensated for their time. Study 3 was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Twenty-seven participants were excluded from the analysis as their behavioral performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 135 participants. We confirm the pattern of results is similar when we include all participants in **Supplementary Table 1** and **5**.

Experimental Procedure

In general, the experimental procedure for Study 3 was very similar to Studies 1 and 2 (including the instructions). However, in order to simulate agents that did not behave deterministically, we fixed β in Eq. 2 to 1.5 instead of 100. As was discussed in Supplementary Results, Study 1, β defines the linearity of the sigmoid function in **Figure 1b**. Decreasing β thus increased the linearity of the slope, meaning that agents often behaved in a manner that was not consistent with their harm aversion parameter, □.

Again, we minimized the possibility that any differences between agents observed in Study 3 could be explained by the order of observations using the methods described in Supplementary Methods, Study 1.

Results

Study 3 replicated all of the findings from Study 2. Again, participants were significantly more uncertain in their beliefs about the bad agent on average ($\overline{\sigma_T}$) relative to the good agent (bad: 0.081±0.001; good: 0.066±0.001; Z = 7.101, p < 0.001; **Supplementary Table 1**). Participants also indicated greater subjective uncertainty in their impression of the bad, relative to good, agent

(bad: 35.609±1.432; good: 30.858±1.665; Z = 3.896, p < 0.001; **Supplementary Table 5**). This translated into faster updating for the bad agent, as demonstrated by a larger ω (bad: -3.468±0.042; good: -3.974±0.043; *Z* = 7.296, p < 0.001; **Supplementary Table 1**).

Study 4: Inferring morality versus competence

Methods

Study 3 demonstrated that the observed asymmetry in learning between good and bad agents is robust to a setting where agents' choices are noisy rather than deterministic. A remaining question is whether the effect is triggered by negative impressions more generally, or specifically by negative impressions about *moral character*. We tested this in a fourth study where participants either inferred on agents' high versus low morality or agents' high versus low competence.

Participants

Two-hundred and eighty U.S. residents were recruited from AMT and randomized to either the morality condition or the competence condition. All participants provided informed consent and were compensated for their time. The study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Thirty-one participants from the morality condition and twenty-nine participants from the competence condition were excluded from the analysis as their behavioral performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 109 participants in the morality condition and 111 participants in the competence condition. We confirm the pattern of results is similar when we include all participants in **Supplementary Table 1** and **5**.

An *a priori* power analysis indicated that the study required 104 participants in each condition to have 80 percent power to detect a moderate effect (0.4) in a nonparametric between-groups analysis. Thus, our study was sufficiently powered to observe an effect in our between-groups design.

Experimental Procedure

Like our previous studies, participants randomized to the morality condition predicted the moral choices of a bad agent ($\kappa = 0.3$) and a good agent ($\kappa = 0.7$). Instead of predicting which of two options was chosen by the agents, in Study 4 participants predicted whether the agents would accept or reject a sequence of offers of a certain amount of money at the expense of a certain number of shocks to the victim (**Figure 3a**). Thus, if the offer is accepted, the agent receives the indicated amount of money and the victim receives the indicated number of shocks. However, if the offer is rejected, the agent receives no money and the victim receives no shocks. Trial sequences were created in a similar manner to Study 2, where we created pairs of trials that were matched in their informational value for the good and bad agent (**Supplementary Figure 4**) and presented minimal differences in learning trajectories for an ideal Bayesian observer. As in Studies 1 and 2, β in Eq. 2 was fixed to 100 to simulate agents that were completely deterministic in their behavior.

In the competence condition, participants predicted whether agents would succeed or fail at scoring a certain number of points in a certain amount of time in a series of basketball games (**Figure 3b**). To manipulate competence, we created agents who differed in their ability to score points. This was parameterized as tau (τ) and represents the agent's skill level. When $\tau = 0$, agents are extremely skilled and can effortlessly score points in minimal amounts of time; as τ approaches 1, agents become weakly skilled and require increasing amounts of time to score a single point. We created the agents to behave identically to the agents in the morality condition, such that one agent had a low basketball skill ($\tau = 0.7$) and the other agent had a high basketball skill ($\tau = 0.3$). Effectively, this meant that the low-skill agent required more time to score each point than the high-skill agent. The trial sequences for the competence condition were identical to the trial sequences for the morality condition. Thus, accepting an offer of \$15.50 in exchange for 4 shocks in the morality condition was analogous to successfully scoring 4 points in 15.50 minutes in the competence condition.

As in the previous studies, on every third trial, participants indicated their general impression of the agent's morality (or basketball skill) on a scale ranging from 0 = nasty (or beginner) to 1 = nice (or expert), and how uncertain they were about each impression.

For the previous studies participants were motivated to learn about the moral character of the agents because they were instructed that they would later have to decide whether to trust the agents in a one-shot trust game that could earn them additional money. Because this motivation was less relevant to participants randomized to the competence condition, we chose to omit this instruction entirely for Study 4. Instead, participants were instructed to learn well about the behavior/performance of the agents because the more accurate their predictions, the more money they could gain. This money was paid out to participants as a bonus after completing the task. The trust game was additionally included at the end of the task to be used as a manipulation check (see **Supplementary Table 7** for results).

We chose to examine competence learning in a task focused on basketball ability, rather than intelligence or social ability, because past work has shown that the latter are not independent from moral character impressions[11,12]. Conversely, we expected moral character impressions to be independent from basketball ability. This was supported in a supplementary pilot study (N=97), where participants rated the moral character and athleticism of two agents who decided how much money they were willing to pay to prevent an anonymous victim from receiving 10 painful electric shocks. One agent (the *bad* agent), indicated that they would require \$4.30, and the other agent (the *good* agent), indicated that they would require \$23.40. The order that the agents were presented was randomized across participants. As expected, participants rated the bad agent as significantly less moral than the good agent (bad: 37.07±2.226; good: 81.47±2.300; $Z = -8.027$, $p < 0.001$), but rated the two agents similarly on athleticism (bad: 43.66±1.792; good: 44.91±1.801; $Z = -0.260$, $p = 0.795$).

Statistical Analyses

The main goal in Study 4 was to investigate whether the observed asymmetry in learning about bad versus good agents was specific to *moral* inference. Consequently, we computed for each

participant the difference between good and bad agents for each of our main dependent measures ($\Delta\overline{\sigma_T}$, $\Delta\omega$, Δ uncertainty rating) and compared this to the difference between low-skill and high-skill agents.

Results

As a manipulation check, we examined whether participants trusted the good agent to a greater extent than the bad agent by comparing the amount participants entrusted to each agent in the trust game. As expected, participants in the morality condition entrusted significantly more with the good agent (good: 7.62±0.326) than with the bad agent (bad: 2.80±0.337; Z = -7.034, $p < 0.001$; **Supplementary Table 7**). Participants in the competence condition entrusted slightly more to the low-skill agent (low-skill: 7.65±0.507) than the high-skill agent (high-skill: 6.03±0.380; Z = 2.967, $p = 0.003$), but the difference in trust for high- versus low-skill agents was significantly smaller than the difference in trust for the good versus bad agents (morality, good - bad: 3.368±0.790; competence, low-skill - high-skill: 0.826±0.291; Z = -3.608, $p < 0.001$).

Next, we investigated whether participants indeed learned through trial-and-error about the agents' moral preferences in the task. We analyzed the model's final estimates about each agent's κ ($\hat{\mu}_1^{50}$) for the morality condition, and verified that participants formed beliefs that closely resembled the agent's true κ (bad: 0.287±0.002; good: 0.715±0.002; Z = -9.062, $p < 0.001$). Final estimates about each agent's τ ($\hat{\mu}_1^{50}$) in the competence condition also verified that participants formed beliefs that closely resembled the agent's true τ (low-skill: 0.714±0.002; high-skill: 0.288±0.002; Z = 9.145, $p < 0.001$; **Supplementary Table 1**). Subjective ratings also confirmed sufficient learning by our participants; final ratings indicated the good agent was generally characterized as nicer than the bad agent (bad: 0.362±0.024; good: 0.770±0.020; Z = -8.091, $p < 0.001$), and the high-skill agent was characterized as more experienced in basketball than the low-skill agent (low-skill: 0.153±0.017; high-skill: 0.787±0.014; Z = -9.113, $p < 0.001$; **Supplementary Table 5**).

A rank sum test found a significant difference in $\Delta\overline{\sigma_T}$ between the morality condition and the competence condition (morality condition: 0.007±0.002; competence condition: -0.002±0.001; Z = 3.334, $p < 0.001$). Simple effects analysis revealed that belief uncertainty, $\overline{\sigma_T}$, was significantly higher for the bad agent, relative to good agent in the morality condition (bad: 0.056±0.002; good: 0.049±0.001; Z = 4.219, $p < 0.001$). No differences in $\overline{\sigma_T}$ were observed between agents in the competence condition (low: 0.049±0.001; high: 0.051±0.001; Z = -0.497, $p = 0.619$; **Supplementary Table 1**).

In line with this result, the magnitude of the difference in subjective uncertainty ratings between agents was significantly greater in the morality condition compared to the competence condition (morality condition: 6.082±1.802; competence condition: -2.129±1.872; Z = 4.118, $p < 0.001$). Simple effects analysis demonstrated that subjective uncertainty was significantly greater for the bad agent, relative to the good agent in the morality condition (bad: 29.335±1.598; good: 24.165±1.607; Z = 3.649, $p < 0.001$). No significant differences in subjective uncertainty were observed between agents in the competence condition (low: 18.457±1.227; high: 20.653±1.274; Z = -1.775, $p = 0.076$; **Supplementary Table 5**).

The magnitude of $\Delta\omega$ was also greater in the morality condition, relative to the competence condition (morality condition: 0.324±0.069; competence condition: 0.060±0.069; $Z = 3.392$, $p < 0.001$). Simple effects analysis demonstrated a higher ω for the bad agent relative to the good agent (bad: -4.390 ±0.064; good: -4.714±0.048; $Z = 4.219$, $p < 0.001$), however there was no significant difference in ω between low- and high-skill agents (low: -4.726±0.047; high: -4.665±0.057; $Z = -0.574$, $p = 0.566$; **Supplementary Table 1**).

Study 5: Inferring bad moral character destabilizes beliefs about competence

Methods

We hypothesized that if asymmetries in learning are driven by inferences about immoral *agents*, rather than immoral *actions*, then other aspects of person perception should destabilize following an inference that an agent is bad. Such a mechanism would be advantageous because it is useful to attend to and learn about all aspects of bad people, in order to build a richer model of those who pose a threat. We tested this hypothesis in a fifth study, where we asked whether inferring an agent's moral character as either good or bad could spill over to influence how people learn about that agent's competence. Study 4 showed that people do not learn differently about two agents who significantly differed in basketball skill in the competence task. Thus, in Study 5 we implemented this same task to test whether we could manipulate learning and uncertainty about competence as a function of inferences about the agent's moral character.

Participants

Two-hundred and fifty-nine U.S. residents were recruited from AMT. Participants provided informed consent and were compensated for their time. The study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Seventy participants were excluded from the analysis as their behavioral performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 189 participants. We confirm the pattern of results is similar when we include all participants in **Supplementary Table 1** and **5**.

Experimental Procedure

Participants predicted both the moral choices and basketball performance of two agents in Study 5. One agent was characteristically low in morality (bad) and the other was high in morality (good), however both agents were similarly competent in their basketball skill. Trial sequences were made up of 60 'morality' trials and 40 'competence' trials. On morality trials, participants predicted whether the agent would accept or reject an offer of a certain amount of money at the expense of a certain number of shocks to an anonymous victim. On competence trials, participants predicted whether the agent would succeed or fail at scoring a certain number of points in a certain amount of time during a basketball game.

Trial sequences were created by interleaving morality trials with competence trials such that (a) participants initially predicted three of the agent's moral choices, and (b) every second or third morality trial would be followed by either one or two competence trials (**Figure 4a**). Across trials we randomized whether competence trials were presented after 2 or 3 morality trials, and whether morality trials were presented after 1 or 2 competence trials. Subjective character and uncertainty ratings were collected following every third morality trial, while subjective competence and uncertainty ratings were collected following every third competence trial.

The morality trial sequences were created using the same procedure as referred to in Supplementary Methods Studies 2 and 4, where one agent was significantly more averse to harming the victim ($\kappa = 0.7$) than the other ($\kappa = 0.3$) with minimal differences in learning trajectories for an optimal Bayesian observer.

Although we wanted the good and bad agents to behave similarly in their basketball performance, we sought to ensure that behavior was not identical in the event that participants could recall the previous agent's performance and thus more easily predict that of the second agent observed. Consequently, we simulated one agent to be slightly less competent ($\kappa = 0.45$) than the other ($\kappa = 0.55$), and randomized across participants which competence simulation was paired with which agent (bad versus good). In other words, for half of the participants, the good agent was slightly less competent, while the bad agent was slightly more competent; for the other half, the good agent was slightly more competent, while the bad agent was slightly less competent.

Competence trial sequences were created in a similar manner to morality trial sequences. We first created a set of 19 trials where the values of τ were randomly drawn from a normal distribution around one agent's indifference point (M = 0.55, s.d. = 0.15). Next, we created a set of 19 matched trials around the other agent's indifference point by subtracting each τ value from 1. Again, we sequentially paired trials that were matched in their informational value for each of the two agents (as in **Supplementary Figure 4**), and randomized the order of presentation of each member of the pair. The pairs comprised trials 2-39 of the sequence, while the initial and final trials were fixed to $\tau = 0.5$.

As in Study 4, participants were instructed to learn well about the behavior/performance of the agents because they would receive a financial bonus in proportion to the accuracy of their predictions. The trust game was additionally included at the end of the task to be used as a manipulation check.

Statistical Analyses

The primary goal for Study 5 was to investigate whether asymmetries in learning are driven by inferences about moral character, or by asymmetries in the choices that good and bad agents make. If the observed learning differences are driven by asymmetries in the choices that morally good and bad agents make, then the effects should be restricted to analysis of the morality trials where agents behave differently. However, if the effects are driven by immoral agents, rather than immoral choices, then learning differences should span across morality and competence trials. Consequently, we performed all analyses separately for morality trials and competence trials. We used two-tailed signed tests to confirm group mean parameter estimates differed significantly between good and bad agents on morality trials (replicating findings from Studies 1-4). A similar analysis restricted to competence trials allowed us to investigate whether we could independently manipulate how people form impressions about an agents' competence as a function of their moral character.

Results

First, we investigated whether participants indeed learned through trial-and-error about the agents' moral preferences in the task. We analyzed the model's final estimates about each agent's κ ($\hat{\mu}_1^{50}$), and verified that participants formed beliefs that closely resembled the agent's true κ (mean±SD bad morality: 0.290±0.028; good morality: 0.707 ± 0.003; bad competence: 0.500 ± 0.061; good competence: 0.499 ± 0.061; **Supplementary Table 1**). Specifically, participants inferred that the bad agent required less money to increase shocks to the victim than the good agent (Z = -11.922, p < 0.001), but both agents would spend similar amounts of time to score additional points in basketball (Z = 0.011, p = 0.991). Participants' beliefs about the agents' character also affected their social behavior, as they entrusted the good agent with more money than the bad agent in the trust game (bad: 2.70±0.241; good: 7.90±0.234; *Z* = -10.112, p < 0.001; **Supplementary Table 7**).

Subjective character ratings also confirmed that the bad agent was characterized as nastier than the good agent (mean±SD; bad: 0.346 ± 0.202; good: 0.741 ± 0.173; Z = -11.755, p < 0.001; **Supplementary Table 5**). However, despite the agents' similar basketball performance, participants rated the bad agent as less skilled than the good agent (bad: 0.426 ± 0.214; good: 0.510 ± 0.237; Z = -3.061, p = 0.002). This is consistent with previous work on the halo effect, where impressions created about one trait spread to other traits[12]. Study 4 found that competence inference was not influenced by impressions of basketball skill. Given that Study 4 demonstrated that people do not learn differently about agents who drastically differ in their basketball competence, the observed difference in competence ratings here was not a concern for subsequent analyses. Thus, we can be confident that any observed between-agent differences in competence inference are unlikely to be attributed to asymmetries in impressions of good and bad agents' basketball skill.

Again, participants were more uncertain in their beliefs about the bad agent's morality, both in the model's uncertainty estimates derived from participant predictions ($\overline{\sigma_T}$, bad: 0.063±0.001; good: 0.055±0.001; Z = 5.055, p < 0.001; **Supplementary Table 1**), and their subjective uncertainty ratings (bad: 27.880±1.019; good: 24.209±1.027; Z = 4.127, p < 0.001; **Supplementary Table 5**). Consequently, beliefs about the bad agent's morality were more volatile than beliefs about the good agent's morality, as demonstrated by a higher ω for the bad agent (bad: -4.116±0.046; good: -4.428±0.039; Z = 5.079, p < 0.001; **Supplementary Table 1**).

Confirming our hypothesis, participants also formed more uncertain beliefs about the bad agent's competence ($\overline{\sigma_T}$, bad: 0.065±0.001; good: 0.062±0.001; Z = 3.075, p = 0.002; **Supplementary Table 1**). This finding was also realized in participants' subjective ratings, where they expressed greater uncertainty in their impression of the bad agent's basketball skill (bad: 28.875 ± 0.955; good: 27.277 ± 0.992; Z = 2.323, p = 0 .020; **Supplementary Table 5**). Thus, it is not surprising that participants formed more volatile beliefs about the bad agent's competence (bad: -4.224 ± 0.039; good: -4.327 ± 0.034; Z = 3.030, p = 0.002; **Supplementary Table 1**), relative to the good agent, as indicated by a higher ω (**Figure 4b**). These findings suggest that our observation of more uncertain and volatile beliefs about the bad agent cannot be attributed to asymmetries in the choices that good and bad agents make.

Study 6: Revising impressions when moral preferences change

Methods

Studies 1 through 5 show that beliefs about the morality of bad agents are more uncertain (and thus more volatile) than beliefs about the morality of good agents. Such results suggest that bad impressions are more rapidly updated than good impressions in the face of new, and potentially inconsistent, evidence. We hypothesized that this may reflect a mechanism by which people could revise their impressions of those who we infer threat by promoting cognitive flexibility in the service of cooperative but cautious behavior. Here, we test this prediction directly using an adapted version of the moral inference task.

Participants

Four-hundred and eight U.S. residents were recruited from AMT and 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. Participants provided informed consent and were compensated for their time. The study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Forty-four participants were excluded from the analysis as their behavioral performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 364 participants. We confirm the pattern of results is similar when we include all participants in **Supplementary Table 1** and **5**.

An *a priori* power analysis indicated that the study required 360 participants to have 80 percent power to detect a small to medium interaction effect ($f = 0.175$) in an ANOVA. Thus, our study was sufficiently powered to observe an effect in our between-groups design. We pre-registered our sample size, experimental design, and planned analyses on the Open Science Framework (https://osf.io/5s23d/).

Experimental Procedure

Participants completed a modified version of the moral inference task. In the task, participants predicted a sequence of 36 choices made by a single agent, and on each trial received immediate feedback about their accuracy. Every few trials, participants rated their impression of the agent's moral character and how certain they were about their impression. The study comprised a 2x2 factorial design with moral character (bad versus good) and shift direction (improve versus worsen) as between-subject independent variables.

Moral character: Between subjects we manipulated the moral character of the agent that participants observed (bad versus good). To manipulate moral character, we created agents with different preferences towards harming the victim, similar to our previous studies (bad agent: $\kappa = 0.3$; good agent: $\kappa = 0.7$). For the first 30 trials (phase 1) participants observed the two agents make choices for identical trial sequences. On every trial, the agents faced the same two options, but

because the agents had different preferences towards harming the victim, they often chose differently. We created the sequence of 30 trials using similar methods to those reported in Supplementary Methods, Study 1, and simulated how the agent chose using Equations 1-3. In phase 1, we asked participants to provide subjective character and certainty ratings every 1-3 trials, for a total of 15 ratings.

Shift direction: Because we were interested in how participants update their impressions when an agent's behavior becomes inconsistent with prior evidence, we manipulated the agents' preferences on the final 6 trials of the moral inference task. For half of the participants, the agent became more moral than previously observed in the first 30 trials (improve condition) and for the other half the agent became less moral than previously observed in the first 30 trials (worsen condition). In the improve condition, agents became more harm-averse, and therefore required more money to inflict pain than previously (κ+0.2). In the worsen condition, agents became less harm-averse, and therefore required less money to inflict pain than previously (κ-0.2).

For the final 6 trials (phase 2), participants observed the agents make choices that were inconsistent with their previous preferences. Thus, in the improve condition, agents made prosocial choices where they would have previously chosen antisocially. In the worsen condition, agents made antisocial choices where they would have previously chosen prosocially. Together, this resulted in four conditions, manipulated between subjects: 1) bad agent becomes more moral (bad-improve, $\kappa = 0.3 \rightarrow \kappa = 0.5$), 2) bad agent becomes less moral (bad-worsen, $\kappa = 0.3 \rightarrow \kappa = 0.1$), 3) good agent becomes more moral (good-improve, $\kappa = 0.7 \rightarrow \kappa = 0.9$), and 4) good agent becomes less moral (good-worsen, $\kappa = 0.7 \rightarrow \kappa = 0.5$). In phase 2, we asked participants to provide subjective character and certainty ratings every second trial, for a total of 3 ratings.

In order to minimize the potential influence of prior expectations on participant predictions, we anchored prior expectations through explicit instruction. Specifically, we told participants that on average, people required \$1 per shock to the victim. This prior expectation maps on to $\kappa = 0.5$ (i.e., equidistant from the initial preferences of the good and bad agents).

Statistical Analyses

The primary goal of Study 6 was to investigate whether participants more rapidly update their impressions of bad agents than good agents, particularly when agents show moral improvement. Consequently, we computed the magnitude that participants' impressions updated from phase 1 to phase 2. The update was defined as the difference between participants' phase 2 and phase 1 ratings (update = phase 2 – phase 1). For phase 1 ratings we took the average of the final 3 ratings in phase 1, and for phase 2 ratings we took the average of the 3 ratings in phase 2. We preregistered this definition of the update prior to collecting data (https://osf.io/5s23d//). We conducted a 2 (agent: bad versus good) x 2 (shift direction: improve versus worsen) ANOVA to obtain main effects and interaction effects. Because our dependent measure was not normally distributed we split the data to complement the ANOVA with non-parametric statistics.

Results

First, we investigated whether our results from phase 1 of the task replicated our previous findings from Studies 1-5. Again, participants were significantly more uncertain in their beliefs about the bad agent on average ($\overline{\sigma_T}$) relative to the good agent (bad: 0.076±0.002; good: 0.063±0.001; Z = 6.680, p < 0.001; **Supplementary Table 1**). Participants also indicated greater subjective uncertainty in their impression of the bad, relative to the good, agent (bad: 33.584±1.164; good: 27.945±1.347; Z = 4.362, p < 0.001; **Supplementary Table 5**). This translated into faster updating for the bad agent, as demonstrated by a larger ω (bad: -3.559±0.042; good: -3.928±0.034; Z = 6.577, p < 0.001; **Supplementary Table 1**).

Next, we investigated impression updates following the agents' shift in behavior. As predicted, we observed a main effect of agent on updating, where participants updated their character ratings more for bad agents than good agents (bad: 18.951±1.245; good: 14.928±1.316; F(1,360)=5.124, P=0.024; Z=3.541, P<0.001, **Figure 4d** and **Supplementary Figure 6**). In line with past work showing a negativity bias in impression formation, we also observed a main effect of shift direction, where participants updated their character ratings more when morality worsened than when it improved (worsen: 22.083 ±1.389; improve: 11.468 ±1.010; F(1,360)=37.698, P<0.001; Z=6.372, P<0.001). Finally, there was an interaction between agent and shift direction (F(1,360)=6.803, P=0.009; Chi-squared=57.227, P<0.001), where asymmetric updating was more pronounced when morality improved.

As a secondary analysis, we compared uncertainty before versus after the shift in our 2x2 factorial design. Post-change, for bad agents, uncertainty remained high, regardless of whether the agent's morality improved or worsened (**Supplementary Figures 6** and **7**; improved: Z=0.040, P=0.968; worsened: Z=1.233, P=0.218). However, for good agents, uncertainty increased when morality worsened (Z=-5.507, P<0.001), and decreased when morality improved (Z=2.252, P=0.024). The more uncertain participants became about the good agent whose morality worsened after the shift, relative to before, the more they updated their impression about that agent (Spearman's ρ =0.420, P<0.001).

Study 7: Supplementary study verifying the observed asymmetry does not depend on specific labels used for character ratings

Methods

Participants

One-hundred and twenty-five U.S. residents were recruited from AMT. All participants provided informed consent and were compensated for their time. Study 7 was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Nine participants were excluded from the analysis as their behavioral performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 116 participants. We confirm the pattern of results are similar when we include all participants in **Supplementary Table 1** and **5**.

Experimental Procedure

In general, the experimental procedure for Study 7 was very similar to Studies 1 and 2 (including the instructions). In this study, we tested whether any between-agent effects related to the specific labels that were used to rate the agent's moral character. Consequently, participants indicated their general impressions of the agents on a scale ranging from *bad* (0) to *good* (1) as opposed to the scale from the previous studies which ranged from *nasty* to *nice*. All other aspects of the task were identical to those set out in Supplementary Methods study 1 and 2.

Results

Study 7 replicated all of the findings from our previous studies showing an asymmetry in learning about bad versus good agents. Participants were significantly more uncertain in their beliefs about the bad agent on average ($\overline{\sigma_T}$) relative to the good agent (bad: 0.058±0.002; good: 0.052±0.001; Z = 4.262, $p < 0.001$; **Supplementary Table 1**). Participants also indicated greater subjective uncertainty in their impression of the bad, relative to good, agent (bad: 29.209±1.485; good: 24.602±1.474; Z = 3.207, p = 0.001; **Supplementary Table 5**). This translated into faster updating for the bad agent, as demonstrated by a larger ω (bad: -4.303 ±0.064; good: -4.608 ±0.060; Z = 4.16, $p < 0.001$; **Supplementary Table 1**). Thus, we can be certain that our findings are not dependent on the specific labels that were used on the scale to rate the agents' moral character.

Study 8: Moral character or moral expectations?

A reasonable explanation for why people form more uncertain and volatile beliefs about the moral character of bad agents, relative to good agents, is that people generally expect others to be 'good'. If people have a strong prior expectation that the agents will behave morally, or more like the good agents in our studies, then the bad agents' behaviors will be more surprising. To investigate this hypothesis, we recruited participants in an independent study that investigated how people expect others to behave when faced with the same moral decisions our agents faced in our previous experiments.

Methods

Participants

Thirty U.S. residents were recruited from AMT to participate in a prediction task. All participants provided informed consent and were compensated for their time. Study 8 was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098).

Experimental Procedure

Participants were fully briefed about our previous experiments where two participants arrive at the laboratory, and one of them makes decisions about whether to profit by inflicting shocks on the other. After observing an example trial, participants were asked to indicate how they think most people decided in our previous experiments. Specifically, we asked them to predict which option was most commonly chosen by our participants, for a set of 34 trials. Feedback was not provided throughout the task. Crucially, we incentivized participants to be as accurate as possible in their predictions, because they would be rewarded financially for every choice for which they successfully guessed the majority response.

We modelled participants' predictions using the same decision model that was used to simulate agent choices (**Eq. 1-3**), and extracted how harm averse they expected most people would be in this task, κ_e.

$$V_{harm} = (1 - \kappa_e)\Delta m - \kappa_e \Delta s \quad \textbf{(12)}$$

Results

The priors hypothesis predicts that people will expect others' harm aversion, parametrized as κ_e, to be significantly greater than 0.5. This would demonstrate that people expect others to behave more similarly to the good agent than the bad agent, rendering the bad agent's choices in our task

less expected. In fact, our study reveals that participants expect others to behave slightly more similarly to the bad agent (κ_e = 0.445±0.043) though a one-sample Wilcoxon signed-rank test revealed that this was not significantly different from κ = 0.5 (Z = -1.347, p = 0.178). This study provides vital evidence that, at least within the context of our task, participants do not expect others to behave more similarly to the good agent than the bad agent.

Below, we outline additional results that do not support the hypothesis that asymmetries in learning about good versus bad agents are driven by prior expectations that people will be good.

Additional support against moral expectations

Relationship between subjective priors and behavior:

In Studies 2-7 we asked participants to indicate how nasty or nice they *expect* agents will be, prior to observing either of the agents' choices. If prior expectations that agents will be 'good' increase uncertainty and volatility in beliefs about the bad agent, relative to the good, then we would expect to see greater between-agent differences the nicer participants think agents will be. To investigate, we computed for each participant the difference in belief volatility, ω, between good and bad agents ($\Delta\omega$) and checked for correlations with subjective prior ratings. This analysis was conducted specifically for studies that include a within-subject manipulation of moral character (i.e., bad morality versus good morality).

Across studies we found no consistent relationship between prior beliefs and either of these dependent measures (see **Supplementary Table 9** and **Supplementary Figure 2**). No study showed a significant positive relationship between prior ratings and $\Delta\omega$, as would be predicted by the 'priors hypothesis'. Nonetheless, to investigate the possibility that a sub-threshold relationship really does exist, we conducted a mini meta-analysis on the correlations across all studies that include a within-subject manipulation of moral character. Again, we found no evidence for a relationship between subjective prior expectations and the difference in belief volatility between agents (Z = -0.846, p = 0.398).

Relationship between moral preferences and behavior:

Previous work suggests that people's expectations about others' preferences is related to their own preferences[15]. Consequently, it's possible that people will expect others to perform similarly to how they choose in the task. Prior to observing the agents' choices in Studies 2 and 3, participants indicated how *they* would decide if they were faced with similar decisions to profit from harming an anonymous person. Specifically, participants made a series of 20 hypothetical decisions that involved choosing between less money for themselves plus less shocks for an anonymous person, or more money for themselves at the expense of more shocks for that person. We then fit the same decision model in **Eq. 1-3** to estimate participant's own harm aversion parameter ($\kappa_{subject}$).

$$V_{harm} = (1 - \kappa_{subject})\Delta m - \kappa_{subject}\Delta s \quad \textbf{(13)}$$

If people expect others to have similar preferences to their own, then we would expect participants whose preferences are nicer (i.e., larger values of $\kappa_{subject}$) to show greater between-agent differences in our task, under the priors hypothesis. Participants had an average $\kappa_{subject}$ = 0.445±0.022 in Study 2 and an average $\kappa_{subject}$ = 0.443±0.026 in Study 3. Given that people's own behavior more closely resembles that of the bad agent than the good agent, it is unlikely that the bad agent's behavior is more surprising than the good agent's. In a correlational analysis, we find the opposite relationship between participants' own preferences and $\Delta\omega$ than would be predicted by the priors hypothesis. The nicer participants were, the smaller the effect of character on belief volatility (Study 2: ρ = -0.355, P<0.001; Study 3: ρ = -0.154, P=0.076).

Relationship between generalized trust and behavior:

A prior expectation that people are generally morally good is likely related to beliefs about others' trustworthiness: the greater the expectation that people will be good, the more likely you are to believe others are trustworthy. In Study 1 we asked participants in a pre-testing questionnaire "To what extent do you feel you can trust other people that you interact with in your daily life?". Participants responded on a scale ranging from 1 (*very little*) to 7 (*very much*). The priors hypothesis predicts that the more people generally believe that others are trustworthy, the more volatile beliefs will be for the bad agent relative to the good agent (larger $\Delta\omega$). In fact, we found no relationship between general trust and $\Delta\omega$ (ρ = -0.178, p = 0.300; **Supplementary Figure 3**).

In studies 2 and 3 participants completed a generalized trust scale, consisting of 6 items related to general beliefs about the trustworthiness and kindness of others[16]. For example, "Most people are basically good and kind" and "Most people are trustworthy". Items were rated on a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) and summed for a single measure of 'generalized trust'. Again, we found no significant relationship between general trust and $\Delta\omega$ in Study 1 or 2 (Study 1: ρ = 0.065, p = 0.413; Study 2: ρ = 0.020, p = 0.817; **Supplementary Figure 3**).

Prior expectations of basketball competence versus morality:

Another objection to the priors hypothesis is that prior expectations about morality and basketball competence were very similar (morality = 56.339±1.845; competence = 54.580; Z = 0.678, p = 0.498), yet between-agent differences in belief volatility were restricted to the morality condition. However, prior expectations about skill may have been weaker than those about morality. Thus, this does not rule out the possibility that prior expectations may influence behavior in the morality conditions but not in the competence condition. Consequently, we checked whether participants expressed greater uncertainty in their explicitly stated prior beliefs about an agent's basketball skill relative to moral character. We performed this analysis first for our between-subject design, Study 4, where participants either indicated their certainty about how skilled or how moral they expected an agent would be. We performed a similar analysis for our within-subject design, Study 5, where

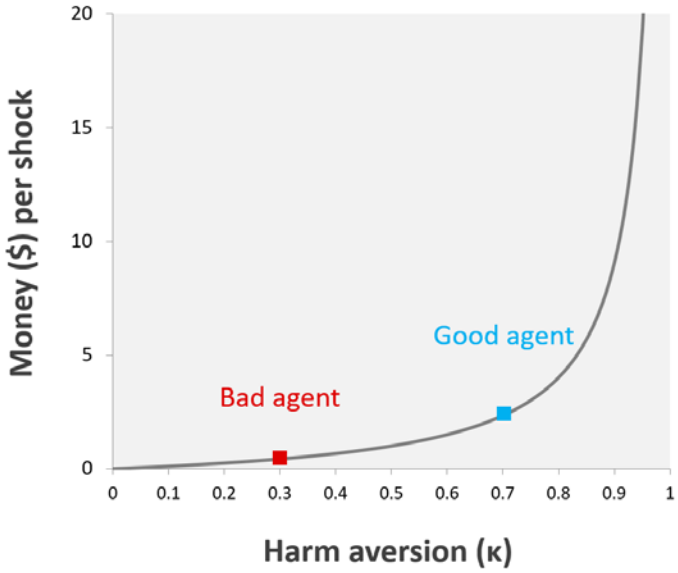
participants indicated their certainty in their expectations about an agent's morality and skill. In both studies, we found no significant differences in how certain participants were in their prior expectations about an agent's basketball skill and morality (Study 4: morality=65.092±2.355, competence=69.089±2.595, Z=-1.536, P=0.125; Study 5: morality=56.079±1.845 , competence=57.217±1.878, Z=-1.246, P=0.213).

Two other pieces of evidence in our previous experiments argue against a priors hypothesis. First, a prior expectation that people will behave morally cannot explain why inferring a bad moral character destabilizes beliefs about basketball competence in Study 5. Second, in Study 6 we anchored participants to expect that most people require $1 per shock to the anonymous victim, consistent with a prior belief of $\kappa = 0.5$. Yet we still find that beliefs about bad agents were more uncertain and volatile than beliefs about good agents. Together, this evidence does not support the hypothesis that asymmetries in learning result from prior expectations that people will be moral. While we find no evidence for the priors hypothesis, future work should investigate the alternative hypothesis - whether inferences about threat destabilize beliefs about agents - through a direct manipulation of perceived threat while holding behavior constant.

Supplementary Figures

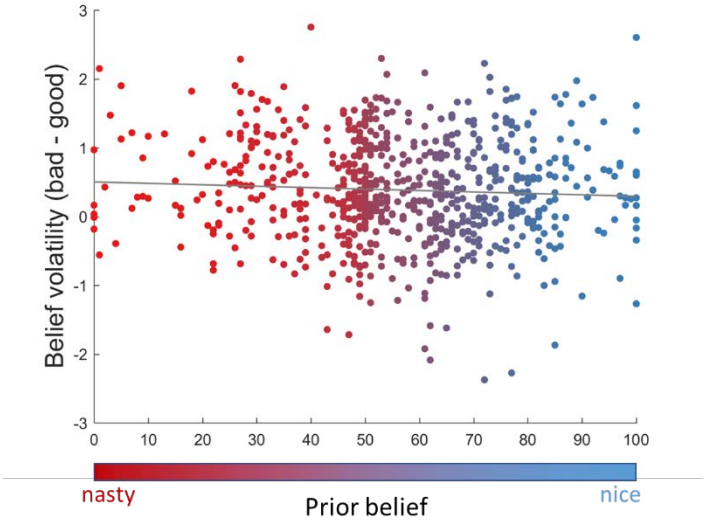
Supplementary Figure 1.

Relationship between the harm-aversion parameter, κ, and the amount of money agents were willing to accept per additional shock. Money per shock is plotted against the harm aversion parameter, κ, which can range from 0 to 1. The bad agent requires less money per shock than the good agent.



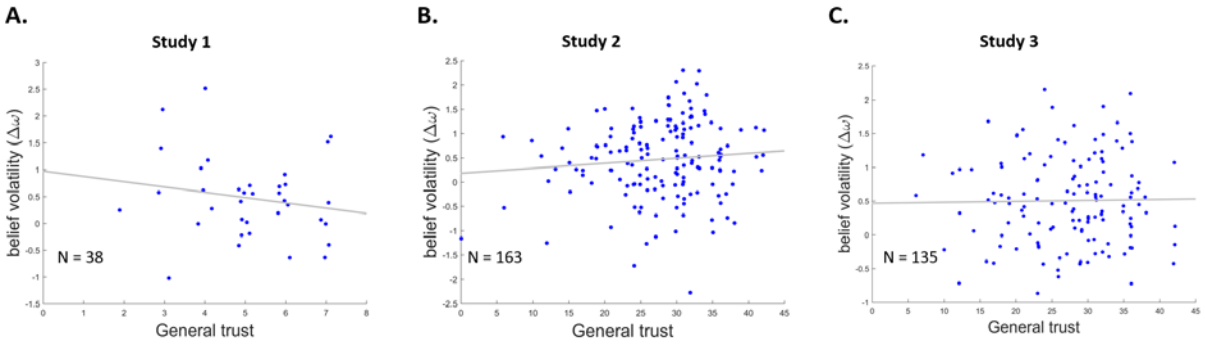
Supplementary Figure 2.

Relationship between prior beliefs and the difference in volatility estimates between agents (Δω) across studies 2, 3, 4 (morality condition), 5, and 7.



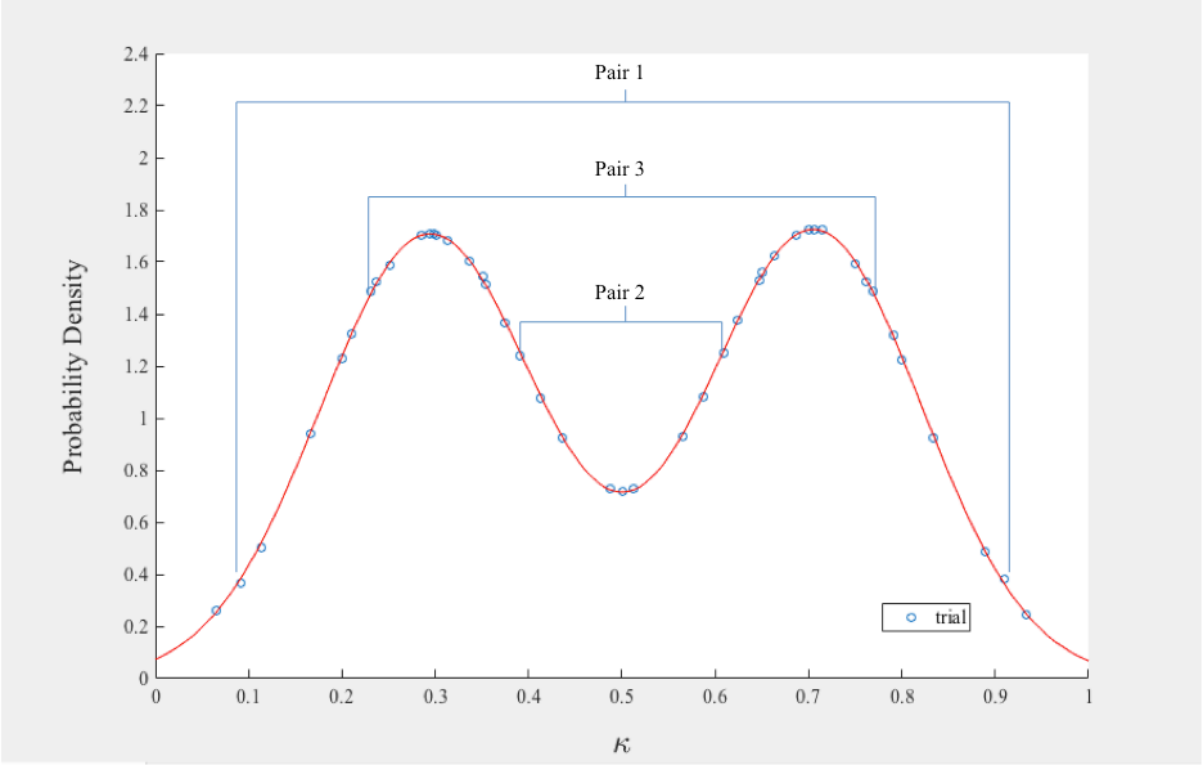
Supplementary Figure 3.

Relationship between general trust and volatility difference between agents (Δω, higher values represent larger volatility for bad agent relative to good agent) for study 1. (A) ρ = -0.178, p = 0.300, study 2 (B) ρ = 0.065, p = 0.413 and study 3 (C) ρ = 0.020, p = 0.817.



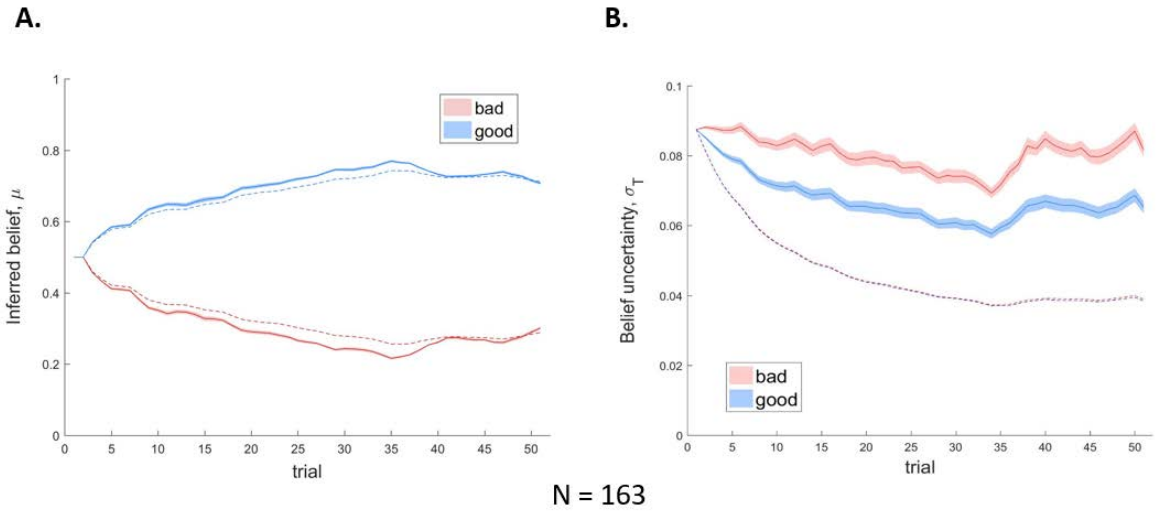
Supplementary Figure 4.

Graphical depiction of the optimized trial sequence. To minimize the possibility that differences between agents could be attributed to the order of observations we created pairs of trials that were matched in informational value for the good and bad agent. Each pair comprised trials with mirrored κ values: one member of the pair was randomly drawn from a normal distribution around the good agent's indifference point, and the other member was the mirrored deviation from the bad agent's indifference point (1 – κ). This resulted in a bimodal distribution of trials and ensured that a trial that was highly informative about one agent was sequentially paired with a trial that was equally informative about the other agent.



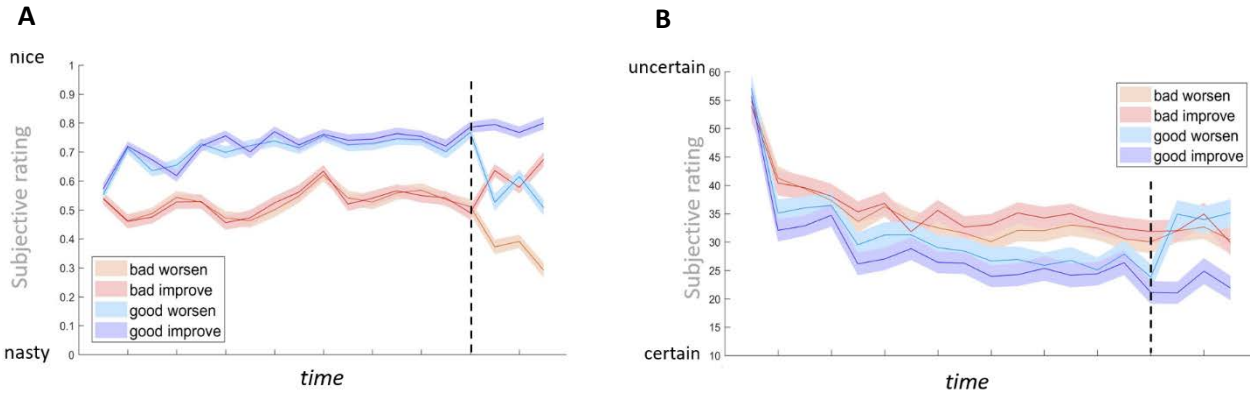
Supplementary Figure 5.

(**A**) Trajectory of model estimates of inferred beliefs ($\hat{\mu}$) about each agent's κ for each trial, averaged across all participants in Study 2 (solid lines) and for an optimal Bayesian learner (dotted lines). (**B**) Trajectory of belief uncertainty estimated by the model averaged across participants in Study 2 (solid lines) shows that participants are more uncertain throughout the experiment for the bad agent. Participants are more uncertain for both agents than for the optimal Bayesian learner (dotted lines), which has an identical trajectory for good and bad agents due to the symmetric task design.



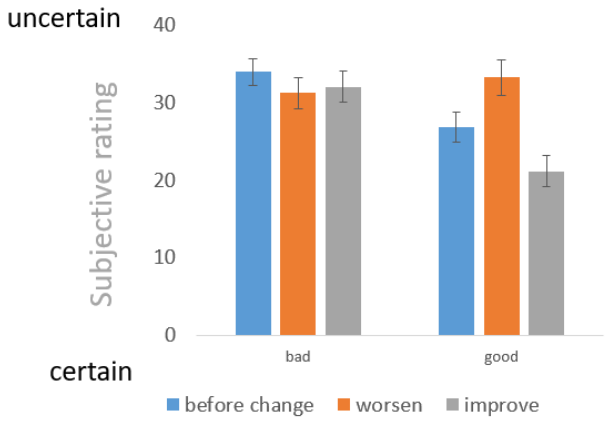
Supplementary Figure 6.

Temporal evolution of subjective character ratings (A) and uncertainty ratings (B) from Study 6. Dashed line represents the point at which the agent's behavior worsened or improved. In phase 1 (to the right of the dotted line) ratings were made every 1-3 trials, for a total of 15 ratings. In phase 2 (to the left of the dotted line) ratings were made every second trial.



Supplementary Figure 7.

Uncertainty before and after preferences shifted, Study 6. Before the agents' preferences shifted, participants indicated greater subjective uncertainty about their character impressions of the bad agent relative to the good agent. Uncertainty remained high after preferences shifted for the bad agents. For good agents, uncertainty increased when morality worsened and decreased when morality improved.



Supplementary Tables

Supplementary Table 1 provided separately here.

Supplementary Table 2.

Details of each learning model used.

Model	Notes	Estimated parameters
1 Learning rate Rescorla Wagner	Beliefs are symmetrically updated, with a single learning rate for each participant.	α = Learning rate β = Prediction noise
2 Learning rate Rescorla Wagner	Beliefs are asymmetrically updated, with separate learning rates for positive versus negative outcomes, for each participant.	α_{pos} = Learning rate positive outcomes α_{pos} = Learning rate negative outcomes β = Prediction noise
HGF	A two level model, with one estimated parameter governing the volatility of beliefs at the second level, and a second estimated parameter governing the prediction noise.	ω = Tonic volatility β = Prediction noise

Supplementary Table 3.

Magnitude of volatility difference; comparison between subjects and an optimal Bayesian learner.

Variable	Study	Subject or Bayesian	mean ± SEM	Test Statistic	p-value	effect size (r)
Belief volatility (Δω)	Study 1	Subject	0.433 ± 0.121	3.200	0.001	0.519
		Bayesian	0.015 ± 0.011			
	Study 2	Subject	0.446 ± 0.061	7.382	<0.001	0.578
		Bayesian	0.021 ± 0.012			
	Study 3	Subject	0.506 ± 0.056	7.659	<0.001	0.659
		Bayesian	-0.007 ± 0.002			
	Study 4	Subject morality	0.316 ± 0.069	4.579	<0.001	0.439
		Bayesian morality	-0.049 ± 0.015			
		Subject competence	-0.060 ± 0.069	-1.266	0.206	N.S.
		Bayesian competence	0.045 ± 0.015			
	Study 5	Subject morality	0.313± 0.059	4.558	<0.001	0.332
		Bayesian morality	0.042 ± 0.002			
		Subject competence	0.103 ± 0.042	3.277	0.001	0.238
		Bayesian competence	0.001 ± 0.003			
	Study 7	Subject	0.304± 0.067	5.090	<0.001	0.473
		Bayesian	-0.037 ± 0.015			

Supplementary Table 4.

Model accuracy (%)

		bad	bad s.d.	good	good s.d.	average
Study 1*		87.00	4.842	86.29	5.375	86.65
Study 2		75.00	7.996	68.70	9.561	71.85
Study 3		69.40	7.521	64.50	9.257	66.95
Study 4	morality	77.59	7.374	78.92	8.722	78.26
	competence	80.68	9.659	77.94	6.877	79.31
Study 5	morality	77.70	6.976	79.10	9.014	78.40
	competence	76.30	9.496	76.50	9.412	76.40
Study 6		73.44	10.971	66.59	12.309	70.02
Study 7		77.31	7.726	77.76	9.760	77.54
Average		77.16		75.14		**76.15**

* Study 1 was conducted in the laboratory and included participants recruited from Oxford's Psychology Research recruitment scheme. All subsequent studies were conducted online, with participants recruited from Amazon's Mechanical Turk (MTurk). Because MTurk studies typically feature a larger amount of noise, the model is more accurate for Study 1 than subsequent studies.

Supplementary Table 5 provided separately here.

Supplementary Table 6.

Linear regression investigating the effects of moral character (agent) on uncertainty controlling for time.

Study 1

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	32.667	1.2419	26.304	8.2267e-122
time	-1.3086	0.10935	-11.968	2.4724e-31
bad_agent	8.5201	1.0714	7.9525	4.0491e-15

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.068019	0.00062682	108.51	0
time	-0.00037405	1.9192e-05	-19.49	8.2979e-81
bad_agent	0.011129	0.0005539	20.092	1.6079e-85

Study 2

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	31.827	0.65658	48.474	0
time	-0.86005	0.057809	-14.878	4.053e-49
bad_agent	8.7528	0.56641	15.453	8.958e-53

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.072005	0.00039184	183.76	0
time	-0.00019591	1.1997e-05	-16.329	1.8074e-59
bad_agent	0.013443	0.00034626	38.823	1.9794e-315

Study 3

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	39.755	0.78017	50.956	0
time	-0.98857	0.068691	-14.392	5.7759e-46
bad_Agent	4.4793	0.67303	6.6555	3.1558e-11

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.073087	0.00031938	228.84	0
time	-0.00026265	9.7785e-06	-26.86	7.4073e-155
bad_agent	0.014283	0.00028222	50.607	0

Study 4

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	30.236	0.57226	52.837	0
time	-0.90032	0.043558	-20.669	1.9249e-92
moral_condition	2.0323	0.60426	3.3634	0.00077372
low_agent	-2.7833	0.58957	-4.721	2.3883e-06
moral_condition:low_agent	7.9522	0.85455	9.3057	1.6994e-20

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	29.04	0.57677	50.35	0
time	-0.31132	0.014551	-21.395	1.2996e-98
morality_condition	3.5128	0.60103	5.8446	5.2908e-09
low_agent	-2.1961	0.59829	-3.6706	0.00024369
morality_condition:low_agent	7.365	0.84998	8.6649	5.4793e-18

Study 5 – morality trials

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	34.923	0.48325	72.267	0
time	-0.33428	0.012116	-27.589	1.2027e-159
bad_agent	3.6545	0.4192	8.718	3.4481e-18

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.062677	0.00024774	252.99	0
time	-0.00026206	6.3334e-06	-41.378	0
bad_agent	0.0077317	0.00021937	35.246	6.519e-265

Study 5 – competence trials

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	38.231	0.55317	69.112	0
time	-0.52409	0.02072	-25.294	2.865e-133
bad_agent	1.4897	0.50115	2.9725	0.0029671

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.072121	0.00024859	290.12	0
time	-0.00047413	9.4859e-06	-49.982	0
bad_agent	0.0027261	0.000219	12.448	2.1445e-35

Study 6

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	66.698	0.61734	108.04	0
time	0.64645	0.061215	10.56	7.5942e-26
bad_agent	-6.2111	0.52898	-11.741	1.6979e-31

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.076009	0.00038911	195.34	0
time	-0.00040903	1.9664e-05	-20.801	1.8021e-94
bad_agent	0.010275	0.00034042	30.184	4.1468e-193

Study 7

Subjective uncertainty rating

	Estimate	SE	tStat	pValue
(Intercept)	31.183	0.71088	43.866	0
time	-0.2386	0.020942	-11.394	1.3136e-29
bad_agent	4.5381	0.61161	7.42	1.4294e-13

σ_T

	Estimate	SE	tStat	pValue
(Intercept)	0.065468	0.00035494	184.45	0
time	-0.00050919	1.0867e-05	-46.855	0
bad_agent	0.0057238	0.00031365	18.249	2.277e-73

Supplementary Table 7.

Trust Game & agent comparison.

	Amount entrusted (mean ± SEM)		*Z*-statistic	p-value	effect size (*r*)
Study 2	bad:	3.82 ± 0.265	-8.522	<0.001	0.667
	good:	7.74 ± 0.243			
Study 3	bad:	3.36 ± 0.303	-7.589	<0.001	0.653
	good:	7.15 ± 0.295			
Study 4	bad:	2.80 ± 0.337	-7.034	<0.001	0.674
	good:	7.72 ± 0.326			
	low-skill:	6.02 ± 0.380	2.967	0.003	0.280
	high skill:	5.19 ± 0.381			
Study 5	bad:	2.70 ± 0.241	-10.112	<0.001	0.736
	good:	7.90 ± 0.234			
Study 6	bad:	6.59 ± 0.259	-2.055	0.040	0.161
	good:	7.32 ± 0.248			
Study 7	bad	3.79 ± 0.354	-5.957	<0.001	0.520
	good	6.97 ± 0.327			

Supplementary Table 8.

Trial-wise updating of character ratings; comparison between good and bad agents.

Study	Bad agent		Good agent		test-statistic	p-value
	mean	SEM	mean	SEM		
Study 1	7.369	0.655	6.880	0.759	0.558	0.577
Study 2	9.780	0.602	7.909	0.452	2.787	0.005
Study 3	10.103	0.610	9.577	0.522	1.140	0.254
Study 4, morality	8.933	0.679	6.840	0.453	2.579	0.010
Study 4, competence	5.818	0.477	6.707	0.553	-2.299	0.021
Study 5, morality	7.183	0.367	6.477	0.358	2.433	0.015
Study 5, competence	6.461	0.229	5.348	0.240	5.407	0.000
Study 6*	12.445	0.628	10.615	0.602	2.460	0.014
Study 7	7.821	0.478	5.539	0.472	5.677	0.000

*Between-subjects

While we did not find larger impression updating for the bad agent relative to the good agent in studies 1 and 3, we see the same pattern of effects. A parametric meta-analysis, including the results from all within-subject studies (thus, excluding Study 6 and the competence condition from study 4) yielded significant results ($Z = 7.461$, $p < 0.001$).

Supplementary Table 9.

Correlation between prior trait rating and the difference in belief volatility (Δω) between good and bad agents.

		Prior trait rating* (mean ± SEM)	Δω correlation ρ (p-value)
Study 1		n.a	
Study 2		48.724 ± 1.610	-0.179 (0.022)
Study 3		52.007 ± 1.934	-0.029 (0.738)
Study 4	morality	56.954 ± 1.841	0.175 (0.068)
	competence	54.955 ± 1.404	0.106 (0.268)
Study 5	morality	61.355 ± 1.341	0.020 (0.788)
	competence	56.735 ± 1.192	0.033 (0.610)
Study 7		60.621 ± 1.750	0.006 (0.952)

*prior trait ratings are collected prior to observing any outcomes. For the morality task, participants are asked to indicate how nasty or nice they expect the agent to be on a scale ranging from 0 = *nasty* to 100 = *nice*. For the competence task, participants are asked to indicate how skilled they expect the agent to be in basketball on a scale ranging from 0 = *beginner* to 100 = *expert*. Study 6 was omitted from this analysis because this was a between-subjects, rather than within-subjects, design. Thus, we cannot compute Δω for Study 6.

Supplementary Table 10.

Prior mean and variance of the perceptual and response model parameters.

Parameter	Notes		mean	variance
ω	Constant component of the tonic volatility at the second level. Represents the temporal evolution of x_2. *Estimated in native space.*		-4	1
Predictions (x_1)	Predictions are a sigmoid transformation of x_2, and so do not have prior values.	$\mu_{1:}$	none	none
		$\sigma_{1:}$	none	none
Probabilities (x_2)	The prior mean on x_2 (prior belief about agent's harm-aversion, κ) was fixed to a neutral point that was equidistant from the true κ value of both agents. Estimated in logit space.	$\mu_{2:}$	0.5	0
	The prior variance on x_2 was fixed to ensure that any differences in learning about good and bad agents derived from the model could not result from differences in the prior estimates. Estimated in log-space.	$\sigma_{2:}$	0.35	0
β	Constant component that describes how sensitive prior beliefs are to the relative utility of different outcomes, or the prediction noise. Estimated in log-space.		1	1

Supplementary Table 11.

Model comparison. Sum log-model evidence (LME) for each study.

	HGF	1 learning rate RW	2 learning rate RW
Study 1	-1662.70	-1855.68	-1905.96
Study 2	-7599.38	-7849.20	-7759.33
Study 3	-7291.61	-7543.39	-7099.73
Study 4 morality	-4050.18	-4820.50	-5230.21
Study 4 competence	-3983.84	-5315.96	-5791.98
Study 5 morality	-8149.24	-9252.11	-9916.40
Study 5 competence	-6021.07	-7167.02	-8085.64
Study 6	-5361.11	-5567.04	-5568.75
Study 7	-4375.62	-4859.85	-5250.32
Total	**-48494.75**	**-54230.75**	**-56608.30**

Supplementary Table 12.

Correlation between model free parameters ω and β. Analysis investigating (a) the relationship between ω and β for the bad agent, (b) the relationship between ω and β for the good agent, (c) the relationship between Δω and Δβ between good and bad agents

	Bad agent		**Good agent**		**Bad - Good**	
Study	ρ	**P**	ρ	**P**	ρ	**P**
Study 1	0.002	0.992	-0.008	0.960	0.315	0.055
Study 2	0.272	0.000	0.316	0.000	0.472	0.000
Study 3	0.288	0.001	0.156	0.072	0.338	0.000
Study 4, morality	-0.292	0.002	-0.434	0.000	-0.229	0.017
Study 4, competence	-0.331	0.000	-0.408	0.000	-0.171	0.071
Study 5, morality	-0.131	0.073	-0.039	0.596	0.098	0.180
Study 5, competence	-0.288	0.000	-0.253	0.000	0.016	0.830
Study 6*	0.363	0.000	0.118	0.117	n/a	n/a
Study 7, morality condition	-0.175	0.061	-0.507	0.000	-0.039	0.673

*Because Study 6 is between-subject and we did not have a within-subject manipulation of agent, we cannot compute the difference between agents (Δω and Δβ).