I judge you by your profit: Judgments of effort exerted by self and others are influenced by task contingent rewards

Max Rollwage*[1,2,3], Franziska Pannach[3], Ulf Toelch[4], Igor Kagan**[6,7], Arezoo Pooresmaeili**[3,6]

[1] Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

[2] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London, United Kingdom

[3] Perception and Cognition Group, European Neuroscience Institute, University Medical Center Göttingen, Göttingen, Germany

[4] Biological Psychology and Cognitive Neuroscience, Freie Universität Berlin, Berlin, Germany

[5] Berlin Institute of Health, Berlin, Germany

[6] Decision and Awareness Group, German Primate Center (DPZ), Göttingen, Germany

[7] Leibnitz Science Campus "Primate Cognition", Göttingen, Germany

* Correspondences should be addressed to:

Max Rollwage

max.rollwage.16@ucl.ac.uk

** IK and AP are joint last authors

Abstract

Estimating the effort someone has spent on a task is a core dimension for evaluating own and other's actions. Interestingly, it has been shown that self-judgments of effort are influenced by the magnitude of obtained rewards (Pooresmaeili et al., 2015). However, it is unclear whether the influence of reward on effort estimations is limited to self-judgments (i.e. a form of self-serving bias) or whether reward information is invariably incorporated when judging effort, thus extending also to social contexts. Here we show that people also integrate reward magnitude when judging the effort exerted by another person. Participants (N=48, 24 pairs) performed an effortful task interleaved with a partner. In half of the trials, participants performed the task themselves and rated their own effort. In the other half, participants rated the other person's effort after watching them performing the task. After each trial but before the effort rating, both participants were informed of the gained reward for that trial. Our results show that higher rewards led to higher estimations of exerted effort. Importantly, this was true for self- as well as other-judgments. In line with a Bayesian cue integration framework, reward magnitude had a stronger effect when judging the effort of another person, compared to self-judgments, since people should have more reliable internal estimates of own effort. Supporting this claim, computational modelling revealed that reward information and objective performance criteria related to the exertion level were combined in a Bayes optimal manner to form effort estimates. Interestingly, the extent to which rewards influenced effort judgments was positively correlated with conservative world-views, suggesting links between this effect and broader social attitudes.

Keywords: effort judgment, reward, self-serving bias, Bayesian, conservative attitudes

Every action comes with a cost and we constantly evaluate in retrospective manner whether a behavior was „worth it“. Effort judgments are particularly important in a social context as effort is one of the core dimensions for evaluating behavior of others. For instance, people that contributed more effort to a cooperative task are perceived as more likeable (Tjosvold et al., 1981), while free-riders, who do not contribute to the joint endeavor, are even punished (Fehr and Gaechter, 1999; Fehr and Gachter, 2002). Outside the lab effort estimations influence broader social attitudes, as for instance the support for welfare benefits is strongly linked to the perception that the people who receive those benefits have made an effort to find a job (Feldman, 1983). Although retrospective effort estimations have an important impact on different aspects of our lives, decision neuroscience has so far mainly focused on prospective effects of effort on behavior. The common finding of these studies is that effort is processed as a value discounting factor that is traded off against potential rewards when making choices (Walton et al., 2003; Skvortsova et al., 2014; Apps et al., 2015).

A recent study took a different approach and asked whether such an interaction extends to situations when effort is judged in retrospect, after the receipt of the reward (Pooresmaeili et al., 2015). Intuitively people have the impression to hold accurate representations about the effort they and others have spent. In contrast to this intuition, self-judgments of effort were influenced by the magnitude of obtained rewards, and thus effort estimations were biased. However, it is unclear whether the influence of reward on effort estimations is limited to self-judgments (i.e. some form of self-serving bias) or whether reward information is invariably incorporated when judging exerted effort in other contexts.

Interestingly, in the study of Pooresmaeili et al. (2015) the integration of rewards into effort-judgments was dependent on the contingency between reward and task difficulty, with a higher

influence of reward when it had a strong association with the task difficulty. Therefore, rewards only influenced effort-estimations if they were correlated with the executed effort. Thus, the integration of reward into effort ratings resembled Bayesian cue integration (Ernst and Banks, 2002; Kording and Wolpert, 2004), according to which different sources of information are combined, weighted by their reliability, to infer the true state of the world. In such a paradigm, the consideration of rewards provides an additional source of information for estimating the exerted effort, rather than an irrational bias. If this interpretation holds true, influences of rewards on effort estimations should not be limited to self-judgments but should also be present for other-judgments, potentially showing even a stronger effect size as people should have less reliable internal representations for the effort exerted by others compared to their own effort.

Whether optimal or not in terms of cue integration, an effort estimation biased by obtained reward could have a profound impact on social attitudes. For instance, the perception of expended effort for people with high incomes could be inflated, and vice versa, deflated for people with low income. Such an effect might serve to justify economic hierarchies and contribute to stabilization of existing inequalities (Pooresmaeili et al., 2015). Nevertheless, so far it is unclear whether the integration of reward into effort estimations is related to such social attitudes.

Here, we investigate whether people integrate reward magnitude when judging the effort another person has exerted. We present the data of 48 subjects (24 pairs) performing an effortful task, interleaved with a partner, and rating either their own or the partner's exerted effort. The task, adapted from Pooresmaeili et al. (2015), consisted of pressing two buttons as fast as possible to push a ball up a virtual ramp (**Figure 1A**). The ball rolled back with a constant displacement per time frame, simulating a gravity force. The difficulty of a given trial was

determined by the gravity force (6 difficulty levels that were pseudo-randomly presented). Subjects were informed about the fact that the task contained different difficulty levels represented by varying gravity force. On trials in which the ball was successfully pushed up the ramp, the participant who conducted the trial was rewarded. The rewards were contingent on the task difficulty (with values drawn from six Gaussian distributions with means depending on the difficulty, from 1.5 to 6.5 cents, and SD 1.2), but crucially, within one difficulty level the reward was randomly selected from the corresponding distribution. In most of the trials (70%), the reward was presented before subjects rated their effort but in some of the trials (30%) the reward was presented after participants had judged the effort. Trials in which reward was shown after the effort ratings served as a reference, since in those trials no influence of reward on effort ratings should be present. In half of the trials, participants performed the task themselves and rated their own effort (self-judgment). In the other half, participants rated the other person's effort (other-judgment) after watching them perform the task. On other-judgment trials participants observed on their screen how the ball moved up the virtual ramp and additionally heard a beep tone that corresponded to every key press through headphones (**Figure 1B**). Thus, participants received information about the task difficulty and the exerted effort of the other player.

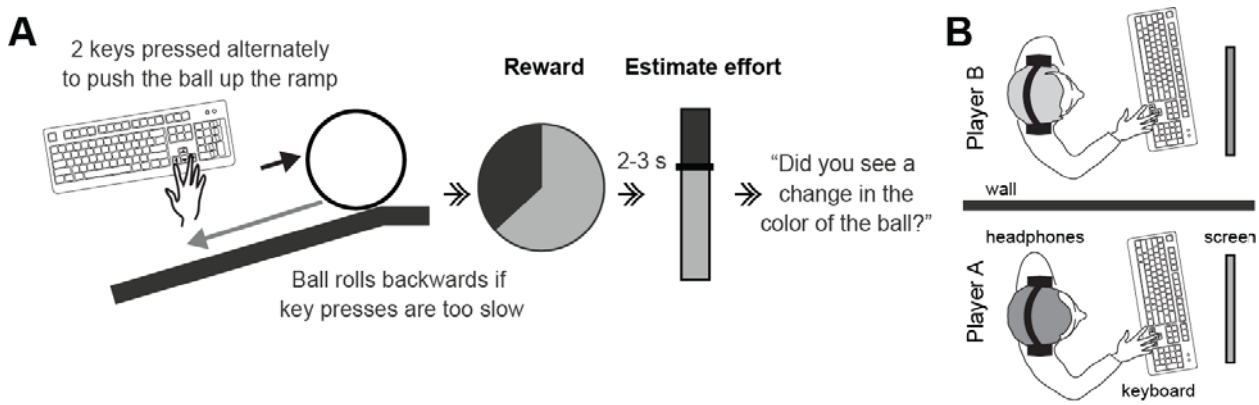


Figure 1. **Experimental set up and task. (A)** Participants were asked to move a ball up a ramp by performing fast, alternating key presses. A gravity force was simulated, displacing the ball backward by a constant amount on each display frame. We used six levels of task difficulty, corresponding to the amount of ball displacement per frame. After the ball was successfully pushed all of the way to the top of the ramp, participants received a monetary reward, where reward amount was contingent upon task difficulty. Subjects rated their effort by shifting the position of a sliding bar. In half of the trials participants performed the task themselves and rated their own effort. In the other half, participants rated the other person's effort after watching them performing the task. **(B)** Subjects were seated next to each other in the same room, separated by a partition wall. When the partner was playing the game subjects saw on their screen how the ball was moved up the ramp and heard via the headphones a beep tone for every key press the other person was conducting.

We first investigated the influence of reward magnitude on effort estimations by regression analysis of trial-to-trial variation in reward magnitude and effort estimation. The analysis was conducted separately for the self- and the other-judgments of each participant and only trials in which reward information was presented before the effort rating were considered. The regression slopes were calculated between the trial-to-trial variation of reward (rewards were z-scored within each difficulty level) and estimated effort (effort ratings were z-scored within each difficulty level).

Higher rewards were associated with higher effort ratings (**Figure 2A,B**), everything else being constant. Across subjects, there was a significant influence of reward variation on effort estimation for self-judgments (one-sample signed test, $p = 0.002$, **Figure 2C**) as well as for other-judgments (one-sample signed test, $p = 0.00002$, **Figure 2C**). Interestingly, the effect of rewards on effort-estimations was significantly stronger for other- than self-judgments (mean slopes of

0.14 and 0.1 respectively, Wilcoxon signed rank test, p =0.029, **Figure 2C**). Nevertheless, the amount of reward integration for self- and other-judgments was highly correlated across subjects (r=0.77, $p<10^{-11}$; see **Suppl. Figure S1**), indicating that integrating reward information into effort judgment is a general tendency of individuals, independent of the type of judgment that is required.

Figure 2. **Influence of reward on effort judgments for self and others.** For estimating the influence of reward on effort estimations, the trial-by-trial fluctuations in reward were used as predictors in a linear regression for estimated effort, separately for self- and other-judgments of each subject. **(A)** Data of one representative subject is shown as scatter plot and linear regression between trial by trial variation in reward and effort for self- (left) and other-judgments (right). Trial by trial reward magnitude and effort rating were z-scored within each difficulty level, so that values represent variation in comparison to the average of the difficulty level. **(B)** The individual (standardized) regression slopes for self and other ratings are presented. The example subject presented in (A) is indicated by the arrows. **(C)** Group average of slopes ± standard errors are presented. *p<0.05, , **p<0.01, ***p<0.0001

Having established an influence of reward magnitude on self- and other-judgments of effort, we next investigated how rewards were integrated with other information sources related to the exerted effort to form the final effort ratings. To this end, we compared predictions of five computational models with the behavioral data. The focus of these models was on how effort estimations, based on performance-related information (E_p) and on reward information (E_r), are combined to form the final effort rating (**Figure 3A**).

The exerted effort in this task can be estimated based on performance-related information such as the amount of key presses and the duration of a given trial. Those quantities were observable for subjects in every trial, both for self- and other-judgments. Assuming that participants' subjective effort estimations show some correspondence with performance-related information, it seems plausible that subjective effort estimations should be based (to some degree) on the exerted key presses and the elapsed time of a trial (as shown in **Figure 3B** and **Figure 3C**). We can leverage this to deduce presumed subjects' internal effort representations based on such observable quantities. Thus we can define the expected effort estimation given performance-related information (key presses and elapsed time) as follows:

(1) $E_p = \beta_0 + \beta_k * \text{key press} + \beta_t * \text{time} + \varepsilon_p$

where β_0, β_k and β_t and were fitted to the trials in which the effort was rated *before* the reward was revealed, giving an unbiased estimation of effort-judgments.

Moreover, based on the task design, rewards were contingent on exerted effort. Therefore, effort judgments can be predicted based on reward. We can define the expected effort estimation given reward (E_r) as follows:

(2) $E_r = \beta_0 + \beta_r * \text{reward} + \varepsilon_r$

Analogous to Eq(1), we derived β_0 and β_r from the trials in which the effort was rated *before* the reward was revealed. We then applied E_p and E_r to the trials where effort was rated *after* the reward was revealed, to get a trial-by-trial estimation of subjects' internal effort representation based on either performance related information (E_p) or reward information (E_r) only.

Both effort estimations, E_r and E_p, are associated with some uncertainty, representing how well exerted effort can be inferred from those different sources. The uncertainty associated with E_r and E_p can be formalized as the mean squared error:

(3) $\sigma^2_p = \Sigma\ \varepsilon_p{}^2$

and

(4) $\sigma^2_r = \Sigma\ \varepsilon_r{}^2$

In each trial, E_r and E_p can differ by a certain amount ($\Delta E \neq 0$). We consider several models how these two informational sources might be integrated into the actual effort rating E.

Model 1 (performance only model) assumes the actual effort rating is only based on the E_p without taking reward information into account. Thus, the E_p would directly result in the actual effort rating E:

(5) $E= E_p$

Model 2 (reward only model), in contrast, only relies on reward information (E_r) and completely ignores performance information:

(6) $E= E_r$

Model 3 (simple average model) assumes that both sources of information are considered and that E is an equally weighted average of E_r and E_p:

(7) $E= 0.5*E_p+ 0.5*E_r$

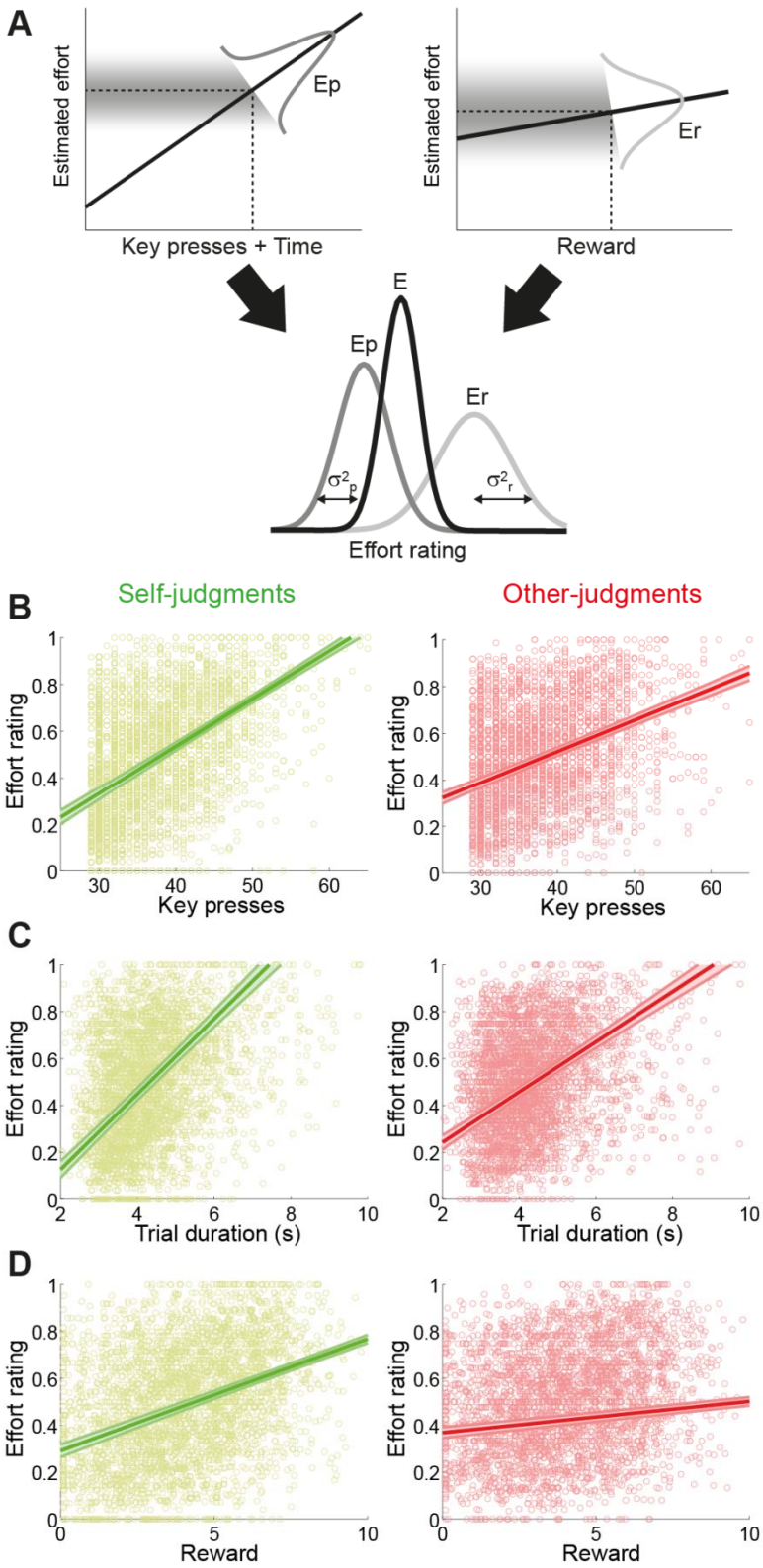


Figure 3. **Different information sources are integrated to form a final effort rating. (A)** Two types of information can be combined to derive the final effort estimation (E): E_p is the effort estimation based on performance-related information (key presses and trial duration), whereas E_r is the effort estimation based on reward related information. On every trial, E_p and E_r depend on the specific amount of key presses, time and reward. Both estimates will differ by some degree. According to Bayes optimal models, the influence of each signal on E depends on its reliability, represented by the width of the distributions (σ^2_p and σ^2_r). **(B)** Relation between the number of key presses and effort ratings for self-judgments (left) and other-judgments (right), based on the trials in which effort was rated *before* the reward was revealed. Scatter plots (light colors) show the corresponding data over all subjects. The regression line shows the average slope for all subjects ± the group standard error of the individual slopes. **(C)** Similar to (B) for the relation between trial duration and effort ratings. **(D)** Similar to (B) for the relation between reward magnitude and effort ratings. Note that this relation effectively captures the correlation between reward and effort-rating built-in by the task design: Since trials that are more difficult were associated with higher rewards and these trials were subjectively rated with higher exerted efforts, reward magnitude carried information about the subjective estimations of exerted effort in these trial.

So far, the models did not take the reliability of the information into account. However, **Model 4** (Bayesian average model) assumes that the variance of the different sources of information used in order to weight E_p and E_r in a Bayes optimal way (see **Figure 3A**):

(8) $E = \omega_p * E_p + \omega_r * E_r$

where $\omega_r = 1/\sigma^2_r / (1/\sigma^2_r + 1/\sigma^2_p)$ and $\omega_p = 1 - \omega_r$. The variances σ^2_r and σ^2_p are derived as explained above (Eq(3) and Eq(4)) and the weights are directly derived from variances.

Moreover, it might be possible that E_r and E_p are combined, but with a different weighting than predicted by any of the previous models. Thus, in **Model 5** (flexible weighting model) we set the weighting parameter ω_r as a free parameter and $\omega_p = 1 - \omega_r$.

The model evaluation was done by computing the maximum-likelihood fits to the trial-to-trial data of individual subjects, separately for self- and other-judgments. The quality of fits was compared by the Bayesian Information Criterion (BIC; Burnham and Anderson, 2002; Lewandowsky and Farrell, 2010).

The model comparison revealed that for both, self- and other-judgments, the Bayesian average model had the lowest BIC (**Table 1**), showing that this model provides the most accurate and parsimonious explanation of the data. For self-judgments the second best model was the flexible weighting model, whereas it was the simple average model for other-judgments.

***Table 1*. Results of model comparison.** BIC: Bayesian Information Criterion. BIC values are summed across 48 subjects. The Bayesian average model had the lowest BIC score and is therefore ranked highest by the model comparison. ΔBIC: Difference in BIC of each model minus the BIC of the best fitting model (BIC_{model}-BIC_{best}).

	Self-judgment			**Other-judgment**		
Model	BIC	ΔBIC	Mean R^2	BIC	ΔBIC	Mean R^2
Effort only	14053.9	+73.8	.315	152031.4	+291.3	.184
Reward only	15329.7	+1349.6	.193	15538.3	+626.3	.091
Average	14092.1	+111.9	.305	14921.4	+9.7	.188
Bayesian average	**13980.1**	**0**	**.324**	**14911.8**	**0**	**.192**
Flexible weighting	13997	+16.9	.343	14999.6	+87.9	.206

These results provide additional evidence that reward information is integrated when forming effort-estimations about self and others. Moreover, subjects integrate different sources of information weighted by their reliability, as predicted by Bayesian cue integration. Therefore, the integration of rewards into final effort ratings appears optimal in terms of information usage as rewards are combined with performance-related information based on the reliability with which each information source can be used to infer the exerted effort.

After having demonstrated that reward influences the estimation of exerted effort of self and others, we asked whether this behavioral effect showed any meaningful transfer to broader social attitudes. We hypothesized that reward-biased effort estimations, although optimal in terms of information integration, could be associated with acceptance of economic inequalities and reluctance to social change. Such attitudes show a considerable overlap with political conservatism, and it has been proposed that preference for social stability and acceptance of social inequalities are the two core aspects of conservative world-views (Jost et al., 2003; Jost et al., 2009). In order to probe a relation between our behavioral effects and broader social attitudes, we assessed several questionnaires related to conservative world-views (Pratto et al., 1994; Funke, 2005; see Methods for detailed description) and conducted a factor analysis on these scales using maximum-likelihood estimation. Based on the scree-test (Cattell, 1966) we identified a sharp drop in eigenvalue after the first component (see **Suppl. Figure S2**), indicating that one factor, explaining 76% of the variance, was the best and most parsimonious solution for the covariance structure of the questionnaire data. We interpreted this factor as “conservative world-views” based on the factor loadings (**Figure 4A**). This factor represented acceptance/support of economic inequality, reluctance to social change and obedience to established authorities.

We extracted factor scores for every subject to relate those with our behavioral effect of reward integration into effort-judgments. Due to the high correlation between reward integration for self- and other-judgments (cf. **Suppl. Figure S1**), we used the sum of both effects (from the model free analysis: $slope_{self}+slope_{other}$, see **Figure 2**) as indicator for the extent to which subjects were biased in their effort estimations by rewards.

As we hypothesized, people with stronger influence of reward magnitude on effort-estimations displayed more conservative world-views ($\beta=0.41$, $p=0.005$, see **Figure 4B**), underlining an interesting link between this behavioral effect and social attitudes. Importantly, this association between conservatism and reward influence could also be shown when separately considering self-judgments ($\beta=0.32$, $p=0.03$) and other-judgments ($\beta=0.44$, $p=0.002$). It is remarkable that an association between a low-level behavioral laboratory task and high-level social attitudes could be found. In particular, the size of this effect is surprising since the influence of reward explained the variance of conservative attitudes to a similar extent as well-established socio-demographic predictors such as age, gender, education or income have been shown to do (Sidanius et al., 1994; Saucier, 2000; Ortoleva and Snowberg, 2015).

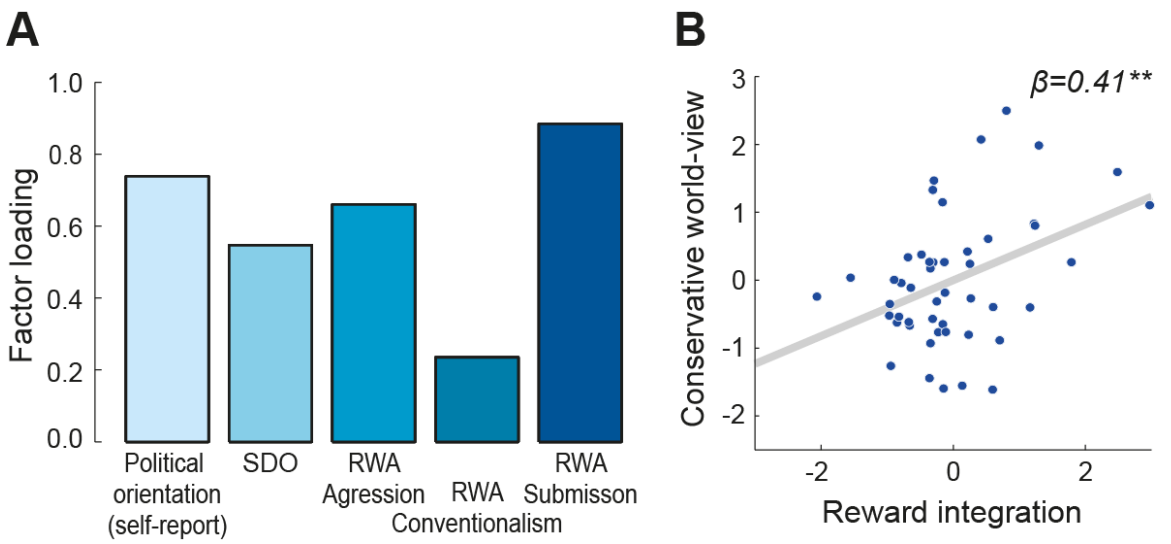


Figure 4. **Conservative world-views (factor scores) are positively correlated with the amount of reward integration into effort-estimation. (A)** A factor analysis over multiple questionnaires revealed that one factor (based on scree-test) which can be interpreted as conservatism is explaining the interrelation between these scales. SDO: social dominance orientation; RWA: right-wing authoritarianism. **(B)** Stronger influence of rewards on effort estimations is positively associated with conservative world-view. The sum of the slopes for self- and other-judgments ($slope_{self}$+$slope_{other}$) is used as indicator for reward integration. Since we controlled for empathy and perspective taking in the regression analysis, we present values of reward integration from which we regressed out influences of these covariates. **p<0.01

Since we have shown that the integration of rewards into effort-estimations can be optimal with respect to the use of available information, we asked whether the increased integration of reward information for conservatives is an irrational bias or whether it is in line with Bayes optimal cue integration. One possibility is that conservative group could either show less reliable effort-estimations based on the available performance information (key presses and time), in which case it would be beneficial to weight the reward information stronger. Alternatively, conservatism could lead to integration of reward information above the extent that would be Bayes optimal.

To this end, we investigated whether conservatism was associated with the weights of reward information (ω_r) predicted by the Bayesian averaging model. There was a significant

association between ω_r and conservative world-views for self- (β=0.3, p=0.04) but not for other-judgments (β=0.21, p=0.15), indicating that the increased incorporating of rewards into effort estimations for self-judgments is in line with Bayesian cue integration. This increased weight for reward information was driven by less reliable (self-) effort estimations based on performance information, indicated by a positive association between uncertainty σ^2_p and conservative world-views (β=0.27, p=0.038, one-tailed). Since the increased reward integration within the conservative group for other-judgments was not correlated with the weights of reward information (ω_r) predicted by the Bayesian model, we investigated whether conservative people deviated from Bayes optimality for those judgments. To this end, we conducted the model comparison for other-judgments separately for people with low and high conservatism scores (based on a median-split). The group with low conservatism scores was best described by the Bayesian average model (difference in BIC relative to next best model: 12.0), however the group with high conservatism scores was best described by the simple average model (BIC=7253.6) followed by the Bayesian average model (BIC=7256). In line with this, the extent to which each subject's other-judgments were captured by the Bayesian average model (BICweight, see Methods for a detailed description) was negatively correlated with conservative attitudes (β=-0.26, p=0.039, one-tailed).

These results indicate that the increased incorporation of rewards by conservative people is in line with Bayes optimality for self-judgments, whereas when judging someone else's effort conservative people deviate from Bayes optimality. We note however that these results might depend to some extent on the exact formulation of the used models and the investigated model space.

The weighting of different sources of information by their reliability is the optimal way of combining information. However, here we show that the increased incorporation of reward information is driven by less reliable effort estimates based on performance criteria. We suggest that the estimation of exerted effort based on such objective criteria represents a form of metacognitive ability (Fleming et al., 2012): the correspondence between the subjective effort rating and objective indicators of exerted effort shows that people are able to accurately monitor and report their exerted effort. It would be interesting to probe whether an improvement of this metacognitive ability (e.g. by training) might result in a reduced incorporation of reward information when rating one's effort, as would be predicted by the Bayesian framework. The negative correlation between conservatism and reliability of performance-based estimates might indicate that this group showed a reduction in this facet of metacognition, leading to less insight about the true effort they have exerted, a possibility that future studies could investigate further.

Integrating reward information into effort ratings can be seen as a form of uncertainty reduction. Interestingly, there is a solid body of literature suggesting uncertainty aversion as one of the core psychological foundations of conservative attitudes (see Jost et al., 2003 for an extensive review). It might be promising to revisit those existing psychological findings by using a computational approach, to see whether such findings could be accommodated within the Bayesian framework. Such fine-grained analysis might be helpful in understanding the psychological underpinnings of social attitudes.

Although a Bayesian framework can explain why conservatives are more influenced by rewards when judging their effort, they deviate from this optimality when judging someone else's effort. Moreover, even for self-judgments we have shown that effort ratings of conservatives are less related to performance-based effort estimates (key presses and time). In some situations,

such exaggerated influence of reward can thus lead to a biased perception of people's exerted effort. Our results suggest that people with conservative views might judge other people's effort more through the received outcomes, indicating that in everyday life they would adhere to the view that people with low income have also exerted less effort, and conversely, rich people deserve their wealth since they have worked harder, perhaps ignoring other contributing factors. A fine-grained understanding of the building blocks contributing to social attitudes holds the potential for counteracting such social biases and overcoming unfounded beliefs.

Acknowledgements

We thank Aurel Wannig and Caedyn Stinson for their comments on the manuscript. This work was supported by a seed fund grant from Leibniz ScienceCampus *Primate Cognition*, Göttingen, Germany to IK and AP. AP was supported by an ERC Starting Grant (716846, *Rewarded Perception*).

Authors' contribution

MR, IK and AP conceptualized and designed the task. MR and UT chose and devised the online questionnaire. MR and FP conducted the experiments. MR analyzed the data. MR, IK and AP wrote the manuscript.

Methods

Participants

The final sample composed 48 subjects (27 females, age 20-37 years, mean: 25.8 years). Ten subjects were excluded from the original sample (N=58) as their effort ratings were not significantly influenced by task difficulty (i.e. no significant effect of a trial-by-trial linear regression between task difficulty and effort ratings), indicating that they rated the effort randomly. Participants were paid 6 Euro per hour, plus a bonus for their performance in the experiment, resulting in a total mean payment of 20 ± 5 Euro. All subjects gave written consent about their attendance. The study was approved by the local ethics committee of University Medicine Göttingen, proposal number **15/7/15**.

Experimental design

The behavioral experiment was adapted from Pooresmaeili et al. (2015) in order to test two subjects simultaneously. Participants pressed two keys of the keyboard, as fast as possible, to push a ball up a virtual ramp before rating the exerted effort. Trials were presented interleaved between both subjects. In half of the trials participants performed the task themselves and rated their own effort (self-judgment). In the other half, participants rated the other person's effort (other-judgment) after watching them performing the task.

While the participants tried to push the ball up the ramp with fast alternating key presses, the ball rolled back by a constant amount during each frame of the display, simulating gravity force. This gravity force differed pseudo-randomly between trials and created the six difficulty levels of the task (individually adjusted for each participant, see below). If the participants managed to push the ball up the ramp within 10 seconds they were rewarded for this successful trial (only the participant who performed the trial was rewarded). The amount of reward was

contingent on the task difficulty and the actual reward value was drawn from Gaussian distributions with means ranging from 1.5 to 6.5 cents and standard deviation of 1.2 cents. In most of the trials (70%) the reward was presented before subjects rated their effort and in some of the trials (30%) the reward was presented after participants had judged the effort. Trials in which reward was shown after the estimation of effort served as a reference, since in those trials no influence of reward was present.

The subjects were seated inside the same room, next to each other, separated by a partition wall. Each subject in a pair performed 200 trials resulting in a total of 400 trials which were grouped into 10 experimental blocks (each containing 40 trials). The subjects saw on their screen how the ball was moved up the ramp, independently of whether they performed the task themselves or whether the other person was playing. Moreover, subjects wore headphones and each key press was accompanied by a beep tone. Hence, subjects had visual and auditory information about the key presses. In trials where the participants conducted the task themselves they judged their own exerted effort and in trials when the other person performed the task, they estimated the effort that the other person had performed. Subjects did not see how the other person rated the effort, in order to prevent reciprocal influences.

Stimuli and Task

The stimulus presentation and registration of behavioral responses was controlled by one personal computer. Stimuli were produced with MATLAB and the Psychophysics Toolbox. Each trial consisted of a stimulus (ball and ramp), a reward, and an effort rating display all shown on a black background (a trial timeline is shown in **Figure 1A**). The stimulus display contained the ball (radius: three visual degrees), initially at the starting part of the ramp (ramp length: 19 visual degrees; both ball and ramp had a light gray color). The ball was displaced up the ramp with

consecutive alternate key presses (left and right arrow keys) until it reached the upper plateau. Each key press resulted in a constant amount of displacement (0.87 visual degrees per key press) and was counteracted by a gravity force of variable strength that displaced the ball backward. Moreover, each key press caused a beep tone which was played on the headphones of both participants. To determine the levels of used gravity force individually for each subject, at the beginning of the experiment we asked each subject to push the ball up the ramp by pressing both keys alternately and consecutively as fast as they possibly could. To motivate the participants, they were told that they would receive 1 Euro bonus if they managed to push the ball up the ramp. Ninety percent of the gravity force necessary to counteract the maximum number of key presses in a limited time (10 s) determined the maximum gravity force used in the experiment. This estimation was done for each participant alone; the other participant in a pair waited outside the room. Based on this individualized estimate, six equally spaced gravity levels were defined and used in the experiment. A trial was aborted if key presses did not occur fast enough (maximum pause allowed was 2 s). If participants were able to successfully push the ball all of the way up, they received a monetary reward, with the amount contingent on task difficulty (gravity force level). Reward magnitude was defined based on six Gaussians with means of 1.5, 2.5, 3.5, 4.5, 5.5, and 6.5 cents with a SD of 1.2 cents. Reward display consisted of a pie chart that depicted subjects' reward as a proportion of maximum reward possible and a number that showed the reward in Arabic numerals. Effort rating display consisted of a slider, and participants were instructed to set the slider at a position that represented their experienced effort during a trial proportionate to the maximum effort they could experience during the experiment. In 70% of trials, the reward display was shown immediately after the stimulus display, whereas in 30% of trials it occurred after the rating of effort. Before the actual experiment started,

participants conducted 36 training trials (which were not analyzed) to get familiar with the experiment and to learn about the effort-reward contingency. Because task difficulty and reward were the only parameters that varied across trials, we intended to distract participants from the main purpose of this experiment to prevent ad hoc strategies of relating reward and effort. For this reason we introduced a second task into the paradigm where subjects were asked to report whether they had seen a brief (21 ms or three frames) color change on the ball (to green, red, or blue), which occurred in 50% of the trials. This question was asked randomly in 20% of trials.

Questionnaires

The set of questions about social attitudes and political orientation were administered via online questionnaire which the participants filled out at least 24 hours before they performed the behavioral experiment. We used a political orientation self-report (on a scale from 0=”very liberal” to 100=”very conservative”) to assess explicit conservative orientation. As indirect indicators for conservatism we assessed social dominance orientation (Pratto et al., 1994), reflecting a general preference for hierarchical intergroup relations, and right-wing authoritarianism (Funke, 2005), measuring authoritarian submission, authoritarian aggression, and conventionalism (Altemeyer, 1981). Since we were interested in the underlying factor, linking these scales, we conducted a factor analysis over these scales, using a maximum-likelihood estimation. The amount of factors was based on the Scree-test (Cattell, 1966; see **Suppl. Figure S2**), indicating that one factor was sufficient to explain the interrelation between these questionnaires. Factor scores of the first factor were extracted to relate those to behavioral measures.

As conservatism, especially right-wing authoritarianism and social dominance orientation, were reported to correlate negatively with empathy (Pratto et al., 1994; Lippa and Arad, 1999),

and empathy could arguably have an impact on our behavioral effects of interest, we assessed empathy and perspective-taking as covariate (Maes et al., 1995).

Statistical analysis

All presented results were tested by using two-tailed statistics, if not otherwise stated. Results from regression analysis are always reported as standardized betas.

Outlier values of effort ratings (3 standard deviations above or below the participant's mean effort) were removed from the analysis, since some subjects reported in the debriefing to have occasionally rated extreme effort values to test their assumptions about what the aim of the experiment was (e.g. whether they could influence the received reward by their effort ratings).

For the relation between trial-to-trial variation in reward magnitude and effort ratings, both were z-scored within difficulty levels, resulting in values representing variation in reward magnitude and effort rating with respect to the average of the corresponding difficulty levels. After having created these within difficulty variations we pooled them across all difficulty levels and conducted a robust linear regression (robust regression toolbox in MATLAB) with reward as predictor and effort rating as dependent variable.

For comparison of the computational models, each model was fitted to participants' z-scored effort ratings, separately for self- and other-judgments, and BIC scores were calculated. For model comparison the sum of individual BIC scores was used.

To investigate the relationship between conservative world-views and influence of rewards on effort-estimations linear regression analysis was used. Factor scores of the first factor from the factor analysis were used as indicator for conservative world-views. The sum of slopes for self- and other-judgments between reward magnitude and effort estimations was used as indicator for the influence of rewards on effort-estimations. The influence of rewards on effort

estimations was used as predictor and conservative-world views as dependent variable in the linear regression analysis, controlling for empathy and perspective-taking as covariates.

To investigate whether conservative world-views were related to weights of reward-information (ω_r) predicted by the Bayesian average model, we conducted linear regression analysis with conservative world-views as predictor and ω_r as dependent variable. To further investigate whether this association was driven by less reliable effort estimations based on performance information we conducted a linear regression between conservative world-views and $\sigma^2_{p.}$

To investigate whether there is an association between conservative attitudes and the extent to which subjects behavior was in line with Bayesian cue integration we calculated the BIC-weight of the Bayesian average model for every subject as follows:

$$\Delta\text{Bayes average} = \text{BIC}_{\text{Bayes avergae}} - \min(\text{BIC})$$

$$\text{BICweight}_{\text{Bayes average}} = \frac{\boldsymbol{e}^{(-.5*\Delta\textbf{Bayes average})}}{\sum_i \boldsymbol{e}^{(-.5*\Delta_i)}}$$

The Bayesian weights of all of the models in a set sum up to 1 and indicate the probability for each model to be the best model for the data.

We then calculated a linear regression with conservative world-views as independent variable and $\text{BICweight}_{\text{Bayes average}}$ as dependent variable.

Reference List:

Altemeyer B (1981) Right-wing authoritarianism. . Winnipeg, Canada: University of Manitoba Press.

Apps MA, Grima LL, Manohar S, Husain M (2015) The role of cognitive effort in subjective reward devaluation and risky decision-making. Sci Rep 5:16880.

Burnham K, Anderson D (2002) Model selection and multimodel inference: a practical information-theoretic approach New York: Springer.

Cattell RB (1966) The Scree Test For The Number Of Factors. Multivariate Behav Res 1:245-276.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Fehr E, Gaechter S (1999) Cooperation and Punishment in Public Goods Experiments. In: CESifo Group Munich.

Fehr E, Gachter S (2002) Altruistic punishment in humans. Nature 415:137-140.

Feldman S (1983) Economic Individualism and American Public Opinion. American Politics Quarterly 11:3-29.

Fleming SM, Dolan RJ, Frith CD (2012) Metacognition: computation, biology and function. Philos Trans R Soc Lond B Biol Sci 367:1280-1286.

Funke F (2005) The Dimensionality of Right-Wing Authoritarianism: Lessons from the Dilemma between Theory and Measurement. Political Psychology 26:195-218.

Jost JT, Federico CM, Napier JL (2009) Political ideology: its structure, functions, and elective affinities. Annu Rev Psychol 60:307-337.

Jost JT, Glaser J, Kruglanski AW, Sulloway FJ (2003) Political conservatism as motivated social cognition. Psychol Bull 129:339-375.

Kording KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427:244-247.

Lewandowsky S, Farrell S (2010) Computational Modeling in Cognition: Principles and Practice: SAGE.

Lippa R, Arad S (1999) Gender, personality, and prejudice: The display of authoritarianism and social dominance in interviews with college men and women. Journal of Research in Personality 33:463-493.

Maes J, Schmitt M, Schmal A (1995) Gerechtigkeit als innerdeutsches Problem: Werthaltungen, Kontrollüberzeugungen, Freiheitsüberzeugungen, Drakonität, soziale Einstellungen, Empathie und protestantische Arbeitsethik als Kovariate. Trier: Univ., Fachbereich I, Psychologie.

Ortoleva P, Snowberg E (2015) Overconfidence in Political Behavior. American Economic Review 105:504-535.

Pooresmaeili A, Wannig A, Dolan RJ (2015) Receipt of reward leads to altered estimation of effort. Proc Natl Acad Sci U S A 112:13407-13410.

Pratto F, Sidanius J, Stallworth LM, Malle BF (1994) Social dominance orientation: A personality variable predicting social and political attitudes. Journal of Personality and Social Psychology 67:741-763.

Saucier G (2000) Isms and the structure of social attitudes. J Pers Soc Psychol 78:366-385.

Sidanius J, Pratto F, Rabinowitz JL (1994) Gender, Ethnic Status, and Ideological Asymmetry:A Social Dominance Interpretation. Journal of Cross-Cultural Psychology 25:194-216.

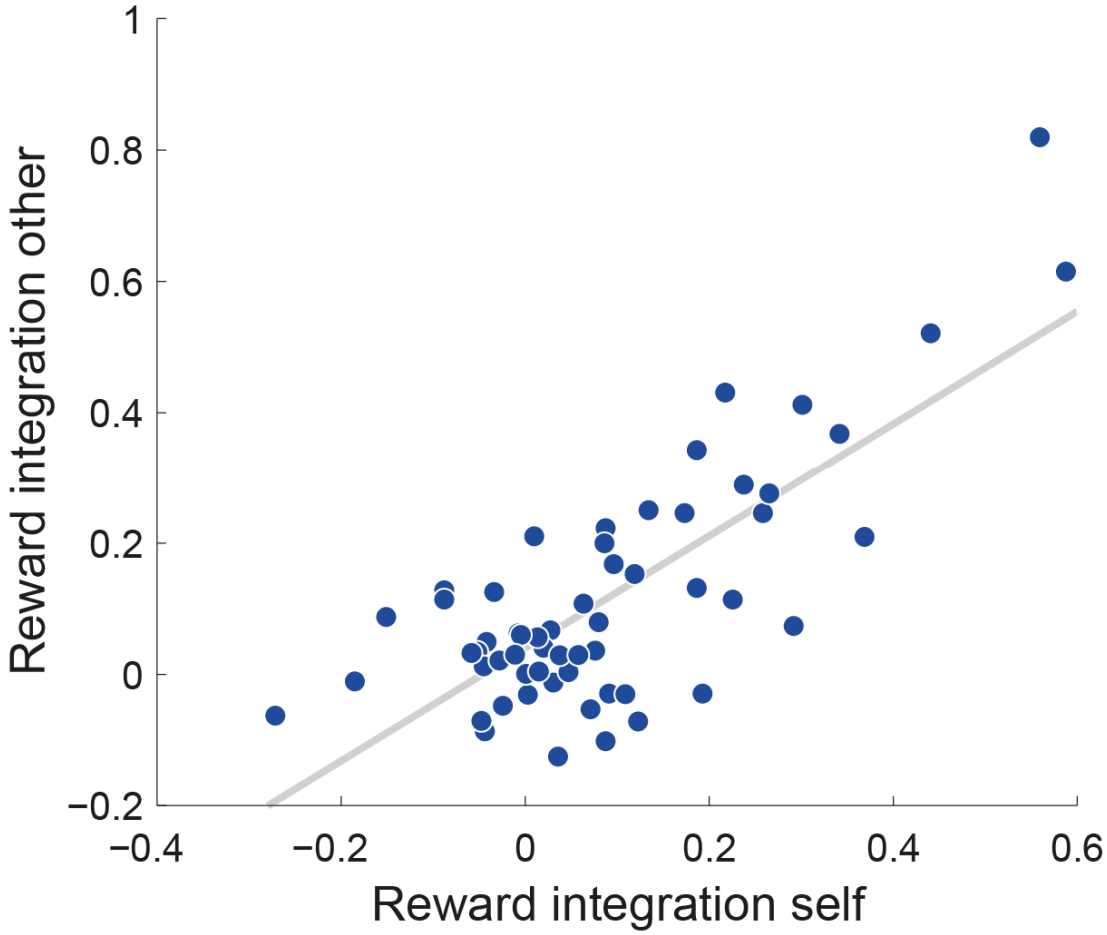
Skvortsova V, Palminteri S, Pessiglione M (2014) Learning to minimize efforts versus maximizing rewards: computational principles and neural correlates. J Neurosci 34:15621-15630.

Tjosvold D, Johnson DW, Johnson RT (1981) Effect of partner's effort and ability on liking for partner after failure on a cooperative task. The Journal of Psychology: Interdisciplinary and Applied 109:147-152.

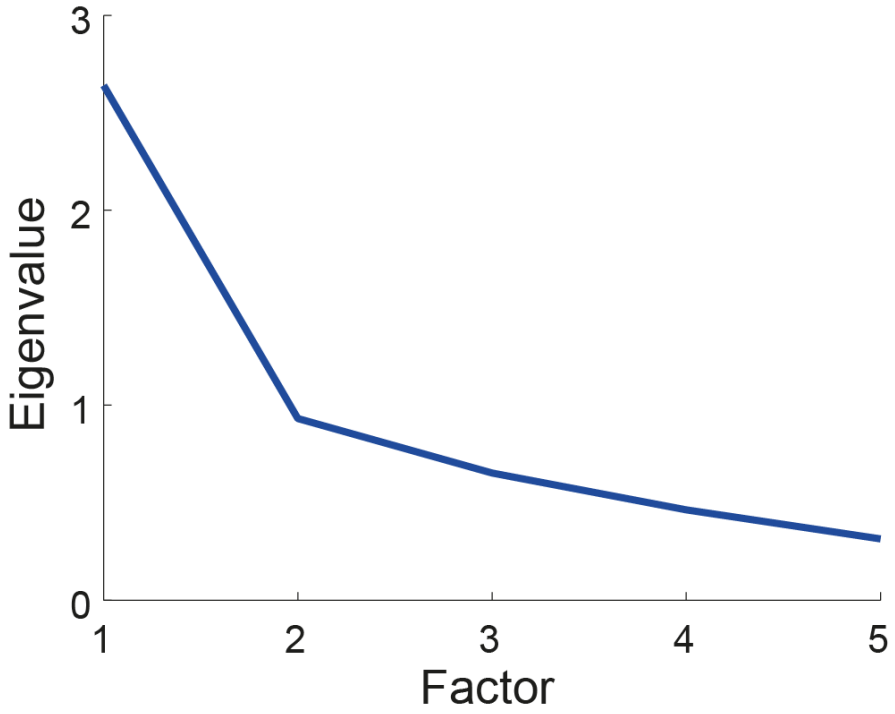
Walton ME, Bannerman DM, Alterescu K, Rushworth MF (2003) Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. J Neurosci 23:6475-6479.

Supplementary Information

Supplementary information contains two supplementary figures.



Supplementary Figure S1. **The amount of reward integration for self- and other-judgments was highly correlated across subjects.** The standardized regression slopes between reward magnitude and effort ratings are used as indicator for reward integration. This measure was calculated separately for self- and other-judgments for every subject.



Supplementary Figure S2. **The scree plot shows that one factor is enough to capture the interrelation between several questionnaires used to assess conservative world-views.** Eigenvalues for up to five possible components are presented. There is a sharp decline in eigenvalue from the first to the second factor, forming a clear "elbow".