Running Head: SOCIAL REVERSAL LEARNING

Observing Others Stay or Switch – How Social Prediction Errors Are Integrated into Reward Reversal Learning

Niklas Ihssen[a], Thomas Mussweiler[b] & David E. J. Linden[a]

[a]Department of Psychology, Durham University, Durham DH1 3LE, UK

[b]Department of Psychology, University of Cologne, 50931 Cologne, Germany

[c]School of Psychology & MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff CF10 3AT, UK

Corresponding author:

Dr. Niklas Ihssen
Department of Psychology, Durham University, Wolfson Building, Queen's Campus
Stockton-on-Tees TS17 6BH, United Kingdom
Telephone number: +44 (0) 191 33 40808
e-mail: niklas.ihssen@durham.ac.uk

Keywords: Reversal learning; social influence; reward; prediction error; similarity

Abstract

Reward properties of stimuli can undergo sudden changes, and the detection of these 'reversals' is often made difficult by the probabilistic nature of rewards/punishments. Here we tested whether and how humans use social information (someone else's choices) to overcome uncertainty during reversal learning. We show a substantial social influence during reversal learning, which was modulated by the type of observed behavior. Participants frequently followed observed conservative choices (no switches after punishment) made by the (fictitious) other player but ignored impulsive choices (switches), even though the experiment was set up so that both types of response behavior would be similarly beneficial/detrimental (Study 1). Computational modeling showed that participants integrated the observed choices as a 'social prediction error' instead of ignoring or blindly following the other player. Modeling also confirmed higher learning rates for 'conservative' versus 'impulsive' social prediction errors. Importantly, this 'conservative bias' was boosted by interpersonal similarity, which in conjunction with the lack of effects observed in a non-social control experiment (Study 2) confirmed its social nature. A third study suggested that relative weighting of observed impulsive responses increased with increased volatility (frequency of reversals). Finally, simulations showed that in the present paradigm integrating social and reward information was not necessarily more adaptive to maximize earnings than learning from reward alone. Moreover, integrating social information increased accuracy only when conservative and impulsive choices were weighted similarly during learning. These findings suggest that to guide decisions in choice contexts that involve reward reversals humans utilize social cues conforming with their preconceptions more strongly than cues conflicting with them, especially when the other is similar.

1. Introduction

Adaptive behavior depends on learning and retaining associations between specific stimuli or responses on the one hand and positive or negative outcomes (reward or punishment) on the other. In a complex and dynamic environment organisms must also adequately respond to sudden changes in those associations and re-learn established contingencies. A widely used experimental tool to study this process in animals and humans is reversal learning (Cools, Clark, Owen, & Robbins, 2002; Dias, Robbins, & Roberts, 1996; Jones & Mishkin, 1972). In a typical setup, human participants learn to choose one of two simple visual stimuli by receiving monetary rewards for correct responses (stimulus A) and being punished by monetary loss for incorrect responses (stimulus B). After a variable number of trials, these contingencies are reversed so that the participant will be rewarded for choosing B and be punished for choosing A. Trial-by-trial choices in this task can be predicted by the algorithms of reinforcement learning models which are based on the calculation of reward prediction errors (Jocham, Neumann, Klein, Danielmeier, & Ullsperger, 2009).

1.1 Social information and decision-making

Critically, in real-life situations learning of reward contingencies is not only achieved by trial-and-error and reward prediction errors but also by social learning, that is, by observing the choices of other agents who are exposed to the same or similar decisional contexts. In the majority of everyday choice situations (e.g. choosing between alternative products or services) social information is readily available either through behavioral observation of others or through active gathering of information (e.g. consumer reviews). Observational factors can be expected to become especially important if well-established behavioral choice routines need to be revised because the expected outcome is not received or experienced as less rewarding. In such situations the possibility of a change in the underlying reward probabilities (e.g., the quality of the usually preferred product/service has changed)

will evoke decisional uncertainty which is a potent trigger for ‘social reality testing’, that is, the reliance on others to resolve ambiguity (Festinger, 1950). The literature to date has ignored whether information about others' choices affects responding to sudden changes in reward properties of a stimulus as implemented in the reversal learning task. This is surprising given the evidence that other basic cognitive processes, such as perceptual judgments, can profoundly be shaped by social influence (Asch, 1956; Baron, Vandello, & Brunsman, 1996).

Social influence can be governed, on the one hand, by socio-normative mechanisms, originating from the influenced person's motivation to gain social approval if the influencing person is present (as in Asch's classic line discrimination studies). On the other hand, it can also arise in the absence the influencing person and social pressure, being motivated by informational needs (Deutsch & Gerard, 1955) and the wish to resolve ambiguity to optimize one's outcomes. Such informational social influence is likely to operate in choice decisions involving uncertain rewards and a few studies have begun to document social influences on probabilistic reward learning. However, these studies used fixed (Biele, Rieskamp, Krugel, & Heekeren, 2011; Burke, Tobler, Baddeley, & Schultz, 2010) or gradually changing (Behrens, Hunt, Woolrich, & Rushworth, 2008; Cooper, Dunne, Furey, & O’Doherty, 2012) reward contingencies rather than a setup involving unexpected reversals.

1.2 Predictions for the use of social information during reversal learning

The primary goal of the present studies was to explore the use of social information (observed choices by another agent) during reversal learning, specifically, how such social influence is mediated by the (i) type of observed choice behavior (conservative versus impulsive) and (ii) the similarity of the observed agent.

The differentiation between conservative and impulsive choices during reversal learning arises as a result of the task-inherent combination of probabilistic reward and possibility of reversals. In other words, even if reward contingencies have not changed, correct choices are occasionally punished by monetary losses (so-called Probabilistic Errors,

ProbErrs). Consequently, after each punishment occurring against the background of correct responses an individual has to decide whether to switch to the other stimulus (taking the punishment as indicator of reversed contingencies) or whether to stay with their previous choice (taking the punishment as a ProbErr). Accordingly, choices in trials that immediately follow ProbErrs and reversals can be classified as reflecting either a *conservative* or an *impulsive* type of choice behavior. Stay responses correspond to conservative choices as the agent relies on accumulated information about a specific choice option – which has been gathered across a number of trials before the unexpected punishment – rather than trying a new option. This can be seen in analogy to an *exploitative* decision-making strategy in multi-armed bandit problems, in which multiple choice options with varying pay-offs are available (Cohen, McClure, & Yu, 2007). Conversely, switch responses can be seen as *impulsive* choices (Fineberg et al., 2010), reflecting an abrupt change of choice behavior based on single events without consideration of the previous choice history. Importantly, adaptive behavior during reversal learning requires the dynamic use of both types of behavior. Although impulsive responses manifest as errors in the trial(s) after ProbErrs (= post-ProbErr choice) they lead to correct choices in the trials(s) after true reversal events. Vice versa, conservative responses increase accuracy after ProbErrs but lead to incorrect choices (‘perseverations’) after reversal events. The key question addressed in the present framework was whether *observing* someone else making conservative choices affects our own choices differently than observing someone else making impulsive choices.

Diverging effects for observed conservative versus impulsive choices can be predicted from findings about the biased use of information during individual decision-making. Thus, it is possible that learners take into account only social information conforming to their preconception or expectation about the correct versus incorrect stimulus (established before the other player’s choice is observed). This preconception is based on the learner’s more frequently chosen stimulus in a given reversal episode and thus usually corresponds to a bias

towards conservative choices made by the observed other. Such selective use of social information would parallel a 'confirmation' bias described in the context of individual decision-making (Nickerson, 1998). Conversely, a social influence bias towards the other's impulsive choices may arise if observational reversal learning is expectation-free but driven by the higher saliency of impulsive (switch) responses occurring against the stream of standard (non-switch) choices between two reversals.

Apart from the type of observed choice behavior, the present studies aimed to examine social influence on reversal learning as a function of perceived similarity of the observed agent. Similarity has been shown to influence different cognitive processes across a wide range of phenomena, including decision-making (Kahneman & Miller, 1986). Similarity is also effective in modulating a variety of social behaviors, ranging from the experience of vicarious reward (Mobbs et al., 2009) to cooperative behavior (Mussweiler & Ockenfels, 2013). Pertinent to the present work, the behavior and opinions of similar versus dissimilar others are more likely to be imitated (Guéguen & Martin, 2009). Moreover, requests from similar others are more likely to be complied with than requests from dissimilar others, suggesting that similarity directly affects the degree of social influence (Burger, 2004).

With regard to the role of similarity, we thus hypothesized that any bias in the following of behavioral patterns of the observed agent should be exaggerated (i.e. social learning rates should be increased) if this person shares a characteristic feature with the observing agent.

2. General Methods

2.1 The social reversal learning task

On each trial of the present task, participants observed the response of a (fictitious) other player before they were required to make their own choice. Reversal learning performance was assessed in two different blocks, examining choice behavior without (private/baseline block, Fig. 1A) and with (social block, Fig. 1B) exposure to the choices

made by a (fictitious) previous participant. In both blocks, participants learned to choose one of two simultaneously presented colors (‘blue’ and ‘green’) by receiving monetary reward (+1 pence [p]) or punishment (-1p) contingent on their choice (e.g. +1p for ‘blue’ and -1p for ‘green’).

After a variable number of trials unknown to the participants, reward/punishment contingencies were reversed so that the previously rewarded stimulus was now punished and vice versa (-1p for ‘blue’ and +1p for ‘green’). A varying number (1-3) of ProbErrs was interspersed at random positions between two reversals so that a normally correct response was unexpectedly punished. The total number of reversal and ProbErr events was matched between the different conditions. Using ProbErr trials (and thus a ‘pseudo-probabilistic’ approach) rather than a truly probabilistic design (with correct choices being rewarded, for instance, with a probability of 0.9 and punished with a probability of 0.1) allowed us to control the number of punishments (and thus objective performance of the other player) across social conditions. Overall, the reward ratio in our pseudo-probabilistic setup (including 1-3 punishments per 7-15 trials in Study 1) was comparable to a reward probability of 0.8/0.2, typically used in other probabilistic reward learning tasks.

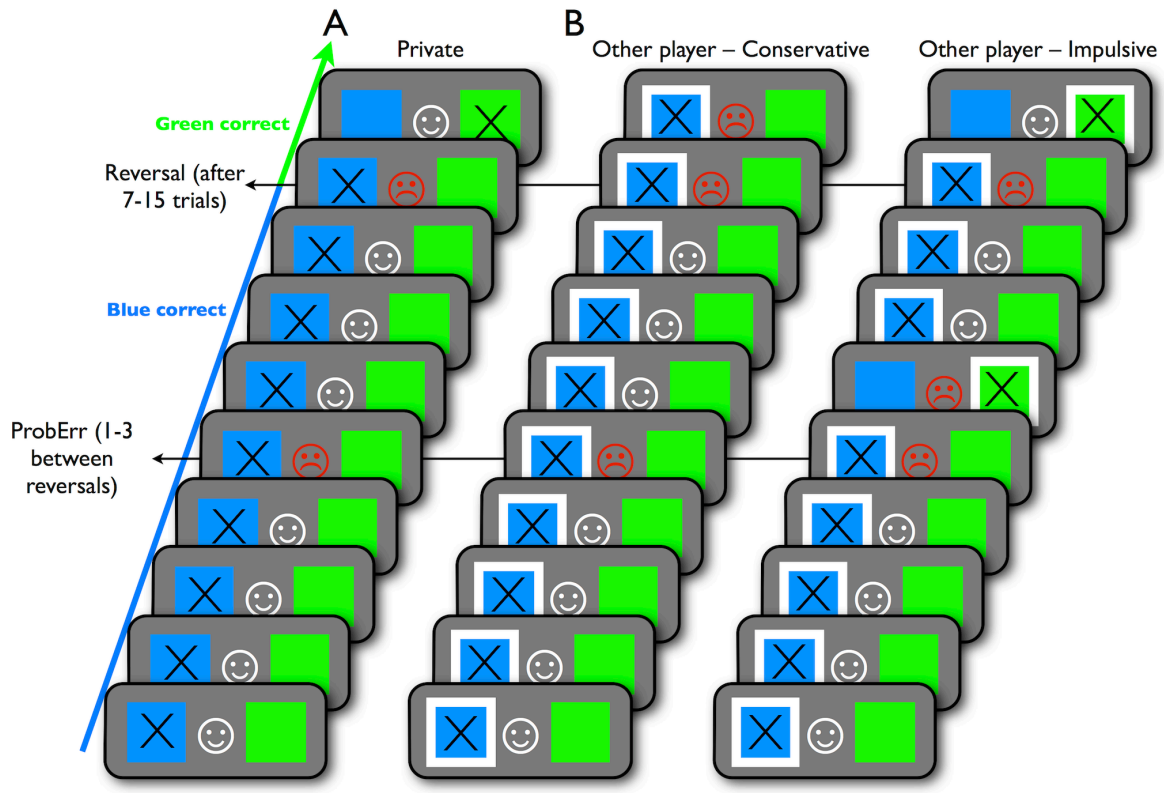

Fig. 1. Design of the social reversal learning task illustrating one reversal episode. (A) Private condition. Crosses illustrate hypothetical choices. Feedback was given instantaneously after a choice had been made. A varying number of probabilistic errors showing 'wrong' feedback after correct choices were presented before each reversal. (B) The two social conditions involving the presentation of conservative versus impulsive choices made by a (fictitious) other player who either shared the same birthday as the participant (similar-other group) or not (dissimilar-other group). The other player's choices were indicated by a white frame surrounding his/her choice before the real participant made his/her choice. Crosses show hypothetical choices reflecting imitative response behavior. Note that participants were exposed to choices made by ONE other player who displayed conservative AND impulsive choices in different reversal episodes. As illustrated, the number of errors made by the other player was balanced between the conservative and impulsive condition.

2.2 Choice behavior and similarity of the other player

In each trial of the social block, choices made by the other player were presented before participants made their own choice. (i) The first experimental manipulation concerned the response of the other player responded in the first trial (+1) after an unexpected punishment. We manipulated the other player's choices to simulate (a) 'conservative' (stay response) versus (b) 'impulsive' (switch response) choice behavior as described above (1.1). After each ProbErr in a given reversal episode (between two consecutive reversals) the other player consistently made either conservative (stay) or impulsive (switch) responses. The order

of conservative/impulsive reversal episodes was randomized with the constraint of yielding a specific number of episodes per condition (see Methods Study 1 and 2). Changing the other player's choice behavior between episodes (but not within episodes) aimed to simulate dynamic changes in decision-making strategies as observed in natural contexts. In addition, after a few reversal events the conservative player perseverated to the previously rewarded stimulus for two or three trials after the reversal. Importantly, inclusion of these 'social perseveration' trials helped to balance the number of 'errors' (wrong choices) committed by the other player in the conservative versus impulsive condition: There were 30 observed conservative versus 20 observed impulsive errors in Study 1 and 12 conservative and 16 impulsive errors in Study 2. Thus, overall following conservative versus impulsive choices was comparably beneficial/detrimental, with a slight disadvantage for conservative choices in Study 1 and the opposite pattern in the replication study (Study 2) so that any effects related to the objective performance of the other player were counterbalanced across the two studies. It should be noted that in contrast to previous studies (e.g. Burke et al., 2010), we did not manipulate and present outcomes of the observed player separately. Nonetheless, as participants were made to believe that the other player was exposed to the exact same reward/punishment contingencies, they could easily infer after each trial whether the observed choice was correct or incorrect.

(ii) The second manipulation was based on an 'incidental similarity' technique (Burger, Messian, Patel, del Prado, & Anderson, 2004) that allowed us to create an association between the participant and the observed agent without providing information relevant to the task: Before the social block participants were presented with a screen showing basic information about the other player (see Methods Study 1). For half of the participants, this information included a date of birth of the observed player matching the day and month of their own birthday (similar-other group), whereas for the other half the birthdays did not match (dissimilar-other group). Previous research has demonstrated that this manipulation

critically determines how information about others influences the self (Brown & Novick, 1992).

3. Study 1

3.1 Methods

3.1.1 Participants

34 volunteers participated in Study 1. Data from two participants were excluded due to the use of wrong response keys or non-compliance with instructions (remaining sample: 27 female, 5 male; mean age $M = 20.82$ years, $SD = 4.67$). All participants (Studies 1 and 2) were members of Cardiff University who gave written informed consent to take part in the studies that had been approved by the School of Psychology's Ethics Committee, Cardiff University. For taking part, participants received course credits or were paid £6 per hour.

3.1.2 Materials and procedure

The screen shown throughout the learning session consisted of two colored squares (blue and green) presented left and right of the center of the monitor on a black background. At all times, participants also could see their accumulated earnings (£), shown centrally below the colors and updated in each trial synchronously with the feedback. The beginning of each choice trial was indicated by the appearance of two (private block) or one (social block) white frame(s) surrounding (one of) the colors. These response cues remained on the screen until the participants made their choice by pressing the '1' or '3' key on the keyboard. A black cross inside the chosen colored square, shown for 500 ms after response onset, indicated the participant's choice. Feedback was then given using a white 'smiley' (correct choice) or red 'frowny' (incorrect choice), centrally presented for 1000 ms. Trials were separated by an Inter-Trial-Interval (ITI) randomly varying between 700 and 1400 ms.

Private and social blocks were presented in counterbalanced order. The private block consisted of 12 reversal episodes (6 blue correct, 6 green correct episodes), containing a pseudo-randomized number of choice trials that varied between 7 and 15 between two

reversals (mean episode length = 11 trials). Within each reversal episode we included 1-3 probabilistic error (ProbErr) trials in which 'wrong'-feedback was given for correct choices (and 'correct'-feedback for wrong choices), even though the reward contingencies had not changed. Across the 12 reversal episodes, we presented 20 ProbErr trials, with 6 episodes containing only 1 ProbErr, 4 episodes containing 2 ProbErrs and 2 episodes containing 3 ProbErrs in random order. Within a reversal episode, ProbErrs could occur at a random position between the third trial after the previous reversal and two trials before the next reversal. We implemented the same trial structure as above in the social block so that each condition was matched with regard to the number of choice trials and the number of ProbErrs. The social block comprised 36 reversal episodes. Before the social block participants received written instructions that they would now be able to observe what an anonymous other player – who received exactly the same learning sequence as they – chose in a previous study. 15 of the participants were presented with a similar player and 17 with a dissimilar player (see below).

In trials after reversals the other player made impulsive (= correct) choices in 24 of the 36 reversal episodes while in the other 12 episodes he/she made conservative choices, perseverating to the previously rewarded color for 2 (6 episodes) or 3 (6 episodes) trials and leading to a total of 30 social perseveration trials. Vice versa for +1 trials after ProbErrs, participants observed conservative choices of the other player in 24 episodes and impulsive choices in the 12 other episodes. The order of episodes with conservative versus impulsive choices was freely randomized. In ProbErr and reversal trials, the other player always responded incorrectly. Except for ProbErr/reversal trials and the trials after ProbErrs/reversals (+1 and perseveration trials), the other player always responded correctly (standard trials).

3.1.3 Social cover story and similarity manipulation

Before the social block, participants were instructed to initiate a 'random generator' on the computer that selected a specific participant from a (fictitious) other study, showing a

rapid sequence of 'subject codes' (e.g. PJ_008) that seemingly stopped at a random point. Participants were told that they would see choices made by the selected participant and received some basic information about the other player on the screen. Based on a pre-experimental demographic questionnaire, the experimenter had manipulated this information shortly before testing while the participants had been waiting in a separate room. It included (i) the other player's 'gender' (either 'male' or 'female' and always matching the participant's own gender), 'date of data collection' in the previous study (always '3 Nov 2010'), and (iii) the other player's birthday. While the birth year was always '1989', the other player's day and month of birth was manipulated to either match the real player's birthday or not (default birthday: '25 Apr').

3.1.4 Data analysis

3.1.4.1 Accuracy analysis

We assessed social influences on reversal learning by calculating for each participant trial-based average accuracy scores (% choices corresponding to the correct color of each reversal episode) as a function of condition (private, conservative and impulsive) and relative trial position within a reversal episode. For calculating accuracy scores for trials surrounding critical events (ProbErrs or reversals), only trials were included that were unaffected by a second ProbErr or reversal either in the preceding trial, in the trial before the preceding trial or in the trial directly following the target trial.

As shown in Fig. 3, apparent differences between the private, conservative and impulsive conditions occurred only in the first two trials (+1, +2) after ProbErr or reversal events, that is, there were no accuracy (switch/stay tendency) differences between the private versus social conditions in standard trials. We used planned comparisons (one-tailed paired t-tests) in post-ProbErr trials +1 and +2 to test for accuracy increases in the conservative condition and accuracy decreases in the impulsive condition, respectively, relative to the private condition. We tested for the opposite pattern (accuracy decrease in conservative,

accuracy increase in impulsive condition) in post-reversal trials +1 and +2. Importantly, by measuring social influence in the conservative versus impulsive condition as a combined pattern of accuracy increase and decrease, any global performance difference between the private and social block was unlikely to confound the present effects. Effect sizes for significant differences between the two social conditions and the private condition were calculated using Cohen's *d* for paired samples ($d = D / SD_D$, where D is the mean difference score and SD_D is the standard deviation of the difference scores).

3.1.4.2 Computational modeling

In addition to analyses of choice behavior averaged across trials, we used computational modeling to investigate the relative contribution of reward-based and social trajectories to drive single-trial choices. Specifically, we tested how well standard reinforcement learning algorithms (Q-learning; Watkins & Dayan, 1992) predicted individual choices in the social block and compared this to several social Q-learning variants that we developed to model observational factors.

Standard Q-learning has successfully been shown to predict responses in a variety of (non-social) learning tasks, including reversal learning (Jocham et al., 2009). It assumes that the choice between two stimuli A and B is determined by action values (Q-values) that are associated with each choice option and can vary between 0 and 1. Q-values are updated after each choice trial taking into account the reward prediction error δ^r, that is, the difference between the observed outcome (= 1 or 0 in binary learning) and the expected outcome (= current Q-value) so that $\delta^r = \text{reward}(t) - Q(t)$. The reward prediction error is additionally weighted by the learning rate α^r so that in trial t + 1, Q is recalculated by $Q(t+1) = Q(t) + \alpha^r * \delta^r(t)$.

Since reward/punishment associations in our task were mutually exclusive, we updated Q-values after each trial for both the blue color (Q_{blue}) and the green color (Q_{green}), with $Q_{green} = 1 - Q_{blue}$. At the start of the private and social learning task, we set both Q-values

to 0.5. Before model predictions were tested against the observed choices, we transformed Q-values into action selection probabilities (P_{blue} or P_{green}), using a softmax function (Sutton & Barto, 1998). The softmax function (see Fig. 2) estimates the probability of the participant choosing a specific option in each trial by adjusting Q-values by an estimate of the 'temperature' β (= degree of randomness in the participant's choices).

In our social Q-learning variants we modeled observed choices of the other player as a 'social prediction error' δ^s. Similar to the standard reward prediction error δ^r, we modeled δ^s as the difference between an internally stored representation of each choice option's value and an external event revising this value. In particular, we defined δ^s as the extent of how much the *observed* choice matches or deviates from the participant's current choice preference as defined by the higher of the two current Q-values, i.e. δ^s = observed choice(t) – Q(t): If Q_{blue} was higher at the start of trial t (before the other player's choice was seen and feedback was given), observing the other choosing blue resulted in a $\delta^s(t)= 1 – Q_{blue}(t)$, whereas a green choice by the other was modeled as $\delta^s(t)= 0 – Q_{blue}(t)$. Vice versa, when $Q_{green}(t)$ was higher than $Q_{blue}(t)$ at the start of trial t, we defined $\delta^s(t)= 0 – Q_{green}(t)$ when the other player chose blue and $\delta^s(t) = 1 – Q_{green}(t)$ when the other player chose green. As a result, conservative choices by the other typically elicited a positive δ^s while impulsive choices evoked a negative δ^s It should be noted that positive (termed 'conservative' δ^s hereinafter) versus negative δ^s (termed 'impulsive' δ^s hereinafter) should not be confused with positive versus negative reward prediction errors (δ^r) in the probabilistic learning literature. In contrast to δ^s whose sign we derived from the correspondence to the dominant choice propensity, the sign of the standard δ^r is defined based on choice outcome (reward versus punishment). Social prediction errors were additionally weighted by the social learning rate α^s.

We used three classes of models, differing in whether and how δ^r and δ^s were combined to update Q-values: (i) *Standard* (non-social) Q-learning: Q-values were updated only by δ^r weighted by α^r (with α^s set to 0), after a choice had been made/outcome had been

received ('Social Ignorer'); (ii) *imitative* learning: Q-values were determined only by δ^s weighted by α^s (with α^r set to 0), after the other player's choice had been observed but before one's own choice was made ('Blind Follower'), (iii) *integrative* learning: Q-values were updated by both δ^r and δ^s.

Recent studies showed that with regard to overall performance measures in probabilistic reward learning tasks, integrative learning strategies are superior relative to purely social (imitative) or non-social (reward-based only) strategies (Burke et al., 2010; Selbing, Lindström, & Olsson, 2014). It should be noted though that in the present studies we did not aim to examine the general adaptivity of different social/non-social learning strategies but rather compared how different types of observed choice behaviors affect the integration of social information into learning and thus the degree of social influence (i.e. the frequency of following observed choices). Specifically, we examined whether learning rates would reflect any biases in the integration of social information, such as increased weighting of expectation-conforming (= conservative) choices or increased weighting of choices made by similar versus dissimilar others.

Based on our expectation of differential effects of observed conservative versus observed impulsive choices (see 1.2), the integrative model was further differentiated into a sub-model that used a single α^s to weight conservative and impulsive δ^s and a sub-model that used different α^s for conservative and impulsive δ^s. Moreover, we distinguished *instantaneous* integrative models (Q-values were updated by δ^s instantaneously, that is, after the other player's response had been observed and before one's own choice was made; in a second step the resulting Q-values were then updated by δ^r after the outcome had been received) and *delayed* integrative models (Q-values were updated by δ^s with a delay, that is, after the outcome had been received and after Q-values had been updated by δ^r). In contrast to the instantaneous models, the delayed models were able to capture information about the correctness of the other player (the magnitude and sign and thus the differential weighting of

δ^s was determined *after* Q-values had been updated by the outcome). Note also that we based δ^s on Q- rather than P-values (action probabilities derived from the softmax function) since the latter yielded on average lower goodness-of-fit values (cf. Lindström & Olsson, 2015). Figure 2 shows a schematic illustration of the integrative models.

We used the Akaike Information Criterion (AIC; Akaike, 1974) based on Log Likelihood Estimates (LLE) as an indicator of each model's ability to predict empirical choices. LLEs use the log of the cumulative product of those action probabilities (P_{blue} or P_{green}, see above) that match the actual choice in each trial t: $LLE = \log(\Pi_t P_{choice}(t))$. We derived LLEs for each participant and model by extracting the highest LLE (= highest goodness-of-fit corresponding to the least negative LLE value) obtained from iterations of free parameters between 0 and 1 using increments of 0.05. LLEs were then transformed into AICs by AIC = -2*LLE + 2*k, where k is the number of free parameters, with lower values indicating superior fit. The rationale of the AIC is to penalise models with a higher number of free parameters in order to counteract the confounding role of model complexity (more complex models usually show a better data fit). AICs and free parameters (reward and social learning rates α^r and α^s, temperature β) were further analyzed by repeated-measurement ANOVAs and paired t-tests.

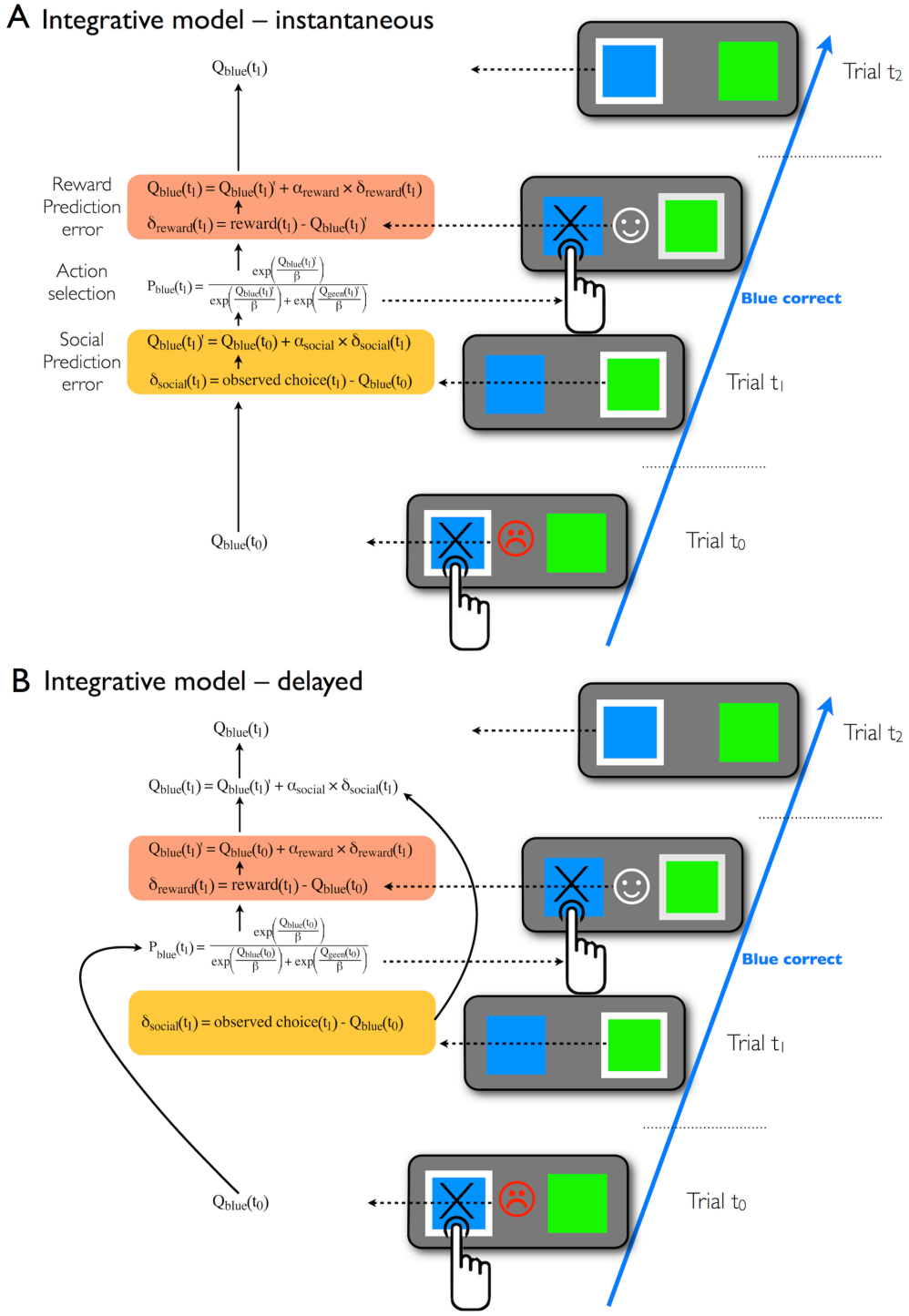

Fig. 2. Schematic illustration of two integrative models used to predict trial-by-trial choices during social reversal learning. (A) Instantaneous integration model. After receiving feedback in trial t_0, in trial t_1 the participant sees the other player either staying with the previous choice or switching to the other color. The resulting social prediction error δ_{social} has a positive (= conservative δ_{social}) or negative (= impulsive δ_{social}) sign and is instantaneously used to update $Q_{blue}(t_0)$ to $Q_{blue}(t_1)'$ taking into account the social learning rate α_{social}. $Q_{blue}(t_1)'$ then determines the action selection probability $P_{blue}(t_1)$ based on a softmax function. After the participant has made his/her choice and the outcome has been received in trial t_1, $Q_{blue}(t_1)'$ is then updated by the reward prediction error δ_{reward} weighted by the reward learning rate α_{reward} to $Q_{blue}(t_1)$. $Q_{green}(t_1)$ is computed as 1 - $Q_{blue}(t_1)$. (B) Delayed integration model. The only difference to the instantaneous integration model is that δ_{social} of trial t_1 is not contributing to the action selection probability $P_{blue}(t_1)$ but becomes only effective after the outcome has been obtained, that is, after $Q_{blue}(t_1)$ has been updated by δ_{reward} to $Q_{blue}(t_1)'$.

3.2 Results

3.2.1 Accuracy after probabilistic errors

In the trial following the first ProbErr in a reversal episode (+1), participants performed better in the conservative condition relative to the private condition, $t(31) = 2.00$, $p = 0.027$, Cohen's $d = 0.35$ (Fig. 3A). We also found a trend-level facilitation for the conservative condition in trial +2 after the first ProbErr, $t(31) = 1.58$, $p = 0.063$, $d = 0.28$. In contrast, we did not find a reduction in accuracy in the impulsive condition in the first two trials after the first ProbErr, ts < 0.30, ps > 0.38. For 'later' ProbErrs in a reversal episode – data was collapsed across 2nd and 3rd ProbErrs due to the small number of episodes presenting three ProbErrs – we found a delayed benefit, showing increased accuracy in the conservative versus private condition in trial +2, $t(31) = 2.38$, $p = 0.012$, $d = 0.42$, but not in trial +1, $t(31) = 0.62$, $p = 0.27$ (Fig. 3B). Again we did not find an accuracy impairment in the impulsive relative to the private condition in trials +1 and +2, ts < 1.0, ps > 0.16.

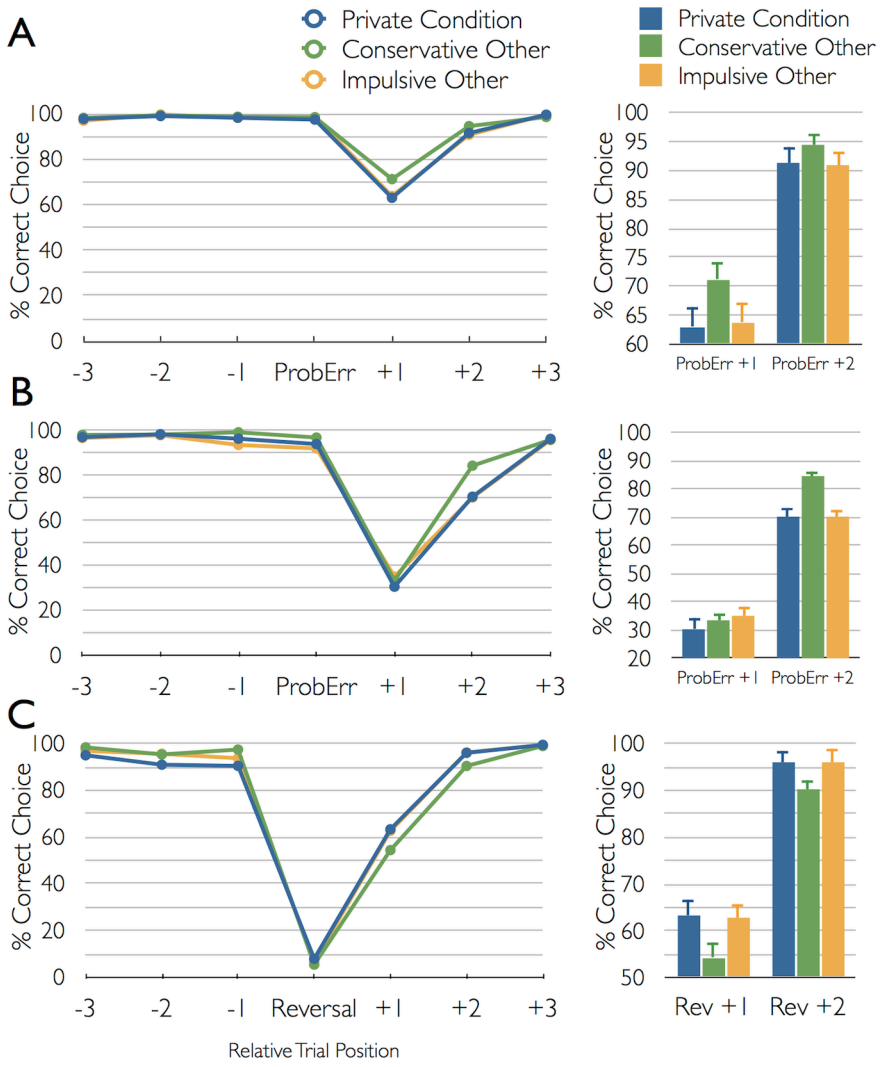

Fig. 3. Choice accuracy (for the truly correct color) as a function of trial position and social influence in Study 1. (A-B) Accuracy for trials -3 before to +3 after probabilistic errors (ProbErr, left) and in more detail for post-ProbErr trials +1 and +2 (right), split for first (A) and late (second/third, B) ProbErrs in a reversal episode. (C) Accuracy for trials -3 before to +3 after reversals (left) and in more detail for post-reversal trials +1 and +2 (right). Error bars show standard errors of the mean.

3.2.2 Accuracy after reversals

Consistent with the post-ProbErr results, social influence was restricted to the conservative other. After reversals this led to a *decrease* of accuracy since the conservative player perseverated to the previously correct color. These 'socially induced perseverations' were apparent in trial +1, $t(31) = 1.93$, $p = 0.032$, $d = 0.34$, and +2, $t(31) = 2.26$, $p = 0.016$, $d = 0.40$ (Fig. 3C). Observing the impulsive player who switched to the newly correct color did not improve accuracy in trials +1 or +2 relative to the private condition, ts < 0.22, *p*s > 0.41.

To summarize, both post-ProbErr and post-reversal results showed that participants frequently followed a conservative (non-switching) other player after their choice had been unexpectedly punished in the previous trial, while an impulsive player was ignored.

3.2.3 Computational modeling

3.2.3.1 Model fit comparisons

Paired t-tests (two-tailed) showed that all integrative models performed substantially better than the standard Q-learning ('social ignorer') and imitative ('blind follower') models, all ts > 2.96, all $ps < 0.007$ (see Table 1 for an overview of AICs). This replicates previous work highlighting that both social (observed choices) and non-social (observed consequences) information are dynamically combined during reward (Burke et al., 2010) or aversive (Selbing et al., 2014) learning tasks.

Table 1. Number of free parameters, means (standard deviations) of Akaike Information Criteria (AIC) and learning rate (α^r and α^s) estimates for reward and social prediction errors (δ) of all models in Study 1.

Model	Free parameters	AIC	α^r	α^s (cons./ imp. δ)	α^s (cons. δ)	α^s (imp. δ)
Standard Q-learning	2	211.96 (54.89)	0.45 (0.08)	–	–	–
Imitative – same α^s	2	373.19 (39.30)	–	0.8 (0.11)	–	–
Imitative – different α^s	3	370.60 (40.22)	–	–	0.93 (0.25)	0.66 (0.18)
Instantaneous integration – same α^s	3	205.40 (54.55)	0.46 (0.08)	0.06 (0.07)	–	–
Instantaneous integration – different α^s	4	205.54 (54.53)	0.46 (0.07)	–	0.09 (0.12)	0.06 (0.07)
Delayed integration – same α^s	3	205.87 (55.43)	0.53 (0.12)	0.11 (0.10)	–	–

Delayed integration – different α^s	4	203.84 (56.38)	0.47 (0.06)	–	0.22 (0.14)	0.05 (0.06)

To further analyze the integrative models, we entered their AICs into a repeated-measurement ANOVA with *time of integration* (instantaneous vs. delayed) and *social learning rate* (same α^s or different α^s for conservative and impulsive δ^s) as factors. We found a main effect of *social learning rate*, $F(1, 31) = 4.53$, $p = 0.041$ and an interaction *time of integration* x *social learning rate*, $F(1, 31) = 5.60$, $p = 0.024$, indicating a better fit for delayed integration for the models using different learning rates specifically (see Fig. 4). Together these results suggest a clear predictive advantage for models integrating reward and social information over models relying on one source information. Modeling also provided some evidence for a differential integration of observed conservative versus impulsive choices but, given the large variability between individual AICs, overall these effects were weak.

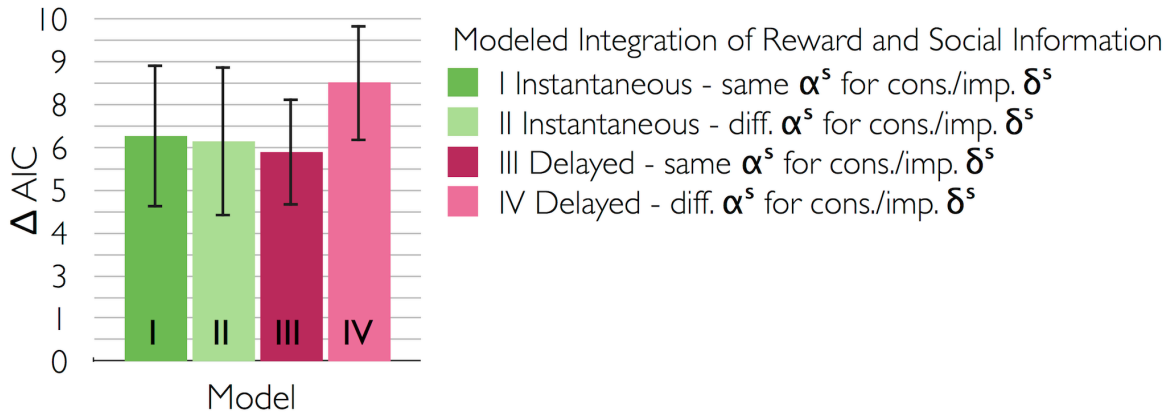

Fig 4. Model fit comparisons in Study 1 using the Akaike Information Criterion (AIC), which corrects for the number of free parameters. The figure shows the differences ΔAIC between the standard Q-learning model, which is based on δ^r only ('social ignorer'), and the 4 integrative models in which both reward and social information are utilized, with the integration occurring either instantaneously or with a delay, and with applying either the same or different learning rates to conservative/impulsive δ^s. A larger ΔAIC indicates superior performance. Error bars show standard errors of the difference.

3.2.3.2 Social learning rates

We further examined individuals' social learning rates (α^s) derived from the free parameter estimations associated with the best performing model (delayed integration of δ^s

and δ^r, different α^s for conservative versus impulsive δ^s). Paired t-tests showed that α^s was substantially higher for conservative δ^s ($M = 0.22$, 95% CI [0.17, 0.27]) versus impulsive δ^s ($M = 0.05$, CI [0.02, 0.07]), $t(31) = 6.92$, $p < 0.001$, $d = 1.22$, suggesting that participants weighted others' responses more strongly if they conformed with their current choice preference. One might argue that conservative δ^s received a larger α^s because they were smaller in magnitude than impulsive δ^s so that net effects on Q-values were balanced. However, the average product of conservative δ^s and conservative α^s (i.e. the net effect on Q-values) was still larger than then average product of impulsive δ^s and impulsive α^s: Mean conservative α^s x mean conservative δ^s = 0.22 x 0.44= 0.10; mean impulsive α^s x mean impulsive δ^s = 0.05 x 0.57 = 0.03. To summarize, the other player's choice was more strongly weighted for conservative versus impulsive social prediction errors, mirroring the 'conservative bias' observed at the level of accuracy data.

3.2.3.3 Incidental similarity

We next tested for increased social learning rates (derived from the best-performing delayed, α^s differentiated model) in those participants who were observing a similar versus dissimilar player. ANOVA of α^s values using *group* (similar versus dissimilar other) and *type of social prediction error* (conservative versus impulsive) showed a significant main effect of type of α^s (with higher α^s for conservative δ^s), $F(1, 30) = 54.16$, $p < 0.001$. We also found a marginally significant interaction between *group* and *type of* α^s, $F(1, 30) = 4.04$, $p = 0.05$. Follow-up tests (one-tailed independent-samples t-tests) revealed that for the best performing model the similar-other group showed a higher α^s for conservative δ^s relative to the dissimilar-other group (see Fig. 5), $t(31) = 2.06$, $p = 0.024$, $d = 0.72$. In contrast, similarity did not lead to a higher α^s for impulsive δ^s, $t(31) = 0.12$, $p = 0.45$. Thus, similarity acted to exaggerate the bias towards expectation-consistent responses of the other player specifically. Importantly, we found no group differences for similarity in any of the other learning parameters, α^r: similar other: $M = 0.47$, $SD = 0.07$; dissimilar other: $M = 0.47$, $SD = 0.06$; $t(30) = 0.30$, $p = 0.77$; β:

similar other: $M = 0.19$, $SD = 0.08$; dissimilar other: $M = 0.19$, $SD = 0.06$; $t(30) = 0.05$, $p = 0.96$. To summarize, interpersonal similarity boosted learning from conservative social prediction errors but did not lead to a general increase of social influence.

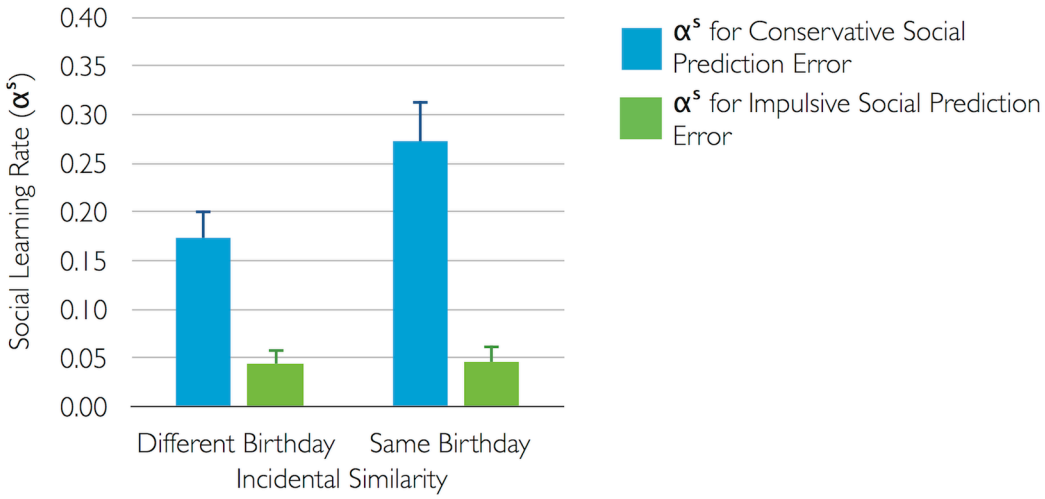

Fig. 5. Mean learning rates for δ^s derived from the best-fitting computational model and split for direction (conservative versus impulsive) of the δ^s and the two groups of participants who observed another player either sharing (similar-other group) or not sharing (dissimilar-other group) their own birthday. Error bars show standard errors of the mean.

4. Study 2

The aim of Study 2 was to control for non-social responses biases resulting from our method of presenting choices made by another player by a lateral, visual cue. In particular, marking one of the two choice options by a white frame may have biased responses at a simple perceptual level through exogenous spatial cueing. In Study 2 we used the same paradigm as in Study 1 – showing a white box around one of the colors that switched or did not switch to the other color after punishments – *without* presenting the social cover story (see 3.1.3) before the task. A spatial cueing account would predict that spatial cueing alone (without social information) will lead to similar results as in Study 1, specifically an accuracy difference between the conditions in which the correct versus incorrect response was cued (i.e., increased accuracy for conservative trials relative to impulsive trials after PEs and vice versa after reversals). In contrast, if omitting social information removes or reduces the

accuracy difference between the differentially cued trials, the effects observed in the previous study could be attributed to social-cognitive processes.

4.1 Methods

4.1.1 Participants

19 students (15 female, 4 male; mean age $M = 20.67$ years, $SD = 1.37$) completed the reversal learning task and received course credits or payment for their participation. All participants provided written informed consent, and the protocol had been approved by the ethics committee of the School of Psychology, Cardiff University.

4.1.2 Materials and procedure

The reversal learning paradigm was identical to Study 1 with the exception of the initial task instruction. Instead of being informed about a randomly selected other player whose responses participants would observe in the social block, they were simply told that they would see additional visual information without any further specification of the nature of this information. As in the previous study, order of the private and 'social' block was counterbalanced across participants.

4.1.3 Data analysis

We calculated % correct choices for the two trials following ProbErrs and reversals as a function of the 'choice' displayed by the white frame (stay [=conservative] versus switch [=impulsive]). As a spatial cueing account would predict maximal differences for trials cueing opposite responses, we computed the difference between stay- and switch-cued trials and directly contrasted this measure with the same difference calculated for Study 1, using between-group comparisons (one-tailed two-sample t-tests). In addition, we compared average accuracy for stay-cued and switch-cued trials with 'private' performance (no lateral visual cue) in the non-social control experiment (Study 2) separately, using paired t-tests.

4.2 Results

As illustrated in Figure 6, reversal learning performance levels in Study 2 were comparable to those in Study 1. Importantly, we found cross-study evidence that accuracy differences between the conservative and impulsive condition in critical trials (stay-cues versus switch-cues) were smaller in the non-social experiment (Study 2) relative to the social version of the same paradigm (Study 1). The difference reached significance for post-PE trials +2 $t(49) = 1.70$, $p = 0.048$, and trend-level when accuracy was computed across PE positions and +1/ +2 trials, $t(49) = 1.43$, $p = 0.080$. For post-PE trials +1 and post-reversal trials, differences did not reach significance, $ts < 1.01$, $ps > 0.15$.

Separate analyses of accuracy data in Study 2 comparing conservative/impulsive trials with private trials provided further evidence that cueing one of the choice options did not modulate choice accuracy. No significant differences between the social and the private conditions emerged for post-PE trials: First ProbErr: all $ts < 0.73$, *all ps* > 0.47; later ProbErrs: conservative trial +1 versus private +1: $t(18) = 0.90$, $p = 0.38$, conservative +2 versus private +2: $t(18) = 1.19$, $p = 0.25$, impulsive +1 versus private + 1: $t(18) = 1.90$, $p = 0.074$, impulsive +2 versus private + 2: $t(18) = 0.57$, $p = 0.58$. Moreover, in the first trial after reversals accuracy was not reduced by observed 'perseverations' (conservative condition), $t(18) = 0.47$, $p = 0.64$, nor was it improved by cued switches (impulsive condition), $t(18) = 0.46$, $p = 0.65$, relative to the private block. Cued switch responses did also not alter accuracy in trial +2 after reversals, $t(18) = 0.02$, $p = 0.98$. The only result comparable to Study 1 was a decrease in performance in the conservative condition relative to the private condition in trial +2 after reversals, $t(18) = 1.88$, $p = 0.077$ (two-tailed).

However, we suspect that this influence on choice accuracy was due to a residual attribution of social properties to the spatial cue by the participants themselves (6 participants actually reported during debriefing after the task that they believed that the white box showed responses by another player). Moreover, while spatial cueing predicts increased accuracy in the conservative relative to the private condition after probabilistic errors, it predicts an

analogous reduction in accuracy for the impulsive condition (which cues switch responses) in those trials, which we did not find in either study. Overall, results of Study 2 thus suggest that presenting choice cues in a non-social context removes (or at least substantially reduces) the specific effects resulting from the experimental manipulations in the previous studies, providing supporting evidence for their social nature.

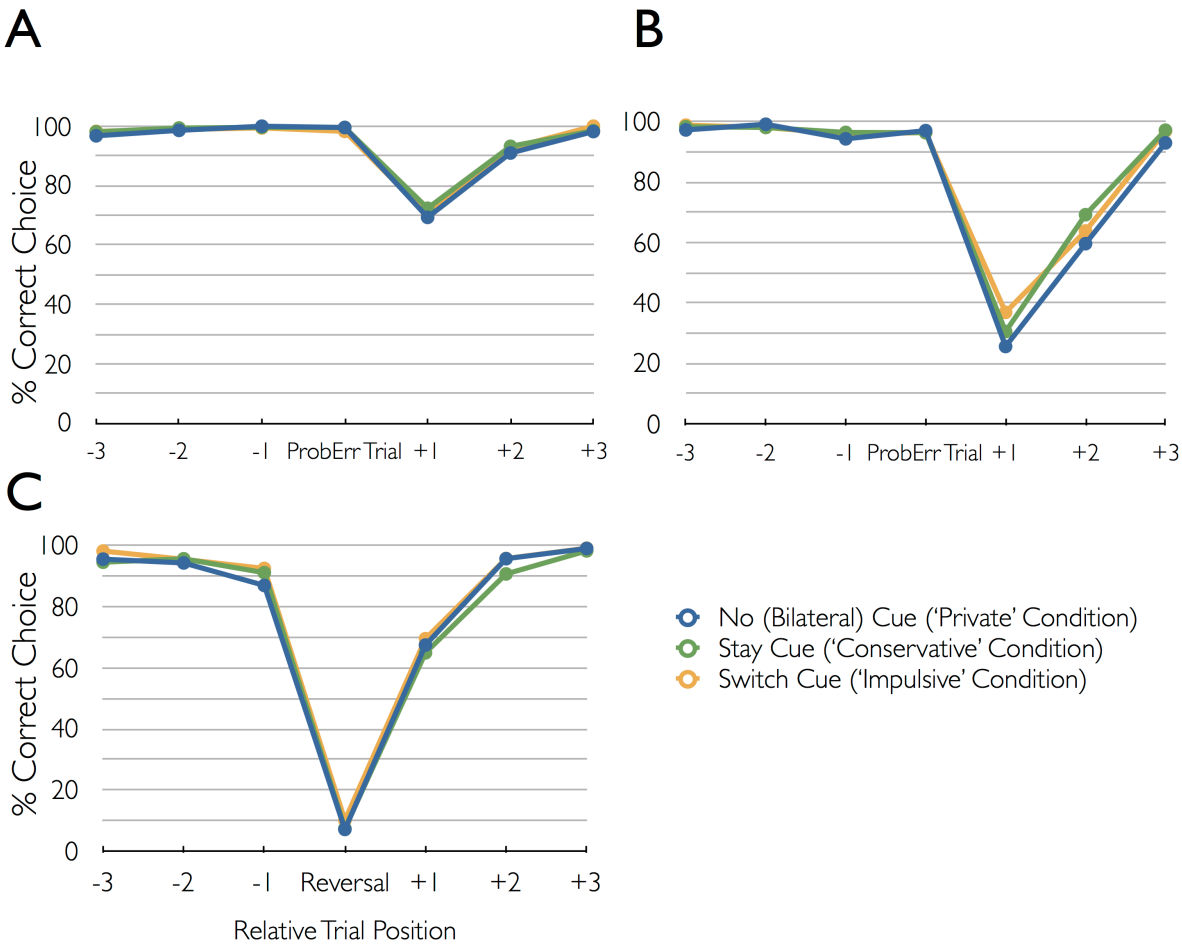

Fig. 6. Choice accuracy as a function of trial position and lateral cue in the non-social control experiment (Study 2). (A-B) Accuracy for trials -3 before to +3 after probabilistic errors (ProbErr), shown for first (A) and late (B) ProbErrs in a reversal episode. (C) Accuracy for trials -3 before to +3 after reversals. Error bars show standard errors of the mean.

5. Study 3

The first goal of Study 3 was to replicate the general finding of social influence on reversal learning (Study 1) using data from a different and larger sample. The second goal of Study 3 was to further investigate the differential effects of observed conservative versus impulsive choices. Specifically, by reducing the number of trials per reversal episode, Study 3

allowed to examine the role of reversal volatility on weighting different types of observed choice behavior. Thus, it is possible that the influence of observed impulsive choices increases with increased reversal volatility, as higher volatility may favor exploratory choice behavior. No similarity manipulation was included in Study 2.

5.1 Methods

5.1.1 Participants

Data were acquired as part of a multimodal genetic imaging project at Cardiff University in which 100 volunteers performed various cognitive tasks while undergoing functional magnetic resonance imaging (MRI). All participants provided written informed consent, and the protocol had been approved by the ethics committee of the School of Psychology, Cardiff University. Reversal learning data from 7 participants were incomplete due to technical problems. All data from the remaining sample (N = 93; 59 female, 34 male, 24.39 years ± 4.54 [SD]) were used for analysis.

5.1.2 Materials and procedure

The overall task structure was identical to the first study but reversals occurred after a fewer number of trials (pseudo-randomized between 7-11, mean episode length = 9 trials). In addition, the following task structure was implemented: (i) The task was split into three blocks, one private block and two social blocks (social blocks were always presented successively), each containing 12 reversal episodes. Each of the two social blocks contained 6 conservative and 6 impulsive episodes. (ii) Due the shorter reversal episodes we reduced the number of ProbErr trials: There were either one or two ProbErrs between two reversals; per condition (private, conservative, impulsive) 8 reversal episodes contained 1 ProbErr and 4 episodes contained two ProbErrs. (iii) In the social blocks half of the post-ProbErr trials (+1) showed conservative choices made by the other, and half of the trials showed impulsive choices. (iv) After each conservative reversal episode, the other player perseverated for one trial (+1), while after each impulsive episode the other player made an impulsive (correct)

choice. (v) The task was not self-paced as in the previous study but stimuli (for instance, presentation of other player's response or reward feedback) had randomized durations in the range of 0.75 s to several seconds.

5.1.3 Data analysis

We used the same methods as in Study 1 to calculate average accuracy scores. Similarly, the same algorithms as in the first experiment were applied to model trial-by-trial choices. The only modifications of the existing functions were as follows: (i) We added a mathematical rule dealing with 'miss' responses – in contrast to the (self-paced) first study, misses could occur in Study 3 as choice cue and feedback stimulus were presented for a limited amount of time. To keep the number of learning trials contributing to LLE scores comparable, an action probability P_{blue}/P_{green} of 0.5 (that is, indecisiveness between blue and green choice) was assumed for missed trials. (ii) The sign of δ^s (positive [= conservative] versus negative [= impulsive]) was not determined by the currently higher Q-value (Q_{blue} versus Q_{green}) but by the more frequent choice of the participant in the last three trials before the predicted trial within a given reversal episode: If the dominant choice in the last three trials was 'blue' at the start of trial t, observing the other choosing 'blue' resulted in a $\delta^s(t)= 1 – Q_{blue}(t)$, while a 'green' choice by the other resulted in as $\delta^s(t)= 0 – Q_{blue}(t)$. Vice versa, when the dominant choice in the last three trials was 'green', observing the other choosing 'blue' resulted in $\delta^s(t) = 0 – Q_{green}(t)$ while a 'green' choice led to $\delta^s(t) = 1 – Q_{green}(t)$.

5.2. Results

5.2.1 Accuracy

Despite the changes in task structure relative to Study 1, paired t-tests confirmed selective facilitation in the conservative condition after ProbErr trials, showing an accuracy benefit in the two trials after the first (trial +1: $t(92) = 5.51$, $p < 0.001$, $d = 0.57$; +2: $t(92) = 3.15$, $p = 0.001$, $d = 0.33$) and second ProbErr (+1: $t(92) = 4.38$, $p < 0.001$, $d = 0.45$; +2: $t(92) = 4.37$, $p < 0.001$, $d = 0.45$), relative to the private condition (see Fig. 7). Again, the (wrong)

impulsive player did not reduce performance after the first (+1: $t(92) = -1.05$, $p = 0.15$; +2: $t(92) = 2.02$, $p = 0.98$) or second ProbErr (+1: $t(92) = -0.86$, $p = 0.20$; +2: $t(92) = 0.49$, $p = 0.69$). For accuracy after reversals, we replicated the detrimental effect in the conservative condition in trial +1 in which the other player perseverated to the previously correct color, $t(92) = -2.93$, $p = 0.002$, $d = 0.30$. Importantly, though, we also found a facilitatory influence of the (correct) impulsive player in trial +1 after reversals showing improved accuracy relative to the private condition, $t(92) = 3.98$, $p < 0.001$, $d = 0.41$.

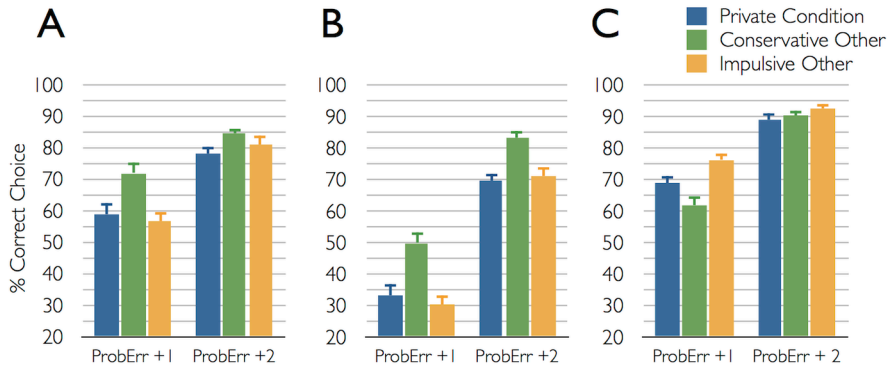

Fig. 7. Choice accuracy (for the truly correct color) as a function of trial position and social influence in Study 2. (A) Accuracy for trials +1 and +2 after the first probabilistic error (ProbErr) in a reversal episode. (B) Accuracy for trials +1 and +2 after the second ProbErr in a reversal episode. (C) Accuracy for trials +1 and +2 after reversals. Error bars show standard errors of the mean.

5.2.2 Computational modeling

Modeling results replicated lower AICs and thus better goodness-of-fit for all four integrative models, relative to the imitative models (all $ts > 9.11$, all $ps < 0.001$, all $ds > 0.95$. Compared to the standard (reward-based) Q-learning model (AIC = 86.28), both instantaneous integration models achieved better fit (same α^s: AIC = 84.56, $t(92) = 3.57$, $p = 0.001$, $d = 0.37$; different α^s: AIC = 85.05, $t(92) = 2.63$, $p = 0.010$, $d = 0.27$) while the AICs of the delayed integration models did not significantly differ from the standard Q-learning model (same α^s: AIC = 85.92, $t(92) = 1.43$, $p = 0.157$; different α^s: AIC = 86.55, $t(92) = -1.01$, $p = 0.319$). In contrast to Study 1, ANOVA of the four integrative models showed a better fit for models using the same α^s for conservative versus impulsive δ^s (main effect *social learning*

rate, $F(1, 92) = 29.76$, $p < 0.001$) and for the instantaneous versus delayed models (main effect *time of integration*: $F(1, 92) = 10.07$, $p = 0.002$). One account for the better fit of models assuming that the same learning rate is instantaneously applied to conservative and impulsive prediction errors – rather than assuming a biased and delayed weighting of conservative choices as in Study 1 – is the relative increase in responses following observed impulsive choices after reversals in Study 3 (see 5.2.1). This likely led to overall higher weights for impulsive δ^s (and thus more balanced weighting of the two types of observed choices) in the iterative free parameter optimization. Such increased learning from impulsive choices was most likely a consequence of introducing a more volatile reward environment with more frequent reversals. This could have amplified participants' exploration propensity (i.e., the spontaneous testing of the alternative, non-rewarded choice option) specifically towards the anticipated end of each reversal episode. Such an interpretation is consistent with the choice accuracy results in Study 3 (see 5.2.1) showing a conservative bias after ProbErrs but an equal influence of conservative and impulsive choices after reversals.

In sum, Study 3 replicated the substantial social influences during observational reversal learning from Study 1 using data from a larger, independent sample. Moreover, Study 3 provides evidence that the bias towards selective learning from observed conservative behavior is mitigated in a less stable decisional context.

6. Simulations

The goal of the simulation analysis was to explore the adaptivity of the choice behaviors implemented in our computational models *within* the present task framework. Specifically, we wanted to compare relative usefulness of the different types of integrative learning (such as differential versus identical weighting of conservative versus impulsive choices), using iterations of learning rate levels and a large number of randomly generated reversal learning sequences.

6.1 Methods

Random learning sequences were constructed within the general design constraints of the present paradigm but included stochastic reward frequencies at different probability levels: We constructed sequences to have 36 reversal episodes with a random length of 7-15 trials each. Within each reversal episode the reward-punishment probability for the correct stimulus was either 0.8/0.2 or 0.6/0.4, and vice versa 0.2/0.8 or 0.4/0.6 for the incorrect stimulus. The other player's response was simulated to be the correct choice (according to the reversal episode) in all trials except in some of the trials following punishments: In half of the reversal episodes the other player switched to the incorrect stimulus after punishments (impulsive behavior), in half of the other episodes, the other player stayed with the previously correct choice (conservative behavior), with the order of conservative/impulsive episodes randomly determined. In addition, after conservative episodes the other played perseverated to the previously correct color for two trials. Average performance of each model was computed for 10,000 permutations (sequences) of the outlined randomization parameters. For each sequence and model we calculated Q-based action selection probabilities P_{blue} and P_{green} according to each model's specific learning algorithms and then computed average % correct choices for each learning rate (α^r or α^s) iterated between 0.05 and 1 (0.05 increments). To reduce the number of free parameters and across models, βs were set to a fixed value of 0.2 corresponding to its average estimated value in data of Study 1. Similarly, for accuracy simulations of the imitative and integrative models α^r was set to a constant value of 0.5.

6.2 Results

For the high reward probability condition (0.8/0.2), the main results were as follows: Standard Q-learning ('social ignorer') reached high accuracy levels of nearly 75% if sufficiently high learning rates were applied (> 0.3). Integrating reward and social information reached comparable accuracy *only* when the same α^s was applied to conservative and impulsive δ^s and when the information was integrated instantaneously rather than with a delay (Fig. 8B-D). Performance based on instantaneous integration and the same α^s peaked at an α^s

of around 0.25, i.e., at substantially higher levels than the modeled α^s for negative δ^s (0.05), suggesting that participants in Study 1 had underweighted impulsive δ^s to optimize performance. Imitative learning did not lead to accuracies levels as high as those of standard-Q learning and instantaneous integration (same α^s). The worst performing model was instantaneous integration with different α^s (Fig. 8C), showing accuracies on average 10% lower than the most successful models.

In the simulations with low reward probability accuracy was substantially decreased across all learning models compared to those with high reward probability. Imitative learning outperformed all other models while performance of standard Q-learning dropped markedly, which can be attributed to the decreased reliability of reward information. Again, within the integrative models we found optimal performance for instantaneous integration using the same α^s. In general, accuracy in the low reward probability simulations increased progressively with increased weighting of social information.

The question arises why these effects – higher adaptivity of ignoring social information when reward information is reliable, higher adaptivity of social learning when reward information is unreliable – are compatible with the notion that the observed agent is exposed to the same reward sequence as the observing agent. However, it should be emphasized that in our simulations the observed agent did not "learn" from the reward outcomes in a technical sense; instead choice behavior (and thus accuracy) was predetermined (see methods 6.1: 100% accuracy in standard trials, fixed number of errors after punishments). Thus, simulated accuracy of social learning was only as good as the fixed validity of social information allowed it to be. General inferences about the adaptivity of social learning thus cannot be drawn. Importantly though, as we mapped the validity of social information in our simulations onto its manipulation in our empirical studies, the simulation results can be used to evaluate whether participants under- or over-utilized reward and social information in the present decision-making framework: Together, our simulation results

suggest that in our paradigm, which was characterized by relatively high reward probabilities, individuals generally over-utilized social information, even though adding social prediction errors to the information that could be gained from reward alone was not necessarily adaptive.

On the other hand, participants under-utilized information from impulsive choices: Simulations demonstrated that social information improved performance only when both conservative and impulsive δ^s were integrated with equal weights, with an optimal α^s of around 0.1-0.3 for both types of prediction errors.

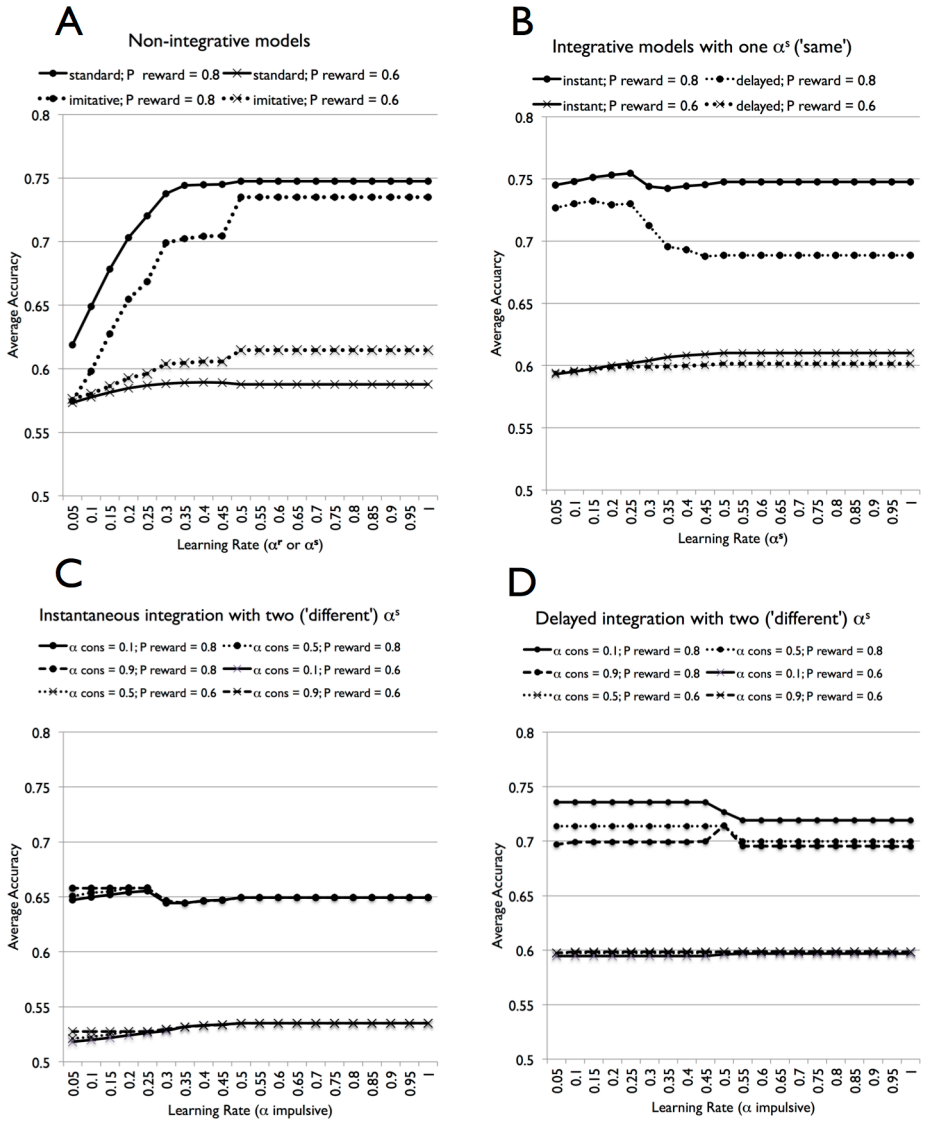

Fig. 8. Results of the simulation analysis. Average accuracy of the different computational models was computed across 10,000 randomly generated reversal learning sequences (each containing 36 reversal episodes with 7-15 trials) that included both social and reward information. Correct responses were

rewarded probabilistically and accuracy was computed for high probability (P reward = 0.8, P punishment = 0.2) and low probability (0.6/0.4) sequences separately. Social information contained in random order conservative (cons) choices in half of the reversal episodes and impulsive choices in the other half. Average accuracy was calculated for different reward (standard Q-learning) or social learning rates (imitative and integrative models) varying between 0.05 and 1. For the integrative models using two α^s, accuracy was computed for low (= 0.1), medium (= 0.5) and high (= 0.9) levels of conservative α^s and an impulsive α^s varying between 0.05 and 1. (A) Results for the non-integrative models (standard Q- and imitative learning), (B)-(D) results for the integrative models.

7. Discussion

The present results demonstrate that observing others critically influences learning of reversing reward contingencies and that individuals integrate an observed agent's choice into learning in form of a 'social prediction error' (δ^s). Specifically, δ^s in combination with the standard reward prediction error (δ^r) jointly explain participants' choices better than models that assume that people rely on either of these sources of information alone (i.e. 'social ignoring' or 'blind following').

Our data demonstrate and replicate the substantial impact of the observed agent's type of choice behavior: Learners follow another player's choice readily if the choice matches their recent choice preference (a 'conservative' choice), even this led to a higher number of perseveration errors. The influence of observed conservative choices was pervasive, affecting decisions after ProbErrs and reversals and conditions with moderate (Study 1) and high (Study 3) reversal volatility. In contrast, observed impulsive choices affected decisions only after reversals and if reversal volatility was sufficiently high (Study 3).

Such selective influence of observed conservative choices – which we demonstrate despite the participants' knowledge that reward contingencies could reverse – has implications for real-life choice situations in which well-established behavioral routines exist that need to be revised or overcome. Our results suggest that in such situations social models in the environment are capable to reinforce the maintenance of established choice preferences, even though these choices are no longer rewarded or the reward is experienced as less pleasant. Such scenarios can be easily mapped, for instance, onto the role of peer models in food choice

behavior during development but also onto the role of social factors in changing unhealthy habits during adulthood.

The notion of a selective influence of (conservative) choices that match established preferences is consistent with a 'confirmation bias', reflecting the tendency to selectively seek or use information that is consistent with one's preconceptions (Nickerson, 1998). The confirmation bias has been recognized as a ubiquitous phenomenon manifesting itself in a variety of domains, ranging from attention to memory and formal reasoning (Nickerson, 1998). In the context of reward-based decision-making, a confirmation bias towards verbal information given *before* learning, was shown to modulate the weighting of outcomes received in the feedback phase (Doll, Hutchison, & Frank, 2011). Moreover, there is an open debate on the extent to which cognitive biases, such as the confirmation bias, reflect mental 'flaws' or have adaptive utility and even lead to more accurate responses than unbiased responses (Gigerenzer, 1991). Our tasks were not designed to allow a quantification of the utility of following social information in general as we deliberately presented sub-optimal choice behavior. Nonetheless, our simulation results suggest that integrating social information in the *present* experimental framework did not necessarily help participants to maximize their earnings, especially not when participants weighted observed conservative choices more than impulsive choices. It is possible, though, that following conservative choices evoked a subjective reduction of uncertainty (not measured in the present studies), which has been shown to mediate social influence in situations of limited stimulus information (McGarty, Turner, Oakes, & Haslam, 1993). The idea of selective weighting of social information to achieve 're-assurance' is also consistent with recent work on the evolutionary bases of conformity (i.e. adoption of behaviors displayed by groups) and social learning, highlighting that participants' confidence and uncertainty critically affect the strength of social influences (Morgan et al., 2012).

Another recent study pertinent to our findings showed that social influence exerted by another observed learner during probabilistic decision-making increased with the observable skill level of the demonstrator (Selbing et al., 2014). Such selective influence of competent others fits well with animal research showing that successful models are more likely to be copied (Laland, 2004). In our study, we balanced performance (inferred skill) levels between the conservative and impulsive conditions, which enabled us to extend these findings by showing that even subjective criteria not linked to competencies, in our case the incidental similarity conferred by a shared birthday, mediate social learning. These effects are consistent with other work showing that incidental similarity increases compliance with requests (Burger et al., 2004) or behavioral mimicry (Guéguen & Martin, 2009). Incidental similarity is assumed to elicit a coarse form of information processing guiding decisions based on heuristics rather than careful considerations of the choice options (Burger et al., 2004). Specifically, salient features indicating similarity will shift attention to other features perceived as similar rather than attributes indicating dissimilarity (Mussweiler, 2003). In accordance with this theory, incidental similarity in our study selectively increased α^s for conservative δ^s, that is, similar players induced an even stronger conservative bias than dissimilar players. The modulation by interpersonal similarity also supports the notion that the current pattern of results can be attributed to social-cognitive processes rather than arising from simple associative learning. Simple associative learning should have resulted in a similar conservative bias in both similarity groups which we did not find.

Consistent with our results, a social modulation of reward-based decision-making was recently shown in two-armed bandit tasks with fixed (non-reversed) reward probabilities (Burke et al., 2010), instrumental conditioning with liquid rewards (Cooper et al., 2012) and advice-based tasks (Behrens et al., 2008; Biele et al., 2011). Furthermore, brain imaging findings suggest that expectancy violations in the social domain engage similar brain regions as prediction errors during reinforcement learning (Harris & Fiske, 2010). Other brain

imaging work has supported the existence of an 'action prediction error' (Burke et al., 2010), reflecting the difference between predicted and observed choices of another agent and underpinned by regionally specific activation changes in the dorsolateral prefrontal cortex. Critically, in contrast to our paradigm, these studies did not compare different attributes and choice behaviors displayed by the observed agent that modulate these computational signals. Accordingly, 'action prediction errors' were modeled as non-signed variables (Burke et al., 2010). Another type of social prediction error has been described for the observation of (independently varied) outcomes received by the other agent and associated with neural responses in ventromedial prefrontal cortex (Burke et al., 2010) and dorsal striatum (Cooper et al., 2012). In our study, we did not vary outcomes for real and observed players independently, making the 'observational outcome prediction error' (Burke et al., 2010) temporally overlap with the (experiential) reward prediction error. Such joint outcomes are assumed to signal prediction errors related to the 'trustworthiness' of the other player and recruit brain regions involved in social evaluation (Behrens et al., 2008). As outlined above, these social-evaluative factors related to the other's competence or trustworthiness may have been captured by the delay component in our models.

We used comparatively simple Q-learning algorithms to model our data. Q-learning has successfully been applied to predict choices during reversal learning by other groups (Jocham et al., 2009). While more sophisticated approaches like Hidden Markov models may alter overall goodness-of-fit results (Hampton, Bossaerts, & O'Doherty, 2006), we do not expect that they would change the *pattern* of results as observed here (e.g. relative explanatory advantage for models with integrative strategies). A promising perspective for future studies would also be to use dynamic rather than fixed weighting factors to model socially mediated decision-making. For instance in Pearce-Hall models (Pearce & Hall, 1980), instead of assigning a constant learning rate that scales prediction errors throughout the learning process, the amount of learning is dependent on attentional deployment

("associability"), which in turn varies depending on prediction errors experienced in preceding trials.

8. Conclusions and Implications

Our findings highlight the pivotal role of social influence in human cognition, demonstrating that even basic mechanisms such as reversal learning are affected by observation of other people's behavior. We show that in scenarios implementing decisional uncertainty by reversing reward contingencies individuals combine social and non-social information to guide their decision. However, individuals learn more strongly from information that is provided by observed agents who are perceived as similar, even if this perception is based on minimalistic and task-irrelevant information (shared birthday). Interpersonal similarity is a potent trigger and modulator of various social behaviors ranging from altruistic punishment (Mussweiler & Ockenfels, 2013) to evaluations of others (Mussweiler & Damisch, 2008). Our results suggest that that similarity modulates even more basic cognitive processes, such as reversal learning.

Learners generally exhibited a 'conservative bias' towards social cues that conform to their preconceptions. We demonstrate how these biases can be incorporated into formalized learning models. Knowledge about such social trajectories in decision-making is important for psychological research into the determinants of human choices in natural contexts, especially if choices are influenced by others, such as in voting or consumer behavior. Moreover, the right balance of reward-based versus social-observational learning is often crucial for the success or failure of education programs or interventions for behavioral change. The formal models proposed here can inform and optimize the design of such programs and interventions, specifically if the approaches include 'role model' behavior to promote learning.

Acknowledgements

Part of this research was funded by the National Center for Mental Health (Cardiff, Wales) hosted by the National Institute for Social Care and Health Research (NISCHR), Wales, and a grant from the bilateral programme between the Economic and Social Research Council of the UK (ESRC) and the German Research Foundation (DFG) (RES-062-23-0946: The Neural Substrates of Social Comparison).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Asch, S. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), Whole No. 416.

Baron, R. S., Vandello, J. a., & Brunsman, B. (1996). The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology*, *71*(5), 915–927. doi:10.1037/0022-3514.71.5.915

Behrens, T., Hunt, L., Woolrich, M., & Rushworth, M. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. doi:10.1038/nature07538.Associative

Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The neural basis of following advice. *PLoS Biology*, *9*(6), e1001089. doi:10.1371/journal.pbio.1001089

Brown, J., & Novick, N. (1992). When Gulliver travels: Social context, psychological closeness, and self-appraisals. *Journal of Personality and Social Psychology*, *62*(5), 717–727.

Burger, J. M., Messian, N., Patel, S., del Prado, A., & Anderson, C. (2004). What a coincidence! The effects of incidental similarity on compliance. *Personality & Social Psychology Bulletin*, *30*(1), 35–43. doi:10.1177/0146167203258838

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14431–6. doi:10.1073/pnas.1003111107

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 933–42. doi:10.1098/rstb.2007.2098

Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(11), 4563–7. doi:20026435

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience*, *24*(1), 106–18. doi:10.1162/jocn_a_00114

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgement. *Journal of Abnormal Psychology*, *51*(3), 629–636. doi:10.1037/h0046408

Dias, R., Robbins, T., & Roberts, A. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, *380*, 69–72.

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(16), 6188–98. doi:10.1523/JNEUROSCI.6486-10.2011

Festinger, L. (1950). Informal social communication. *Psychological Review*, *57*(5), 271–282.

Fineberg, N., Potenza, M. N., Chamberlain, S. R., Berlin, H., Menzies, L., Bechara, A., … Hollander, E. (2010). Probing compulsive and impulsive behaviors, from animal models to endophenotypes: a narrative review. *Neuropsychopharmacology*, *35*(3), 591–604. doi:10.1038/npp.2009.185

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." *European Review of Social Psychology*. doi:10.1080/14792779143000033

Guéguen, N., & Martin, A. (2009). Incidental Similarity Facilitates Behavioral Mimicry. *Social Psychology*, *40*(2), 88–92. doi:10.1027/1864-9335.40.2.88

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(32), 8360–7. doi:10.1523/JNEUROSCI.1010-06.2006

Harris, L. T., & Fiske, S. T. (2010). Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Social Neuroscience*, *5*(1), 76–91. doi:10.1080/17470910903135825

Jocham, G., Neumann, J., Klein, T., Danielmeier, C., & Ullsperger, M. (2009). Adaptive coding of action values in the human rostral cingulate zone. *The Journal of Neuroscience*, *29*(23), 7489–96. doi:10.1523/JNEUROSCI.0349-09.2009

Jones, B., & Mishkin, M. (1972). Limbic Lesions and the Problem of Stimulus-Reinforcement Associations. *Experimental Neurology*, *377*, 362–377.

Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, (2), 136–153.

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior*, *32*(1), 4–14.

Lindström, B., & Olsson, A. (2015). Mechanisms of Social Avoidance Learning Can Explain

the Emergence of Adaptive and Arbitrary Behavioral Traditions in Humans. *Journal of Experimental Psychology: General*, *144*(3), 688–703. doi:10.1037/xge0000071

McGarty, C., Turner, J., Oakes, P., & Haslam, S. (1993). The creation of uncertainty in the influence process: The roles of stimulus information and disagreement with similar others. *European Journal of Social Psychology*, *23*, 17–38.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., … Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science*, *324*(5929), 900. doi:10.1126/science.1170539

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings. Biological Sciences / The Royal Society*, *279*(1729), 653–62. doi:10.1098/rspb.2011.1172

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472–489. doi:10.1037/0033-295X.110.3.472

Mussweiler, T., & Damisch, L. (2008). Going back to Donald: how comparisons shape judgmental priming effects. *Journal of Personality and Social Psychology*, *95*(6), 1295–315. doi:10.1037/a0013261

Mussweiler, T., & Ockenfels, A. (2013). Similarity increases altruistic punishment in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19318–23. doi:10.1073/pnas.1215443110

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. doi:10.1037//1089-2680.2.2.175

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. doi:10.1037/0033-295X.87.6.532

Selbing, I., Lindström, B., & Olsson, A. (2014). Demonstrator skill modulates observational aversive learning. *Cognition*, *133*(1), 128–139. doi:10.1016/j.cognition.2014.06.010

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press.

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.