Annals of the New York Academy of Sciences

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

ORIGINAL ARTICLE

Intention–Outcome Trade-Off in Moral Character Learning

Gaojie Huang[1] | Yongbo Xu[1] | Edmund Derrington[2,3] | Jean-Claude Dreher[2,3] | Chen Qu[1]

[1]Center for Studies of Psychological Application, South China Normal University, Guangzhou, China | [2]Laboratory of Neuroeconomics, Institut des Sciences Cognitives Marc Jeannerod, CNRS, Lyon, France | [3]Université Claude Bernard Lyon 1, Lyon, France

Correspondence: Jean-Claude Dreher (dreher@isc.cnrs.fr) | Chen Qu (fondest@163.com)

Received: 3 August 2025 | **Revised:** 16 December 2025 | **Accepted:** 2 February 2026

Keywords: intention learning | moral character learning | outcome learning | reinforcement learning | trade-off parameter

ABSTRACT

To evaluate an individuals' moral character the intentions behind their actions must be discriminated from the actual outcome. How this is achieved remains unclear. We developed a novel paradigm that dissociates the perception of intentions from outcome. Participants separately predicted and received feedback on agents' intentions and outcomes. They then made moral evaluations of the agent. Four independent experiments (1, 2, 3a, and 3b; 120 participants in total), demonstrated that intentions and outcomes mutually influenced each other during the learning process before integrating them for subsequent moral evaluation. Computational modeling further revealed that intentions biased the predictions of outcomes, while outcomes directly modified the beliefs about intentions. Moreover, participants considered both intention and outcome when making moral evaluations, but placed greater weight on intention, regardless of sampling bias and presentation order. These findings offer new insights concerning how individuals process intention and outcome when learning about others' moral character.

1 | Introduction

Our perceptions of others' morality evolve, and are influenced by new information and feedback concerning people's actions and outcomes. This dynamic process is referred to as moral character learning [1, 2]. When learning others' moral character, intentions and outcomes are two primary dimensions [3, 4]. We believe individuals with good/bad intentions are those who frequently yield positive/negative outcomes. Similarly, we tend to see those who produce good/bad outcomes as individuals with good/bad intentions. However, it is unclear how do individuals process conflicting information when learning another's moral character. Elucidating this process is crucial to understand how individuals identify reliable partners and build effective social connections [5].

Empirical evidence has consistently demonstrated that individuals integrate intentions and outcomes to make moral evaluations [6, 7]. Most studies have examined the influence of intentions and outcomes, and their effects across various moral dimensions or contexts [8–11]. For example, adults are more sensitive to intentions than children [6, 12, 13]. Intentions are often implicit and require inferences whereas the resulting outcome is explicit [14–16]. Furthermore, previous studies have presumed that participants understood information regarding others' intentions exactly as the experimenters intended [9, 11, 17]. This assumption may overlook the potential impact of individual differences. Evidence from studies that examined reasoning style, for instance, suggests that individual differences can significantly moderate the extent to which intentions and outcomes effect moral evaluations [18]. This underscores the need to account for individual variability, and quantify the weight of intentions and outcomes separately in moral evaluations for each participant. Moreover, when learning others' moral character, inferences concerning their intentions can be updated with outcome feedback

Gaojie Huang, Yongbo Xu, Jean-Claude Dreher, and Chen Qu contributed equally to this work.

[19, 20]. Recent study explicitly supports this interaction, demonstrating that learning from outcomes can shape the reliance on moral rules versus cost–benefit reasoning, thereby modulating intention-based judgments [21]. Therefore, intention leaning may be influenced by outcome learning. Whether outcome learning could be influenced by intention learning remains unclear. Thus, presenting intentions and outcomes separately should enable us to examine the interplay between intention leaning and outcome learning in learning others moral character.

Here, we designed and optimized a novel task that dissociated the learning of intentions and outcomes and quantified them separately to determine how they interact. Both intention and outcomes were presented apparently to exclude the possible individual difference in inference. We applied the insight that moral learning, like all forms of social learning, is guided by processes that involve prediction and updating [2, 13, 22]. Examination of the moral learning processes will allow us to elucidate the interplay between intention learning and outcome learning. Do intention and outcome merely bias the prediction process, or do they also influence the updating process by which outcome is used to update the individual's beliefs about the agent's intentions? Previous studies, using reinforcement learning models, found that people preferentially form abstract trait inferences about others through feedback, in addition to making reward associations [20]. We hypothesized that the outcome would influence intention learning by biasing current beliefs regarding the agents' intentions.

In our task, after being presented with an agent's intention and outcome separately, participants make a moral evaluation, that can integrate both intention and outcome. Most previous research employed situational narratives to explore how intention and outcome affect moral evaluation [23–26]. Research shows that when intention and outcome are mismatched, people tend to prioritize intention over outcome in their moral evaluations and punishment decisions [11, 25]. However, little research has attempted to quantify their relative contributions. We introduced an intention–outcome parameter, which calculated relative weights of intention and outcome when participants made moral evaluations. As confounding factors, sampling errors and the order of presentation intention and outcome might influence this parameter [27, 28]. Therefore, we performed additional experiments to exclude them and gain a clearer understanding of the contribution of intention and outcome in moral character learning.

In summary, Experiment 1 built on the principle of harm aversion [29], to elucidate the interplay between intentions and outcomes and calculate their relative weights in moral character learning based on reinforcement learning. Briefly, in each trial, participants were first required to predict the intentions of one of four external agents, two of whom were ill-intentioned and two well-intentioned. The agents then chose to expose the participants to either a high or low risk of receiving an auditory shock. After learning the intention of the agent, participants were then required to predict the actual outcome (shock or not), the probability of receiving a shock being more or less likely, depending on the agent's choice and the roulette wheel set used by the agent. Finally, the participants learned the outcome of the shock lottery, shocks being administered at the end of the experiment, and made moral evaluations of the agents. Experiments 2 and 3 allowed us to control for possible confounds from Experiment 1. These included the relative salience of the different outcomes (shock or no shock) for each agent, so as to have two distinct levels of frequency of a poor outcome, independent of the predicted outcome, potential unbalanced learning and sampling errors, and finally, the order of presentation of the outcome and the intention of the agent.

2 | Experiment 1

2.1 | Methods

2.1.1 | Participants

For each experiment (Experiments 1, 2, 3a, and 3b), we recruited 30 participants of age ranging from 18 to 23 years, 16 of which were female: (Experiment 1: mean age 20.33 ± 1.94 years; Experiment 2: mean age 19.40 ± 1.83 years; Experiment 3a: mean age 19.27 ± 1.17 years; Experiment 3b: mean age 19.8 ± 1.84 years). All participants were students recruited from the university community. Exclusion criteria included being under 18 years of age, having a history of psychiatric or neurological disorders, having participated in an earlier experiment in this series, and having majored in psychology. All participants were remunerated (20, 20, 45, and 20 Yuan for Experiments 1–3b, respectively). The study was approved by the Ethics Committee of our university and was carried out in accordance with the Declaration of Helsinki.

2.1.2 | Experimental Design and Procedure

Upon arriving at the lab, participants were informed they would take on roles as decision-maker (decider) and receiver sequentially in an interpersonal game. The decider session was conducted only as part of the cover story (see Supporting Information). In the receiver session, participants engaged in a moral character learning task, in which they learned to distinguish the intentions and outcomes of four agents (identified as previous participants), in terms of their intentions and the outcomes of these agents' decisions.

Agents presented pairs of blue or orange roulette wheels (two agents for each), with one color pair consistently leading to higher auditory shock (loud abrasive shock) probabilities (negative outcome) and the other to lower probabilities (positive outcome), though participants remained unaware of which color pair carried which risk. Each pair of roulette wheels included a dark- and a light-colored wheel. The likelihood of the participants receiving an auditory shock was higher for the dark-colored roulette wheel. The participants knew this because they had played the role of agent in the previous session. Thus, the participants knew that agents with more positive intentions should choose the light-colored roulette wheel, although there remained a possibility that light-colored wheels could still indicate the negative outcome (shock). Similarly, agents with more negative intentions would choose the dark roulette wheels more frequently, but these might still indicate a positive outcome (no shock). Thus, an agent's preference to choose dark roulette wheels suggested their negative intentions, but the actual outcome, shock or not, remained a

lottery. Participants were repeatedly told there was a difference in the probability of receiving the shock across agents depending on which color pair of roulette wheel used by the agent, and they needed to learn each agent's preference and whether this made them more or less likely to receive a shock. Prior to the main task, shock aversion was calibrated using an induction procedure. Participants were exposed to a 5-s sample of the auditory shock, composed of high-frequency biological buzzing (e.g., mosquitoes), human distress vocalizations, and threatening mechanical noises. Participants rated the unpleasantness on a scale of 1 (not at all) to 9 (extremely unpleasant). All participants reported a rating of 7 or higher, confirming that the stimulus was perceived as effectively aversive. Experiment 1 employed a 2 (intention: positive/negative) × 2 (outcome: positive/negative) within-subject design. Figure 1B details the frequencies of roulette wheel choices and shock outcomes for each agent. The roulette wheel colors assigned to the different agents (blue and orange) were counterbalanced across participants.

The learning task consisted of a Learning and a Testing phase. During the Learning phase, participants predicted and then received feedback regarding each agent's choices and the probable outcome, and then made moral evaluations of the agents according to the information they had learned. The procedure is illustrated in Figure 1A.

For each trial of the Learning phase, an agent's face was shown for 1 s. Below the face, a dark- and a light-colored roulette wheels were displayed, with positions counterbalanced over trials. Participants had 2 s to predict which wheel the agent would choose (intention prediction). The agent's actual chosen wheel was then displayed in the center with the nonchosen wheel marked with an "×" for 2 s (intention feedback). A yellow (auditory shock) and a white lightning bolt (no shock) were then shown below the face, with positions counterbalanced across trials. Participants had 2 s to predict whether they would receive a shock (outcome prediction) by choosing the yellow lightning bolt or not (choosing the white lightning bolt), and then the actual result was displayed for 2 s (outcome feedback). Finally, participants had 3 s to rate the agent's moral character (moral evaluation), from 1 (very bad) to 9 (very good). The Learning phase comprised 200 trials, divided into five blocks of 40 trials each, with each agent appearing 10 times randomly per block.

At the end of each block, participants completed a Testing phase, comprised 60 trials divided into five blocks of 12, in which the faces of all pairs of agents were presented twice. Participants were asked to choose (within 2 s) which agent they preferred. To encourage participants to learn the characteristics of the agents they were informed that their prediction accuracy during the Learning and Testing phase would affect whether they would be subjected to auditory shocks at the end of the experiment. However, in fact no shocks were ever administered to any of the participants. A debriefing session was conducted after the experiment to assess participants' comprehension of the experiment. All participants demonstrated a clear understanding of the procedures and reported no difficulties in following the instructions.

2.1.3 | Analyses

Participants' predictions regarding the wheel an agent would choose were identified as indicators of intention learning. Their prediction of having a shock outcome was identified as an indicator of outcome learning. We conducted a 2 × 2 repeated-measures ANOVA to examine the effects of intention and outcome on intention learning, outcome learning, moral evaluation of the agents, and the participants' choices in the Testing phase.

To further investigate the mechanism of influence and the balance between intention and outcome, we fitted eight reinforcement learning models (see Supporting Information), to account for the interplay between intention and outcome learning in different ways. In these models, we assumed that participants learned both agents' intentions and outcomes based on the Rescorla–Wagner prediction error rule, forming the estimates of expected intention ($IE_{i,t}$, i.e., belief about intention, reflecting participants' belief about agent i choosing dark-colored roulette wheel) and expected outcome ($OE^{C}_{j,t}$, i.e., belief about outcome, reflecting participants' belief about auditory shock induced by the used roulette wheel set j, where $C \in \{DW, LW\}$ represents the dark- or light-colored wheel, respectively). Within the reinforcement learning framework, the learning process of intention and outcome can be seen as a continuous cycle of prediction process and updating process [30]. We proposed two potential pathways of influence between intention and outcome within this framework. Specifically, intention might affect the prediction on outcome either by biasing participant's prediction probability regarding the likely final outcome in current trial (prediction bias), calculated using a softmax function, or by prospectively updating participants' beliefs regarding the agent's outcomes when the roulette wheel the agent had selected was identified (belief updating). Similarly, we assumed that outcome could affect the prediction on intention by biasing the participants' prediction probability with respect to the agent's intention in next trial (prediction bias), or by retroactively adjusting the participants' beliefs regarding the agent's intention, when outcome feedback (shock or no shock) was presented (belief updating). We also took into account whether there was an effect of first impression effect by computing the tendency to predict the occurrence (nonoccurrence) a shock outcome when a dark(light)-colored roulette wheel was chosen by the agent to separate the impact of intention information itself from the bias introduced by the presentation order. This was because intention information (roulette wheel color dark/light) was systematically presented before the outcome information in Experiment 1. The eight models were constructed by combining these three factors: the pathway through which intention affects outcome learning (prediction bias vs. belief updating), the pathway through which outcome influences intention learning (prediction bias vs. belief updating), and whether the first impression effect is considered (Yes vs. No; Figure 2 and Table S1). For these eight alternative models, Akaike information criterion weights (AIC_w), which expresses the relative likelihood of one model over others, was used to assess each model's goodness of fit (Wagenmakers and Farrell, 2004). Thus, a higher AIC_w indicates better model fit. After identifying the best-fitting model of the learning process, the

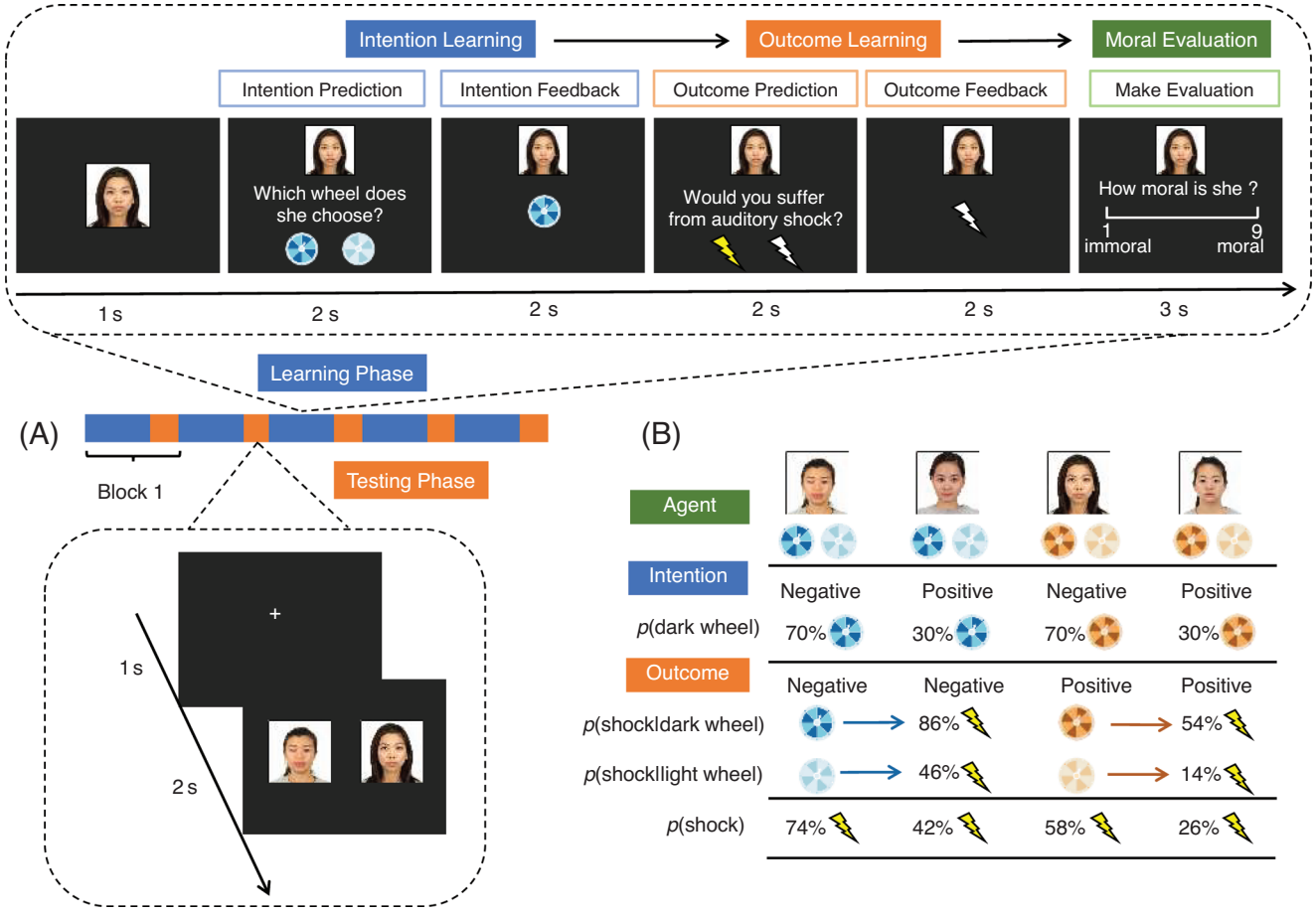


FIGURE 1 | Experimental procedure and setting in moral character learning task. (A) All participants first played the game in the role of the "agent" to ensure that they believed the experimental setting and understood the game. Then participants completed the receiver session. In each trail, a participant was presented with an image of one of four agents and was then asked to predict whether the agent would choose either a dark- or light-colored roulette wheel. The former was associated with a high risk of the participant receiving an auditory shock, the latter with a low risk. They were presented the agent's choice as feedback after 2 s. Then participants were then asked to predict the outcome of the roulette wheel lottery on whether they would receive the shock at the end of the experiment or not. After 2 s they received the outcome of this lottery (shock or not) which was displayed for 2 s. Finally, participants made a moral evaluation of the agent on a 9-point scale (1–9) using keyboard keys. At no point in the experiment did the participant actually receive the auditory shocks. At the end of each block, participants completed a Testing phase in which they were asked to choose which of two deciders they preferred. A highly similar experimental procedure was used in Experiments 2 and 3 with any differences (e.g., whether the participant was required to predict the outcome, or not (Experiment 2) and reversal of the order of presentation of the outcome and intention being reversed (Experiment 3) being specified and explained in the appropriate experimental sections. The agents' faces were selected from the Chicago Face Database [39]. (B) Participants encountered four agents differing in their probability of selecting the dark-colored roulette wheel (agents' intention: 70% for negative vs. 30% for positive) and the shock probabilities associated with their roulette wheel sets (agents' outcome: blue wheels = high shock risk [dark: 86%, light: 46%]; orange wheels = low risk [dark: 54%, light: 14%]). By varying the shock probabilities associated with the roulette wheel sets, the probability of shock outcome was manipulated independently of the probability of selecting the dark-colored roulette wheel. As shown in Panel B, each agent's combined intention–outcome profile in Experiment 1 yielded distinct overall shock probabilities (74%, 42%, 58%, and 26%), which spanned a continuous range rather than forming distinct positive/negative categories. Experiments 2 and 3 systematically adjusted these probabilities to create two discrete valence levels (positive vs. negative) for agents' overall shock outcome (see Tables 1 and 2 for specific configurations).

intention–outcome parameter ω was estimated by maximizing Pearson correlation, and then R^2 was calculated by fitting an optimal linear transformation between model predictions and participant ratings. The intention–outcome parameter reflected the balance between the participants' belief of the agent's intention ($\omega = 1$, when only the intention belief is considered) and outcome ($\omega = 0$, when only the outcome belief is considered) when making a moral evaluation of that agent. Moreover, a permutation test with 1000 random permutations was performed to validate the significant results for intention–outcome parameter.

Finally, we conducted simulation analyses using its best-fitting parameters to validate the absolute performance of the winning model. For each participant, these parameters were treated as a virtual agent to simulate intention and outcome predictions for every trial, repeating this process 100 times with the actual experimental stimulus sequences. This generated 100 datasets of 30 pseudo-subjects. For each dataset, we refit the 2×2 repeated-measures ANOVA analysis as applied to the empirical data to identify whether the winning model successfully reproduced our model-free effects. Finally, we used these simulated datasets to

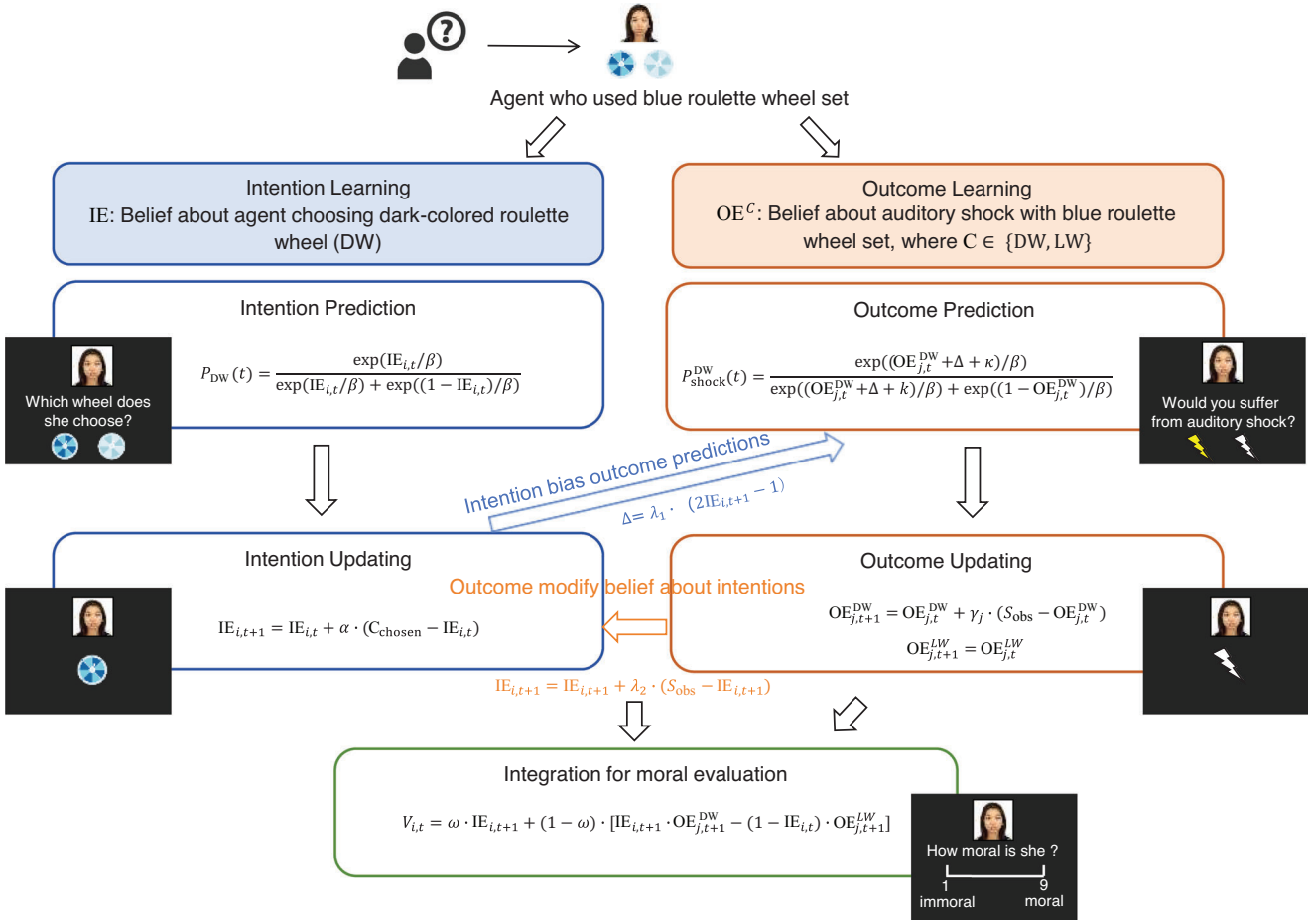


FIGURE 2 | Illustration of best-fitting model (Model 6). We assumed that participants both learned agents' intentions and outcomes, forming the estimates of expected intention ($IE_{i,t}$, i.e., belief about intention, reflecting participants' belief about agent i choosing dark-colored roulette wheel) and expected outcome ($OE^{C}_{j,t}$, i.e., belief about outcome, reflecting participants' belief about auditory shock induced by the used roulette wheel set j, where C ∈{DW, LW} represents the dark- or light-colored wheel, respectively). In this example, participants updated their beliefs based on the Rescorla–Wagner prediction error rule: intention feedback updated the belief about intention $IE_{i,t}$, and outcome feedback updated the belief about outcome $OE^{C}_{j,t}$, with $j = 1$ corresponding to the blue roulette wheel set. Only belief about outcome corresponding to the chosen roulette wheel $OE^{DW}_{j,t}$ is updated, while belief about outcome for the unchosen wheel $OE^{LW}_{j,t}$ remains unchanged. The probability of predicting agent i would choose dark-colored roulette wheel P_{DW} or the presence of an auditory shock P^{DW}_{shock} is calculated using a softmax function. In Model 6, we assumed that intention biased outcome prediction, meaning that the probability of predicting the presence of an auditory shock is influenced not only by belief about outcome, but also weighted by belief about intention. Moreover, the first impression parameter κ was estimated to separate the effect of intention on outcome prediction itself from the bias introduced by the presentation order. The influence of outcome on intention learning occurs through belief updating, where outcome feedback directly updated and modified the belief about intentions. Finally, the updated belief about intention and outcome was integrated into an overall expected morality value for moral evaluation. The intention–outcome parameter ω characterizes the relative weight of belief about intention, with a higher ω indicating a greater emphasis on intention in moral evaluation.

perform parameter recovery to validate that the winning model parameters were reliable. We estimated all the parameters for each simulation dataset, and then examined the correlation between the average recovered parameters and each participant's true parameters (for further details on the computational modeling, see Supporting Information).

2.2 | Results

2.2.1 | Behavioral Results

Participants learned the characteristics of agents' intentions and outcomes (intention prediction accuracy: mean ± SD = 0.59 ± 0.06, $t(29) = 8.39$, $p < 0.001$, Cohen's $d = 1.53$; outcome prediction accuracy: mean ± SD = 0.62 ± 0.05, $t(29) = 13.86$, $p < 0.001$, Cohen's $d = 2.53$). A two-way repeated measures ANOVA on predictions of agents' intentions and outcomes (intention learning and outcome learning) revealed that agents with negative intentions or negative outcomes were perceived to have worse intentions than was actually the case (intention: $F(1,29) = 92.88$, $p < 0.001$, $\eta_p^2 = 0.762$; outcome: $F(1,29) = 7.34$, $p = 0.011$, $\eta_p^2 = 0.20$), with no significant interaction ($F(1,29) = 0.24$, $p = 0.629$, $\eta_p^2 = 0.01$). This indicates that beliefs about agents' intentions were influenced by both intention and outcome (Figure 3A). Similarly outcome predictions were also influenced by both intention ($F(1,29) = 194.24$, $p < 0.001$, $\eta_p^2 = 0.87$) and outcome ($F(1,29) = 13.59$, $p < 0.001$, $\eta_p^2 = 0.32$), with no significant interaction ($F(1,29)$

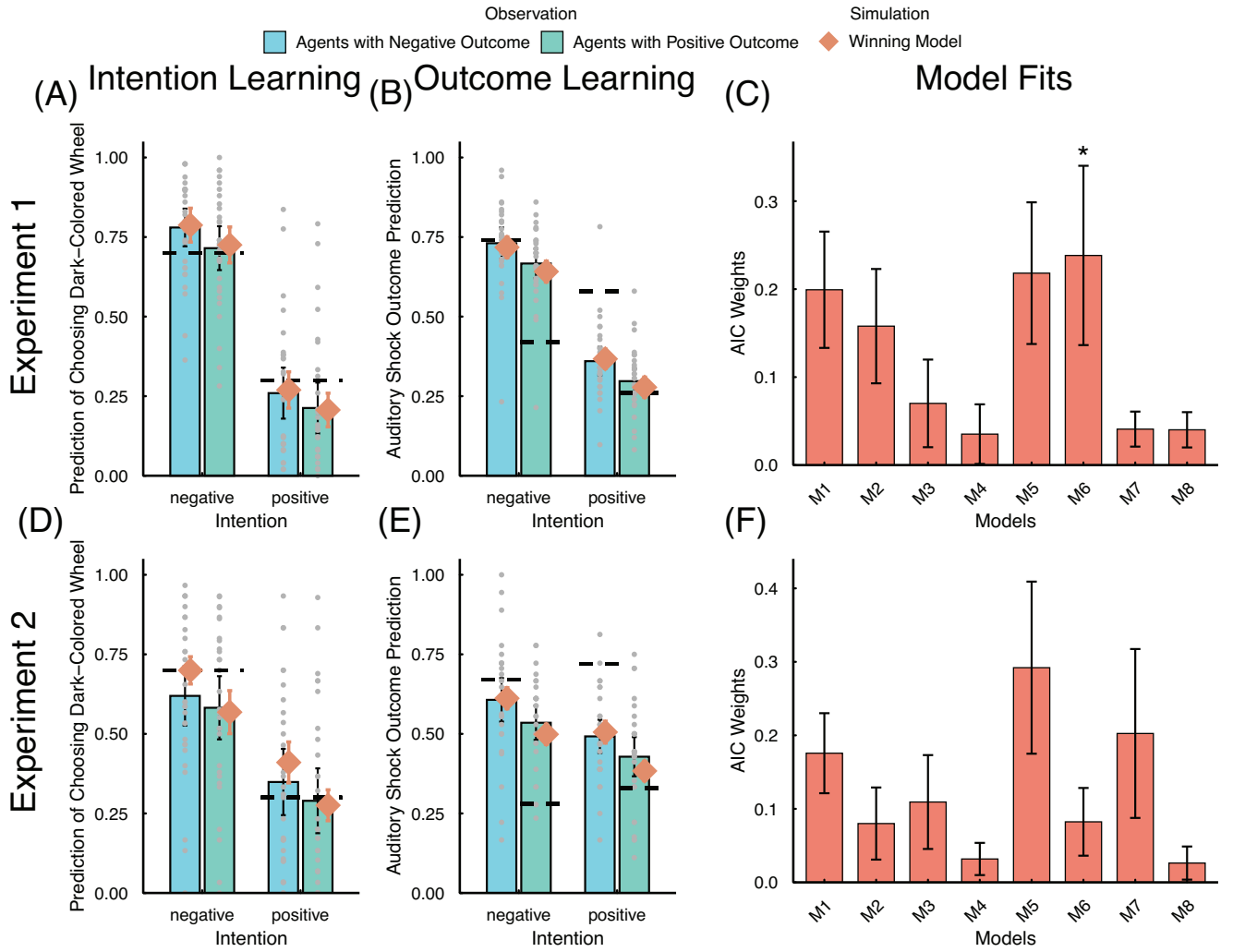


FIGURE 3 | Intention learning and outcome learning in Experiment 1 and Experiment 2. (A, D) Participants' predictions of agents' dark-colored roulette wheel choices quantified intention learning across agents systematically varying in intention (probability of choosing dark wheels) and outcome (auditory shock probabilities determined by orange/blue wheel sets). Dashed lines indicate agents' preset intention probabilities. (B, E) Shock outcome predictions measured outcome learning across agents differing in intention and outcome. Dashed lines denote ground-truth shock probabilities for each agent. (C, F) Model comparison of eight reinforcement learning models testing how intention and outcome learning interact. The models vary based on: (1) the pathway through which intention affects outcome learning (prediction bias [M1–M2, M5–M6] vs. belief updating [M3–M4, M7–M8]), (2) the pathway through which outcome influences intention learning (prediction bias [M1, M3, M5, M7] vs. belief updating [M2, M4, M6, M8]), and (3) whether the first impression effect is incorporated as a covariate (Yes [M5–M8] vs. No [M1–M4]). AIC weights (bar heights) reflect conditional probabilities of model, with asterisks marking the best-fitting model (highest weight). Orange diamonds indicate predictions from winning model. Error bars represent 95% confidence intervals.

< 0.01, $p = 0.960$, $\eta_p^2 < 0.001$; Figure 3B). These findings highlight the mutual impact of intention and outcome on both learning process.

Participants' moral evaluations and partner choices in the Testing phase were analyzed as for intention learning. Participants rated agents with positive intentions or positive outcomes higher in morality (intention: $F(1,29) = 50.46$, $p < 0.001$, $\eta_p^2 = 0.64$; outcome: $F(1,29) = 34.73$, $p < 0.001$, $\eta_p^2 = 0.55$) and preferred them as partners (intention: $F(1,29) = 29.83$, $p < 0.001$, $\eta_p^2 = 0.51$; outcome: $F(1,29) = 6.15$, $p = 0.019$, $\eta_p^2 = 0.18$; Figure 4A,B). No significant interaction effect was found between intention and outcome (moral evaluations: $F(1,29) = 0.26$, $p = 0.618$, $\eta_p^2 = 0.01$; partner choices: $F(1,29) = 2.22$, $p = 0.147$, $\eta_p^2 = 0.07$). These results suggest that both intentions and outcomes influence participants' moral evaluations and partner choices.

2.2.2 | Computational Model Results

To further identify the mechanism of influence between intention and outcome, we fitted eight different computational models to our data. The AIC model selection using AIC_w indicated that Model 6, which proposed intentions biased participants' predictions of agent outcomes while outcome feedback additionally updated their beliefs about agent intentions and the intention–outcome presentation order further modulated outcome prediction, exhibited the highest conditional probability

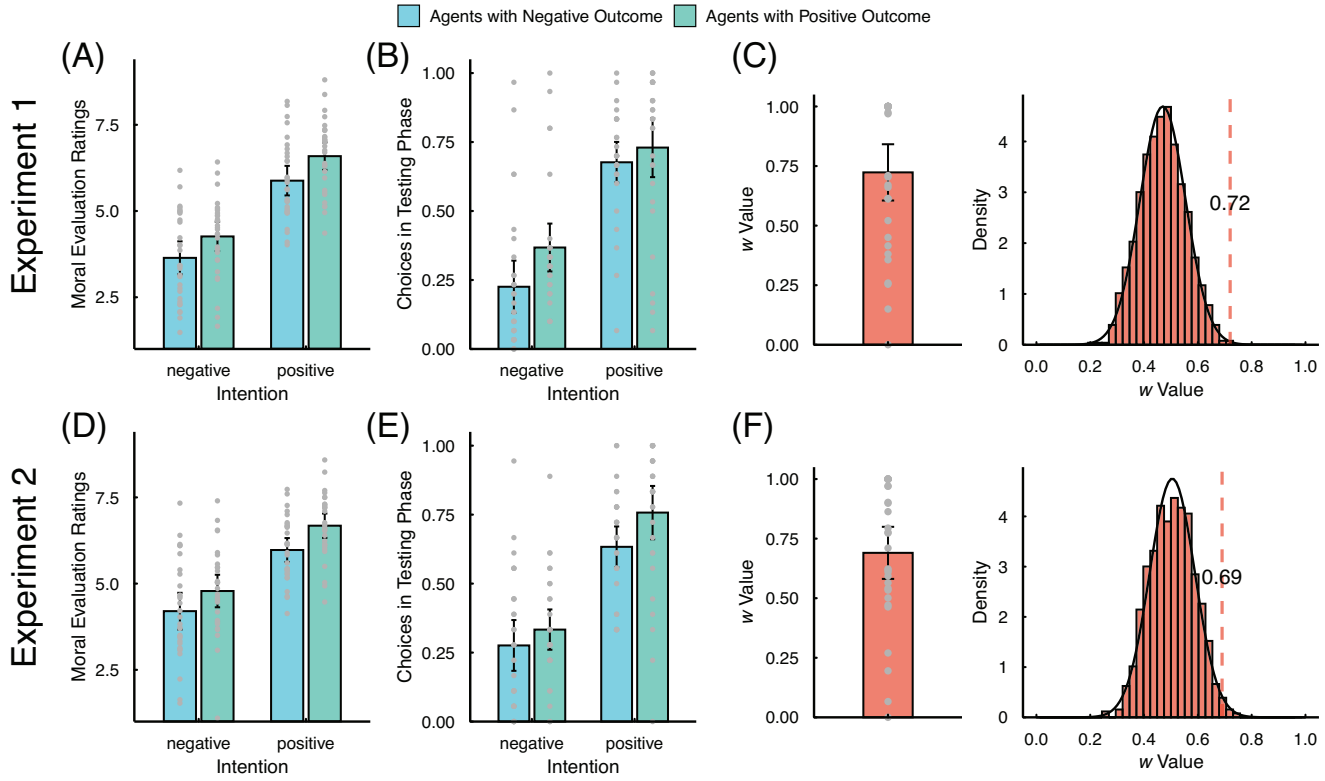


FIGURE 4 | Moral evaluations and choices in the Testing phase in Experiments 1 and 2. (A, D) The degree which participants made moral evaluations on agents varying intention (probability of choosing dark-colored wheel) and outcome (probability of shocking outcome (determined by the roulette wheel sets)). (B, E) The probability of choosing different agents in the Testing phase. (C, F) The intention–outcome parameter estimated based on the winning model in Experiments 1 and 2 and its permutation distribution of the parameter. The red dash line indicated the parameter value obtained from real data. Error bars represent 95% confidence intervals.

(Figure 3C). Furthermore, data simulated using the Model 6 reproduced key patterns of behavioral results, wherein participants perceived agents with negative intentions or negative outcomes to have worse intentions and outcomes than was actually the case (Figure 3A,B). Specifically, the ANOVA analysis recovered significant effects of intention and outcome on both intention learning and outcome learning in 100% of simulations. Moreover, parameter recovery analysis demonstrated moderate to strong correlations between simulated and estimated parameter values (all rs > 0.5, $p < 0.01$, Figure S3), indicating Model 6 can reliably recover participants' parameters on average.

The intention–outcome parameter, which quantified the relative weight of intentions and outcomes in moral evaluation based on Model 6, was then calculated. The model exhibited a good fit to the participants' moral evaluations (mean $R^2 = 0.44$, SD = 0.31). A permutation test showed that the intention–outcome parameter from random data was significantly lower than that for the participants ($p < 0.001$; Figure 4C, right panel), further validating the model. The intention–outcome parameter was significantly higher than 0.5 (mean ± SD = 0.72 ± 0.32, $t(29) = 3.87$, $p < 0.001$, Cohen's $d = 0.71$; Figure 4C, left panel), which indicated that participants placed more weight on intention than outcome when making moral evaluations. This confirms the dominant role of intention over outcome in moral evaluations.

3 | Experiment 2

In Experiment 1, agents' outcome (positive/negative) was manipulated through roulette wheel sets used by them. However, the resultant shock probabilities for the four agents (74%, 42%, 58%, 26%; Figure 1B) failed to segregate into two distinct levels, indicating incomplete dichotomization of outcome valence. This might have diminished the usefulness of outcome information in moral evaluations. Additionally, outcome learning depended on the agents' choices, which led to unequal learning for different agents and potential sampling errors (see Supporting Information). Therefore, Experiment 2 was designed to replicate Experiment 1 while addressing these potential confounding factors.

3.1 | Methods

3.1.1 | Experimental Design and Procedure

The experimental set up was very similar to Experiment 1. However, in Experiment 2, participants played three blocks of 120 Learning trials (30 per agent in each block). Each Learning block was followed by a 12-trials Testing block. Furthermore, participants predicted the intentions of each agent for all trials, but predicted the outcomes for only a subset of trials in

TABLE 1 | The intention and outcome of each agent in Experiment 2.

Agent	Intention	Outcome (wheel set)	*p*(dark-colored wheel)	*p*(shock|dark-colored wheel)	*p*(shock|light-colored wheel)	*p*(shock)
1	Negative	Negative (blue)	70%	8/9 = 89%	4/9 = 44%	12/18 = 67%
2	Negative	Positive (orange)	70%	4/9 = 44%	1/9 = 11%	5/18 = 28%
3	Positive	Negative (blue)	30%	8/9 = 89%	5/9 = 56%	13/18 = 72%
4	Positive	Positive (orange)	30%	5/9 = 56%	1/9 = 11%	6/18 = 33%

Note: Agents' intentions were manipulated by varying the probability of choosing dark-colored roulette (*p*(dark-colored roulette wheel): 70% for negative vs. 30% for positive). Agents' outcomes were manipulated by varying the probability of auditory shock on dark-colored roulette wheels (*p*(shock|dark-colored roulette wheel): 89% for negative vs. 44%–56% for positive) and light-colored roulette wheels (*p*(shock|light-colored roulette wheel): 44%–56% for negative vs. 11% for positive) associated with the roulette wheel sets (blue wheels vs. orange wheels). *p*(shock) indicated the overall auditory shock outcomes for each agent (67%–72% for agent with negative outcome vs. 28%–33% for agent with positive outcome). For example, 8/9 in *p*(shock|dark-colored wheel) means that in nine trials where the agent chose a dark-colored wheel, eight resulted in a shock.

order to control for sampling bias. Specifically, when well(ill)-intentioned agents chose dark(light)-colored roulette wheels (30% of trials, 36 in total), participants predicted and received feedback on every trial. For the remaining trials, when the well(ill)-intentioned agents selected light(dark)-colored roulette wheels, that is, selected the roulette wheels in agreement with their intention bias, participants predicted outcome and received outcome feedback for a random subset of nine of the 30 trials for each agent. Thus, participants made and received outcome prediction and feedback for equal numbers of trials for the well- and ill-intentioned choices by the well- and ill-intentioned agents.

As in Experiment 1, we employed a 2 (intention) × 2 (outcome) within-subject design, with consistent intention manipulation. The probabilities that the well(ill)-intentioned agents selected light(dark) roulette wheels on 70% of trials was fixed. We manipulated outcomes by having agents use roulette wheel sets with high/low auditory shock probabilities (agents with positive/negative outcome). By adjusting the shock probabilities generated by different wheel sets, shock outcome for one well-intentioned agent was fixed at 72% of trials even though they more frequently chose the positive intentioned roulette wheel. Similarly, for one the ill-intentioned agents, the auditory shock outcome was set to 28% of trials despite that agent more frequently choosing the negative outcome biased roulette wheel. In this way, positive and negative outcomes were also artificially balanced across agents that were demonstrating positive and negative intentions (Table 1).

3.2 | Results

3.2.1 | Behavioral Results

Here, we report only the statistical tests regarding our central predictions. Complete results can be found in the Supporting Information.

As illustrated in Figures 2D,E and 3D,E, Experiment 2 replicated the effects of intention and outcome on intention learning, outcome learning, moral evaluations and choices in the Testing phase found in Experiment 1. In Experiment 2, both intention and outcome significantly influenced participants' learning of intentions, outcomes, moral evaluations and partner choices in the Testing phase (Supporting Information). Notably, the pattern of outcome learning across different roulette wheel choices by agents was similar to that observed in Experiment 1 (Figures S1 and S2). This confirms that this pattern reflects the effect of intention on outcome learning rather than sampling bias. Notably, during the debriefing session, although participants understood the experiment, 18 of them mentioned that the discontinuous outcome feedback disrupted their outcome learning, at least to some extent. For this reason, the subsequent experiments (Experiments 3a and 3b) returned to outcome learning on all trials, as in Experiment 1, to ensure a continuous learning process.

3.2.2 | Computational Model Results

As in Experiment 1, eight models were tested to examine the interplay way between intention and outcome learning in moral character learning. Outcomes of trials for which participants made no prediction were excluded from the outcome learning analysis. Model comparison revealed that Model 5, which assumed both intention and outcome bias each prediction process, and that first impression effect modulated outcome prediction, had the highest AIC_w values (Figure 3F), indicating it fit best. Furthermore, data simulated using the Model 5 can capture participants' behavioral pattern (Figure 3D,E). Specifically, intention and outcome effects on both intention learning and outcome learning were perfectly recovered (100% of simulations). Moreover, parameter recovery analysis revealed that the correlations between true and recovered parameters were high (all rs > 0.85, $p < 0.001$, Figure S4), indicating Model 5 can reliably recover participants' parameters on average.

The intention–outcome parameter was estimated based on Model 5. The model showed a good fit to the participants' moral evaluations (mean $R^2 = 0.31$, SD = 0.21). This estimation was further validated by a permutation test ($p = 0.014$; Figure 4F, right panel). Moreover, the intention–outcome parameter remained significantly larger than 0.5 (mean ± SD = 0.69 ± 0.29, $t(29) = 3.56$, $p = 0.001$, Cohen's $d = 0.65$; Figure 4F, left panel), which indicated that the perceived intention of the agent had a dominant role over outcome in moral evaluation of that agent, even when controlling for the confounds.

TABLE 2 | The intention and outcome of each agent in Experiment 3.

Agent	Intention	Outcome (wheel set)	*p*(dark-colored wheel)	*p*(shock|dark-colored wheel)	*p*(shock|light-colored wheel)	*p*(shock)
1	Negative	Negative (blue)	70%	18/21 = 86%	4/9 = 46%	22/30 = 73%
2	Negative	Positive (orange)	70%	10/21 = 48%	1/9 = 14%	11/30 = 37%
3	Positive	Negative (blue)	30%	8/9 = 89%	11/21 = 46%	19/30 = 63%
4	Positive	Positive (orange)	30%	5/9 = 56%	3/21 = 14%	8/30 = 27%

Note: Agents' intentions were manipulated by varying the probability of choosing dark-colored roulette (*p*(dark-colored roulette wheel): 70% for negative vs. 30% for positive). Agents' outcomes were manipulated by varying the probability of auditory shock outcome on dark-colored roulette wheels (*p*(shock|dark-colored roulette wheel): 86%–89% for negative vs. 48%–56% for positive) and light-colored roulette wheels (*p*(shock|light-colored roulette wheel): 46% for negative vs. 14% for positive) associated with roulette wheel sets (blue wheels vs. orange wheels). *p*(shock) indicated the overall auditory shock outcomes for each agent (63%–73% for agent with negative outcome vs. 27%–37% for agent with positive outcome). For example, 18/21 in *p*(shock|dark-colored wheel) means that in 21 trials where the agent chose a dark-colored wheel, 18 resulted in a shock.

4 | Experiment 3

In Experiments 1 and 2, participants first learned about the agents' intentions and then learned about the choice-related outcome. This pattern of presentation may have in itself, biased the moral character learning process to result in the greater importance of intention rather than outcome in the evaluation of moral character. Indeed, previous studies have found that the order in which intention and outcome information is presented does indeed influence the use of intention and outcome for moral evaluation [28, 31]. Therefore, Experiments 3a and 3b were conducted to investigate to what extent the presentation order of intentions and outcomes might influence their relative contributions to moral character learning in our experiment.

5 | Experiment 3a

5.1 | Methods

5.1.1 | Experiment Design and Procedure

Experiment 3a followed the same general procedure as Experiments 1 and 2 with the following exceptions. Each participant performed the task in two different versions. The task of "Intention–Outcome" learning Version (IOV) was identical to Experiment 1, but in the task of "Outcome–Intention" learning Version (OIV), the participants were required to predict the outcome (shock or no shock) first, and after receiving feedback on the outcome, they were asked to predict the intention of the agent (light- or dark-colored roulette wheel; Figure 5A). Finally, participants made their moral evaluation of the agent. In each task, participants completed 120 trials in total, divided into three blocks of 40 trials with the four agents appearing 10 times in each block. The sequence of IOV and OIV tasks and the roulette wheel colors were counterbalanced among participants.

Experiment 3a employed a 2 (intention: positive vs. negative) × 2 (outcome: positive vs. negative) × 2 (presentation order: IOV vs. OIV) within-subject design. Table 2 details the frequencies of roulette wheel choices and shock outcomes for each agent.

5.1.2 | Analyses

In Experiment 3a, we performed a 2 (intention) × 2 (outcome) × 2 (presentation order) repeated-measures ANOVA to examine their effects on intention learning, outcome learning, moral evaluations, and choices during the Testing phase, with a specific focus on the impact of presentation order.

Computational models were fitted separately to the IOV and OIV data to identify the interplay way between intention learning and outcome learning. The models for IOV corresponded to those used in Experiment 1. For OIV, we proposed two conceptual frameworks to model the learning process (see Supporting Information). Models 1–8 follow the framework that participants' learning patterns in OIV are identical to those in IOV, even though they receive feedback on the shock outcome before the Intention information is available. In this case, beliefs about the agents' intentions are formed first (IE), followed by associations with auditory shock indication outcomes (OE^{DW}, OE^{LW}), which would be updated after receiving outcome and intention feedback (Models 1–8). In contrast, Models 9–16 follow the framework that participants' learning patterns align with the order in which intention and outcome were presented. Specifically, participants first formed beliefs about agents' outcomes, updated these beliefs based on outcome feedback (OE; outcome learning). They then updated their beliefs about intentions for the given agent on the basis of the outcome (IE^{S}, IE^{NS}), and made intention predictions concerning the roulette wheel selection of the agent (intention learning; Figure 5B). Consistent with the models in IOV, three factors were considered to examine the interplay way between intention learning and outcome learning when outcome presented first: (1) whether intention influenced outcome learning either by biasing participants' prediction probability of the agent's outcome in next trial (prediction bias [Models 1–2, 5–6, 9–10, 13–14]) or by updating their beliefs about the agent's outcome (belief updating [Models 3–4, 7–8, 11–12, 15–16]); (2) whether outcome influenced intention learning either by biasing the prediction of agent's intention in current trial (prediction bias [Models 1, 3, 5, 7, 9, 11, 13, 15]) or by updating participants' beliefs about the agent's intention (belief updating [Models 2, 4, 6, 8, 10, 12, 14]); and (3) whether the first impression effect is considered (Yes [Models 5–8, 13–16] vs. No [Models 1–4, 9–12]; Table S2). Sixteen models were fitted to capture intention and outcome

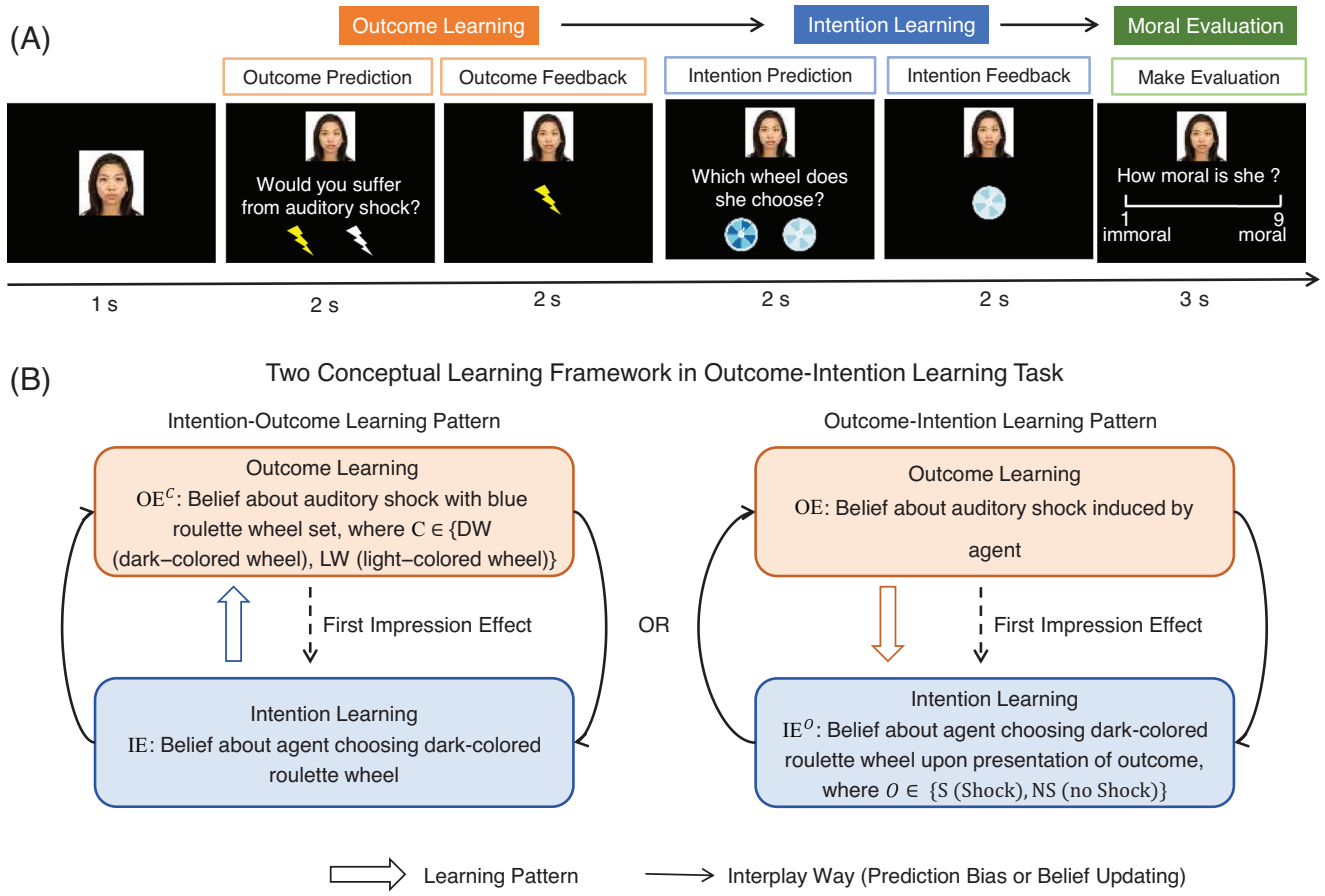


FIGURE 5 | Experiment procedure and computational framework in "Outcome–Intention" learning task. (A) In the "Outcome–Intention" learning task (OIV), the participants were required to predict the outcome (shock or no shock) first, and after receiving feedback on the outcome, they were asked to predict the intention of the agent (light- or dark-colored roulette wheel). (B) Two conceptual learning frameworks were conducted to capture the learning patterns and the interplay way between intention learning and outcome learning in OIV. The first framework (intention–outcome learning pattern) assumed that beliefs about the agents' intentions were initially formed (IE), followed by associations with auditory shock outcomes (OE^C, belief about auditory shock induced by the roulette wheel set [C = DW: dark-colored roulette wheel; C = LW: light-colored roulette wheel] used by the agent), which were then updated after receiving feedback on both outcomes and intentions (Models 1–8). The second framework (outcome–intention learning pattern, Models 9–16) posited that participants learning about agents' intention and outcome followed by presentation order of intention and outcome. Participants initially formed and updated beliefs about the agents' outcome (OE), then predicted and updated the belief about intention for the given agent on the basis of the outcome (IE^O, belief about choosing dark-colored roulette wheel when there was an auditory shock [O = S] or no auditory shock [O = NS]). In both learning frameworks, we examine whether the interplay between intention and outcome learning is through prediction bias or belief updating. Additionally, we consider the potential influence of the first impression effect in both learning frameworks.

learning in OIV, using a stepwise approach at the individual level, similar to Experiment 1. Model validation and parameter recovery analyses were conducted separately for the IOV and OIV paradigms, following the same procedures implemented in Experiment 1. A paired-sample *t*-test was conducted to evaluate differences in intention–outcome parameters between the IOV and OIV conditions.

5.2 | Results

5.2.1 | Behavioral Results

Experiment 3a showed that the participants successfully learned the intention and outcomes of the agents in both the IOV and OIV conditions (intention prediction accuracy: IOV: mean ± SD = 0.57 ± 0.06, $t(29) = 5.86$, $p < 0.001$, Cohen's $d = 1.07$; OIV: mean ± SD = 0.61 ± 0.07, $t(29) = 9.11$, $p < 0.001$, Cohen's $d = 1.66$; outcome prediction accuracy: IOV: mean ± SD = 0.61 ± 0.04, $t(29) = 14.61$, $p < 0.001$, Cohen's $d = 2.67$; OIV: mean ± SD = 0.55 ± 0.05, $t(29) = 5.89$, $p < 0.001$, Cohen's $d = 1.08$).

The 2 (intention) × 2 (outcome) × 2 (presentation order) repeated-measures ANOVA on intention predictions revealed significant main effects for intention ($F(1,29) = 87.75$, $p < 0.001$, $\eta_p^2 = 0.75$) and outcome ($F(1,29) = 41.34$, $p < 0.001$, $\eta_p^2 = 0.59$; Figure 6A,B), and no significant effect for presentation order or any interactions between presentation order and either intention or outcome (*p*s > 0.05). These results were identified as previous experiments that both intention and outcome affected intention learning, regardless of presentation order.

For outcome learning, participants perceived more likely to receive a auditory shock not only for agents with negative out-

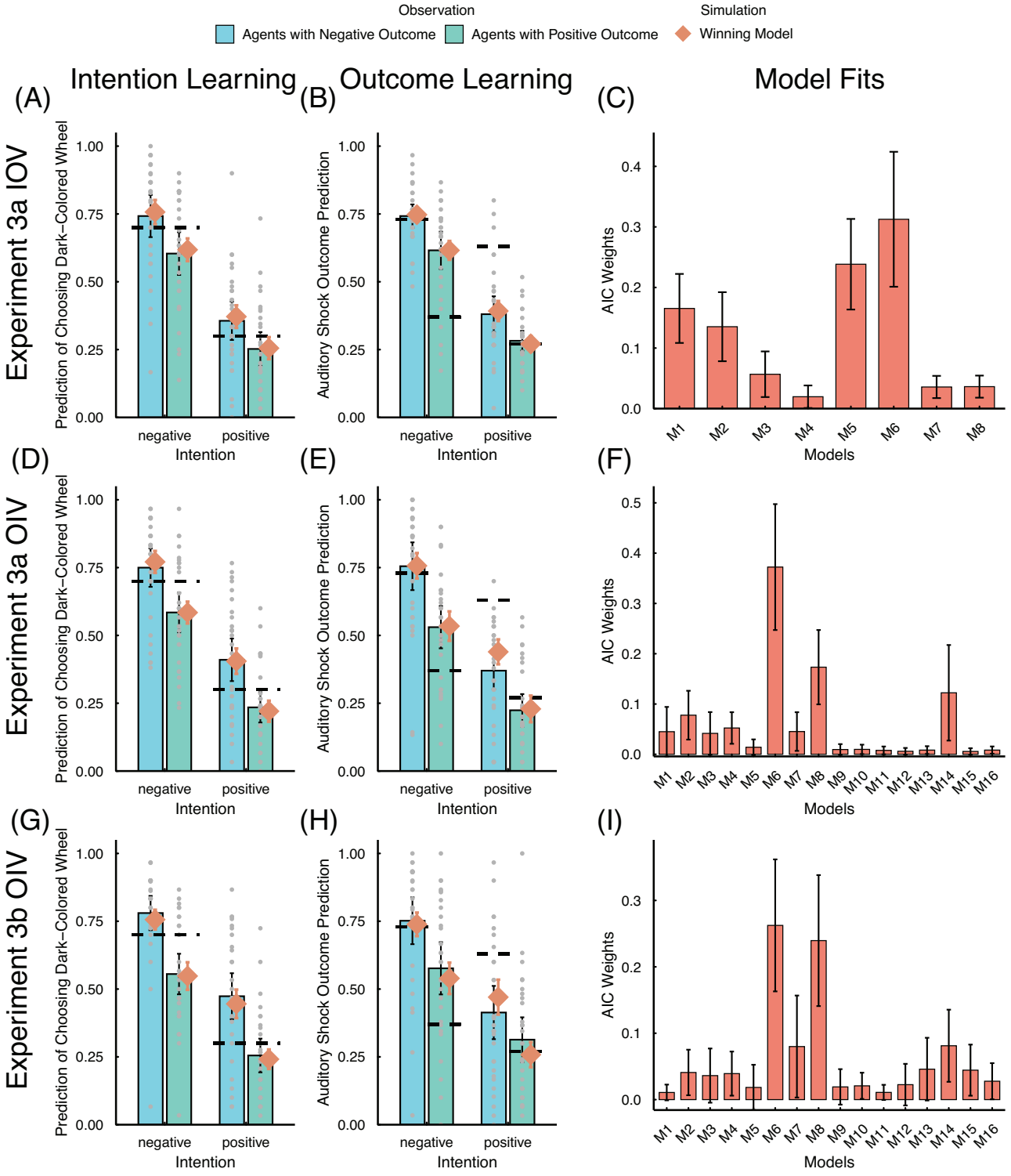


FIGURE 6 | Intention learning and outcome learning in Experiment 3. (A, D, G) Intention learning indicated by participants' predictions of four agents' dark wheel choices. Participants predicted more dark wheel choices for agents with negative intention and negative outcome in both IOV and OIV condition. Dashed lines indicate ground-truth intention probabilities. (B, E, H) Outcome learning measured by shock outcome predictions. Participants predicted the presence of shock outcomes more frequently for agents with negative intentions and negative outcomes in both IOV and OIV condition. Dashed lines represent ground-truth shock probabilities for four agents. (C, F, I) Model comparison using AIC weights in Experiment 3. AIC weights are averaged across participants. Higher AIC weights (red bars) indicate better model fit. Orange diamonds represented predictions from the best-fitting model (highest AIC weight, asterisked). IOV: "Intention–Outcome" learning task; OIV: "Outcome–Intention" learning task. Error bars represent 95% confidence intervals.

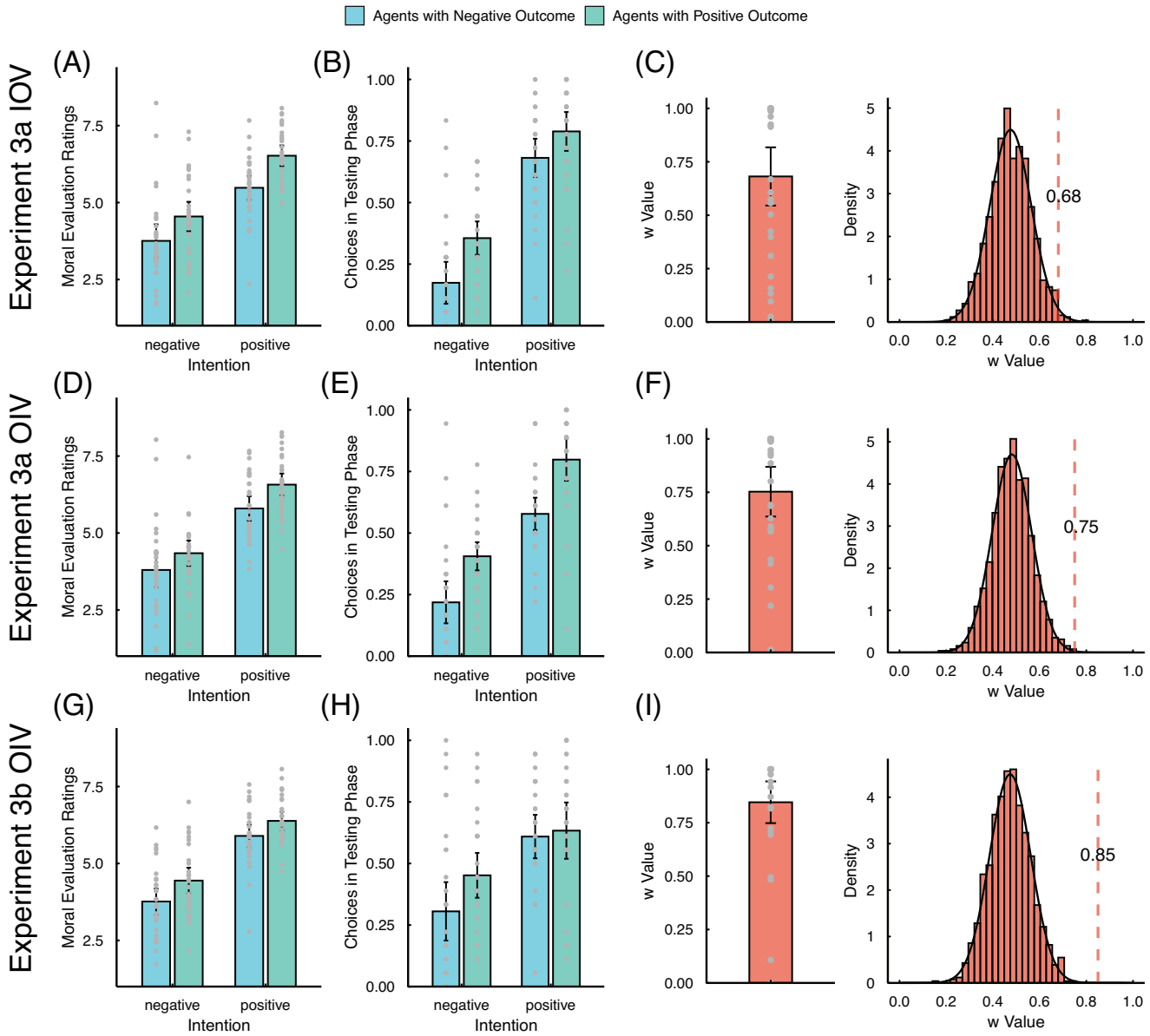


FIGURE 7 | Moral evaluations and choices in the Testing phase in Experiment 3. (A, D, G) The degree which participants made moral evaluations on agents varying intention and outcome. Participants rated agents with negative intention and negative outcome as less moral in both IOV and OIV condition. (B, E, H) The probability of choosing different agents in the Testing phase. Participants are less likely to choose agents with negative intention and negative outcome as partner in the Testing phase in both IOV and OIV condition. (C, F, I) The intention–outcome parameter estimated based on the winning model in Experiment 3 and its permutation distribution of the parameter. The red dash line indicated the parameter value obtained from real data. IOV: "Intention–Outcome" learning task; OIV: "Outcome–Intention" learning task. Error bars represent 95% confidence intervals.

comes ($F(1,29) = 32.10$, $p < 0.001$, $\eta_p^2 = 0.53$) and but also for agents with negative intentions ($F(1,29) = 143.95$, $p < 0.001$, $\eta_p^2 = 0.83$; Figure 6D,E). Unlike intention learning, there was a significant interaction between outcome and presentation order ($F(1,29) = 4.73$, $p = 0.038$, $\eta_p^2 = 0.14$), outcome having a stronger effect in the OIV condition than in the IOV condition. These findings suggest outcome learning was influenced by both intention and outcome, and to a lesser extent by presentation order.

In contrast, with respect to the judgments of the moral character of the agents, participants rated agents with positive intention ($F(1,29) = 61.83$, $p < 0.001$, $\eta_p^2 = 0.68$) or positive outcome ($F(1,29) = 28.21$, $p < 0.001$, $\eta_p^2 = 0.49$) as more moral (Figure 7A,B). No significant effect of presentation order or any interaction effects were found ($ps > 0.05$). However, with respect to partner choices in the Testing phase, participants were consistent with previous studies, as they showed a preference for agents with positive intention ($F(1,29) = 142.66$, $p < 0.001$, $\eta_p^2 = 0.83$) and positive outcome ($F(1,29) = 12.56$, $p = 0.001$, $\eta_p^2 = 0.30$; Figure 7D,E). Beyond these main effects, we also observed an interaction effect between intention and presentation order ($F(1,29) = 5.58$, $p = 0.025$, $\eta_p^2 = 0.16$), where the effect of intention was stronger in the OIV condition than in the IOV condition. These findings suggest that both intention and outcome affected moral evaluations and partner choices, with presentation order having a lesser impact on partner choices.

5.2.2 | Computational Model Results

Here, we report only the statistical tests regarding our central predictions. Complete results can be found in the Supporting Information.

We fitted models for IOV and OIV conditions separately to examine the order of presentation effect on learning patterns and trade-offs between intention and outcome. Replicating Experiment 1, Model 6 best fitted IOV data (Figure 6C), indicating that intention biased outcome predictions, while outcome influenced intention learning through belief updating, and first impression effect modulated outcome prediction (see Supporting Information for details of the computational results in the IOV condition).

In the OIV condition, model comparisons using AIC_w identified Model 6 as the best-fitting model (Figure 6F). The behavioral patterns simulated by Model 6 closely matched the empirical data (Figure 6D,E). Critically, the effects of intention and outcome on both intention learning and outcome learning were fully recovered across all simulations (100%). Parameter recovery analyses further validated the model, showing strong correlations between simulated parameters and their true values (all $r > 0.90$, $p < 0.001$; Figure S6). These results suggest that even when outcomes were presented before intentions, participants primarily associated intention (rather than outcome) with their moral evaluations of agents. Additionally, we consistently observed that intention biased participants' outcome predictions, while outcome influenced the intention updating process, with first impression effect modulated intention prediction in the OIV condition.

We further examined the intention–outcome parameter estimated separately for the IOV and OIV conditions based on Model 6 in each condition. The model indicated a good fit to the participants' moral evaluations in both conditions (IOV condition: mean $R^2 = 0.46$, SD = 0.23; OIV condition: mean $R^2 = 0.47$, SD = 0.28). The permutation test indicated that only eight out of 1000 random permutations produced intention–outcome parameters exceeding the observed value in the IOV condition ($p = 0.008$; Figure 7C, right panel) and none of the randomly generated intention–outcome parameters exceeded participants' actual values in the OIV condition ($p < 0.001$; Figure 7F, right panel), supporting the intention–outcome parameters' validity. We observed that participants consistently weighted intention more when making moral evaluations for both IOV and OIV condition (Figure 7C,F, left panels; IOV condition: M ± SD = 0.68 ± 0.36, $t(29) = 2.72$, $p = 0.011$, Cohen's $d = 0.50$; OIV condition: M ± SD = 0.75 ± 0.31, $t(29) = 4.44$, $p < 0.001$, Cohen's $d = 0.81$). However, the intention–outcome parameter did not show significant difference between IOV condition and OIV condition ($t(29) = -0.84$, $p = 0.410$, Cohen's $d = -0.15$, $BF_{10} = 0.27$). To verify whether our computational model captures the observed presentation order effect, we compared the learning rate parameters between conditions. The analysis revealed that outcome learning rates were significantly higher in the OIV condition compared to the IOV condition, whereas intention learning rates showed no significant difference (see Supporting Information for detailed statistical results). This parametric pattern suggests that the model successfully accounts for the heightened sensitivity to outcome prediction errors in the OIV condition. It is worth noting that testing sequence may be a potential confound, in which participants assigned greater weight to intention in moral evaluations under the OIV condition when the OIV task was conducted before the IOV task ($t(28) = -3.88$, $p < 0.001$, Cohen's $d = -1.42$). However, the testing sequence effect was only observed in OIV condition, but not in the IOV condition ($t(28) = -0.35$, $p = 0.731$, Cohen's $d = -0.13$).

6 | Experiment 3b

6.1 | Methods

To eliminate the potential influence of test sequence on intention–outcome parameter in the OIV condition, Experiment 3b recruited a new sample of 30 participants who only conducted the OIV task, exactly as described in Experiment 3a. We then compared their intention–outcome parameter with the IOV results without effect of testing sequence in Experiment 3a to further investigate this presentation order effect on the balance of intention and outcome for moral evaluations.

6.2 | Results

Here, we report only the results regarding our central predictions. In Experiment 3b, we consistently found that both intention and outcome influenced participants' learning of intention and outcome, moral evaluation and choices in the Testing phase (Figures 6G,H and 7G,H). Computational model results revealed that Model 6 was the best-fitting model broadly in agreement with the results of the OIV condition in Experiment 3a (Figure 6I; see Supporting Information for details).

Furthermore, we compared the intention–outcome parameter differences between IOV condition in Experiment 3a and OIV condition in Experiment 3b, controlling for the effect of testing sequence using an independent sample t-test. The OIV condition showed a marginally significant difference from the IOV condition, $t(58) = -2.005$, $p = 0.050$, Cohen's $d = -0.52$ (Figure 7C,I). A Bayesian independent samples t-test provided anecdotal evidence for a difference ($BF_{10} = 1.38$). These results suggest no meaningful difference in the weighting of intention between IOV condition and OIV condition, despite the effect approaching statistical significance.

7 | General Discussion

Intention and outcome are two primary dimensions in learning about others' moral character. This study developed a task that quantified learning about intentions and outcomes separately to examine their interplay and relative weighting for moral evaluations. Three experiments were conducted to control for potential confounds such as sampling bias, presentation order, and testing sequence. Results showed that participants integrated beliefs about agents' intentions and outcomes when making moral evaluations and partner choices. When learning about agents' moral characters, both intentions and outcomes concurrently influenced participants' intention and outcome learning separately. Computational modeling further revealed distinct interplay patterns between intention and outcome learning.

Across all experiments, intentions were weighted more heavily than outcomes in shaping moral evaluations, which confirms the dominant role of intentions in moral evaluation.

Our findings align with previous research that demonstrates intentions and outcomes are two distinct components integrated into moral evaluations and partner choices [11, 19, 32, 33]. When examining how intention and outcome jointly influence the process of moral learning, two alternative hypotheses were tested. One proposed that individuals initially learn intentions and outcomes separately before integrating them for moral evaluation or partner choices, as suggested by traditional accounts [19, 20]. The other posits that moral outcomes shape intention learning from initial learning stages, and intentions also influence outcome learning from the beginning. Our new paradigm makes both intentions and outcomes equally apparent. Participants predicted and received feedback on intentions and outcomes separately. Thus, this paradigm can eliminate the possible individual differences in inferring implicit intentions. Furthermore, it allows us to address intention and outcome learning separately. As a result, we found that two agents with the same intention were perceived differently depending on whether the outcomes were better or not, indicating intention learning itself was biased by outcome. As for outcome learning, intentions also drove outcome learning. Outcomes with better intentions were predicted to be better, despite the outcomes were not always determined by intentions in current design. Experiment 2 controlled sampling bias in outcome learning and replicated the behavioral patterns from Experiment 1, ruling out that insufficient learning, caused by sampling bias, was a confound. Experiment 3 further reversed the presentation order of intentions and outcomes to test whether the bidirectional effects were influenced by the order of intention and outcome presentation. Participants first predicted and received outcome feedback, and then predicted intentions, and received intention feedback. Results showed the bidirectional influence of intention and outcome remained stale even when the order of information presentation was reversed. Together, this evidence suggests a mutual influence between intention and outcome learning before integration. Intentions shape outcome learning and outcomes bias intention learning.

We used reinforcement learning models to identify how intention learning and outcome learning influence each other, through prediction bias or belief updating. Prediction bias suggests that feedback about either intentions or outcomes biases subsequent predictions about the other. Thus, participants tended to predict better or worse outcomes/intentions when they perceived the other factor to be better or worse, similar to "choice stickiness" effects [34]. Belief updating posits that either intention or outcome feedback would update beliefs about the other. That is, feedback on outcomes not only updated beliefs about outcomes but also updated beliefs about intentions, and vice versa. Prior research has extensively explored the influence of intention on outcome evaluations [11, 16, 25, 26, 32, 35], our focus was on whether outcomes lead to re-assessment of intentions. In our task, outcomes were probabilistic events, determined by the roulette wheels, that mirrored real-life situations where outcomes do not always align with intentions. Thus, positive intentions sometimes produced negative outcomes, and vice versa. We asked whether "good intentions" that yield negative outcomes are re-evaluated, potentially diminishing their perceived goodness. This re-evaluation was measured in our computational model as belief updating. As predicted, we observed distinct patterns of mutual influence between intention and outcome learning. Intentions biased outcome predictions, where better or worse intention lead to better or worse outcome predictions by choice stickiness. Moreover, outcome feedback directly updated beliefs about intentions, where better or worse outcome lead to beliefs modification about intentions. This influence of outcome on intention learning, although counter intuitive, given the probabilistic nature of outcomes, suggests individuals adaptively revise their evaluation of intentions based on observed outcomes [36]. Such direct modification of intention beliefs reflect the process of value shaping, representing a more efficient learning mechanism that facilitates the convergence of moral character impression [37]. This finding is consistent with recent research demonstrating that individuals adaptively adjust their reliance on moral strategies and perceptions according to outcome feedback [21]. Experiment 3 further showed that the mutual influence was robust, regardless of the presentation order of intentions and outcomes. Even when outcomes were presented before intentions, intentions still biased outcome predictions, and outcomes continued to update beliefs about intentions. Additional modeling confirmed that presentation order did not affect learning patterns, with intentions, rather than outcomes, directly associating with beliefs about agents, showing a preponderant role of intention in moral learning.

We proposed an intention–outcome parameter that directly quantifies the relative weight of beliefs about intentions and outcomes in moral evaluations. Unlike prior studies that simply compared agents with good intentions but bad outcomes versus bad intentions but good outcomes [6, 33], our model efficiently accounts for individual differences by assigning each agent a unique parameter, and captures the dynamic interplay between intentions and outcomes. Finally, we consistently found that intentions carried more weight than outcomes, regardless of the presentation order. Even after adjusting agents' final outcomes to improve outcome significance, and controlling for sampling bias, moral evaluations remained predominantly driven by beliefs about intentions. Our findings extend prior work by using a computational modeling framework to provide quantitative evidence for the "privileged status" of intentions in moral evaluation.

This study has several limitations. In Experiment 2, when controlling for sampling bias, outcome learning became discontinuous. In this context, outcome feedback biased intention prediction rather than belief updating. This suggests that the outcome effect may be a context-adaptive strategy. When feedback is intermittent, outcomes appear to bias transient predictions without permanently modifying intention beliefs [37]. However, the boundary of this shift remains unresolved in our study and further research will be needed to clarify this. Moreover, while effective for capturing general learning patterns, reinforcement learning models may oversimplify the process of intention and outcome learning [38]. Further research is needed to evaluate more complex models to account for these learning processes better. Finally, our findings demonstrated that presentation order may asymmetrically influence outcome learning and partner choice (vs. intention learning and moral evaluations). Future research should explore this asymmetry, extending beyond current computational models focused on interplay way between

intention and outcome learning to further clarify how presentation order selectively influence moral learning.

8 | Conclusion

This study advances decades of research on how intentions and outcomes are weighed in moral learning by experimentally dissociating their presentation. Our results indicate that intentions and outcomes are not processed independently before integration. Rather, their mutual influence emerges early in learning. Computational models revealed an asymmetric influence of intentions and outcomes in moral learning. Intentions primarily drive moral character learning, while outcomes indirectly influence moral character learning by modifying the beliefs about intentions. Thus, our study provides a new paradigm and insight to understand how intentions and outcomes are processed in moral character learning.

Author Contributions

Gaojie Huang: conceptualization, methodology, software, formal analysis, validation, resources, data curation, writing – original draft, writing – review and editing, visualization. **Yongbo Xu**: conceptualization, methodology, software, formal analysis, investigation, resources, writing – original draft. **Edmund Derrington**: conceptualization, writing – review and editing. **Jean-Claude Dreher**: conceptualization, writing – review and editing, supervision, funding acquisition. **Chen Qu**: conceptualization, writing – original draft, writing – review and editing, supervision, funding acquisition.

Acknowledgments

This work was supported by the Program for National Natural Science Foundation of China (32171019), the Research Center for Brain Cognition and Human Development, Guangdong, China (2024B0303390003), and the MOE Project of Key Research Institute of Humanities and Social Sciences in Universities (22JJD190004). Additional support was provided by IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investis95%CIents d'Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investis95%CIents d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche to Jean-Claude Dreher (ANR-24-CE37-4261).

Ethics Statement

This study was approved by the Human Research Ethics Committee of the School of Psychology, South China Normal University (Approval No. SCNU-PSY-2021-050). All participants provided written informed consent prior to their participation in accordance with the Declaration of Helsinki.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All primary data and analysis scripts are publicly available https://osf.io/f7qad/?view_only=06c0e4e09bd946f097d41aef949486c8.

References

1. F. Cushman, V. Kumar, and P. Railton, "Moral Learning: Psychological and Philosophical Perspectives," *Cognition* 167 (2017): 1–10, https://doi.org/10.1016/j.cognition.2017.06.008.

2. C. Qu, J. Bénistant, and J.-C. Dreher, "Neurocomputational Mechanisms Engaged in Moral Choices and Moral Learning," *Neuroscience & Biobehavioral Reviews* 132 (2022): 50–60, https://doi.org/10.1016/j.neubiorev.2021.11.023.

3. R. Hartman, W. Blakey, and K. Gray, "Deconstructing Moral Character Judgments," *Current Opinion in Psychology* 43 (2022): 205–212, https://doi.org/10.1016/j.copsyc.2021.07.008.

4. E. L. Uhlmann, D. A. Pizarro, and D. Diermeier, "A Person-Centered Approach to Moral Judgment," *Perspectives on Psychological Science* 10, no. 1 (2015): 72–81, https://doi.org/10.1177/1745691614556679.

5. M. J. Crockett, J. A. C. Everett, M. Gill, and J. Z. Siegel, "Chapter One—The Relational Logic of Moral inference," in *Advances in Experimental Social Psychology*, ed. B. Gawronski (Academic Press, 2021): 1–64, https://doi.org/10.1016/bs.aesp.2021.04.001.

6. M. Gummerum and M. T. Chu, "Outcomes and Intentions in Children's, Adolescents00027;, and Adults00027; Second- and Third-Party Punishment Behavior," *Cognition* 133, no. 1 (2014): 97–103, https://doi.org/10.1016/j.cognition.2014.06.001.

7. Z. Zhang, P. Grocke, and M. Tomasello, "The Influence of Intention and Outcome on Young Children's Reciprocal Sharing," *Journal of Experimental Child Psychology* 187 (2019): 104645, https://doi.org/10.1016/j.jecp.2019.05.012.

8. C. Feng, Q. Yang, L. Azem, et al., "An fMRI Investigation of the Intention-Outcome Interactions in Second- and Third-Party Punishment," *Brain Imaging and Behavior* 16, no. 2 (2022): 715–727, https://doi.org/10.1007/s11682-021-00555-z.

9. B. C. Hilton and V. A. Kuhlmeier, "Intention Attribution and the Development of Moral Evaluation," *Frontiers in Psychology* 9 (2019): 2663, https://doi.org/10.3389/fpsyg.2018.02663.

10. S. Li, G. Huang, Z. Ma, and C. Qu, "Superior Bias in Trust-Related Decisions," *Current Psychology* 42 (2023): 24822–24836, https://doi.org/10.1007/s12144-022-03567-0.

11. H. Yu, J. Li, and X. Zhou, "Neural Substrates of Intention-Consequence Integration and Its Impact on Reactive Punishment in Interpersonal Transgression," *Journal of Neuroscience* 35, no. 12 (2015): 4917–4925, https://doi.org/10.1523/JNEUROSCI.3536-14.2015.

12. S. Li, X. Hao, Y. Mei, Y. Cheng, N. Sun, and C. Qu, "How Adolescents and Adults Learn About Changes in the Trustworthiness of Others Through Dynamic Interaction," *Frontiers in Psychology* 12 (2021): Article 690494, https://doi.org/10.3389/fpsyg.2021.690494.

13. P. L. Lockwood and J.-C. Dreher, "Moral Learning and Decision-Making Across the Lifespan," *Annual Review of Psychology* 76, no. 1 (2025): 475–500.

14. L. M. Hackel, B. B. Doll, and D. M. Amodio, "Instrumental Learning of Traits Versus Rewards: Dissociable Neural Correlates and Effects on Choice," *Nature Neuroscience* 18, no. 9 (2015): 1233–1235, https://doi.org/10.1038/nn.4080.

15. T.-T. A. Nong, C. Qu, Y. Li, et al., "Computational Mechanisms Underlying the Emergence of Theory of Mind in Children," PsyArXiv (2023), https://doi.org/10.31234/osf.io/y876r.

16. R. Philippe, R. Janet, K. Khalvati, R. P. Rao, D. Lee, and J. C. Dreher, "Neurocomputational Mechanisms Involved in Adaptation to Fluctuating Intentions of Others," *Nature Communications* 15, no. 1 (2024): 3189, https://doi.org/10.1038/s41467-024-47491-2.

17. L. Young, A. Bechara, D. Tranel, H. Damasio, M. Hauser, and A. Damasio, "Damage to Ventromedial Prefrontal Cortex Impairs Judgment of Harmful Intent," *Neuron* 65, no. 6 (2010): 845–851, https://doi.org/10.1016/j.neuron.2010.03.003.

18. F. Schwartz, H. Djeriouat, and B. Trémolière, "Judging Accidental Harm: Reasoning Style Modulates the Weight of Intention and Harm Severity," *Quarterly Journal of Experimental Psychology* 75, no. 12 (2022): 2366–2381, https://doi.org/10.1177/17470218221089964.

19. H. J. Cho and L. M. Hackel, "Instrumental Learning of Social Affiliation Through Outcome and Intention," *Journal of Experimental Psychology: General* 151, no. 9 (2022): 2204–2221, https://doi.org/10.1037/xge0001190.

20. L. M. Hackel, P. Mende-Siedlecki, and D. M. Amodio, "Reinforcement Learning in Social Interaction: The Distinguishing Role of Trait Inference," *Journal of Experimental Social Psychology* 88 (2020): 103948, https://doi.org/10.1016/j.jesp.2019.103948.

21. M. Maier, V. Cheung, and F. Lieder, "Learning From Outcomes Shapes Reliance on Moral Rules Versus Cost–Benefit Reasoning," *Nature Human Behaviour* (2025), https://doi.org/10.1038/s41562-025-02271-w.

22. L. Zhang, L. Lengersdorff, N. Mikus, J. Gläscher, and C. Lamm, "Using Reinforcement Learning Models in Social Neuroscience: Frameworks, Pitfalls and Suggestions of Best Practices," *Social Cognitive and Affective Neuroscience* 15, no. 6 (2020): 695–707, https://doi.org/10.1093/scan/nsaa089.

23. F. Cushman, "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment," *Cognition* 108, no. 2 (2008): 353–380, https://doi.org/10.1016/j.cognition.2008.03.006.

24. J. Geipel, C. Hadjichristidis, and L. Surian, "Foreign Language Affects the Contribution of Intentions and Outcomes to Moral Judgment," *Cognition* 154 (2016): 34–39, https://doi.org/10.1016/j.cognition.2016.05.010.

25. P. Y. Hirozawa, M. Karasawa, and A. Matsuo, "Intention Matters to Make You (Im)Moral: Positive-Negative Asymmetry in Moral Character Evaluations," *Journal of Social Psychology* 160, no. 4 (2020): 401–415, https://doi.org/10.1080/00224545.2019.1653254.

26. A. Sarin, D. A. Lagnado, and P. W. Burgess, "The Intention-Outcome Asymmetry Effect: How Incongruent Intentions and Outcomes Influence Judgments of Responsibility and Causality," *Experimental Psychology* 64 (2017): 124–141, https://doi.org/10.1027/1618-3169/a000359.

27. K. Gray and J. E. Keeney, "Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality," *Social Psychological and Personality Science* 6, no. 8 (2015): 859–868, https://doi.org/10.1177/1948550615592241.

28. L. Leloup, G. Meert, and D. Samson, "Moral Judgments Depend on Information Presentation: Evidence for Recency and Transfer Effects," *Psychologica Belgica* 58, no. 1 (2018): 256, https://doi.org/10.5334/pb.421.

29. F. Cushman, K. Gray, A. Gaffey, and W. B. Mendes, "Simulating Murder: The Aversion to Harmful Action," *Emotion (Washington, DC)* 12, no. 1 (2012): 2–7, https://doi.org/10.1037/a0025071.

30. S. Palminteri, G. Lefebvre, E. J. Kilford, and S. J. Blakemore, "Confirmation Bias in Human Reinforcement Learning: Evidence From Counterfactual Feedback Processing," *PLOS Computational Biology* 13, no. 8 (2017): e1005684, https://doi.org/10.1371/journal.pcbi.1005684.

31. C. F. Surber, "Development Processes in Social Inference: Averaging of Intentions and Consequences in Moral Judgment," *Developmental Psychology* 13, no. 6 (1977): 654–665, https://doi.org/10.1037/0012-1649.13.6.654.

32. B. G. Babür, Y. C. Leong, C. X. Pan, and L. M. Hackel, "Neural Responses to Social Rejection Reflect Dissociable Learning About Relational Value and Reward," *Proceedings of the National Academy of Sciences* 121, no. 49 (2024): e2400022121, https://doi.org/10.1073/pnas.2400022121.

33. L. Young and R. Saxe, "The Neural Basis of Belief Encoding and Integration in Moral Judgment," *Neuroimage* 40, no. 4 (2008): 1912–1920, https://doi.org/10.1016/j.neuroimage.2008.01.057.

34. W. Kool, S. J. Gershman, and F. A. Cushman, "Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems," *Psychological Science* 28, no. 9 (2017): 1321–1333, https://doi.org/10.1177/0956797617708288.

35. A. Falk, E. Fehr, and U. Fischbacher, "Testing Theories of Fairness—Intentions Matter," *Games and Economic Behavior* 62, no. 1 (2008): 287–303, https://doi.org/10.1016/j.geb.2007.06.001.

36. L. M. Hackel, D. A. Kalkstein, and P. Mende-Siedlecki, "Simplifying Social Learning," *Trends in Cognitive Sciences* 28, no. 5 (2024): 428–440, https://doi.org/10.1016/j.tics.2024.01.004.

37. H. Suganuma, K. Katahira, H. Ohtsuki, and T. Kameda, "How Social Learning Enhances—or Undermines—Efficiency and Flexibility in Collective Decision-Making Under Uncertainty," *Proceedings of the National Academy of Sciences* 122, no. 48 (2025): e2516827122, https://doi.org/10.1073/pnas.2516827122.

38. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction* (MIT Press, 1998).

39. D. S. Ma, J. Correll, and B. Wittenbrink, "The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data," *Behavior Research Methods* 47, no. 4 (2015): 1122–1135, https://doi.org/10.3758/s13428-014-0532-5.

40. E. J. Wagenmakers and S. Farrell, "AIC model selection using Akaike weights," *Psychonomic Bulletin & Review* 11, no. 1 (2004): 192–196, https://doi.org/10.3758/BF03206482.

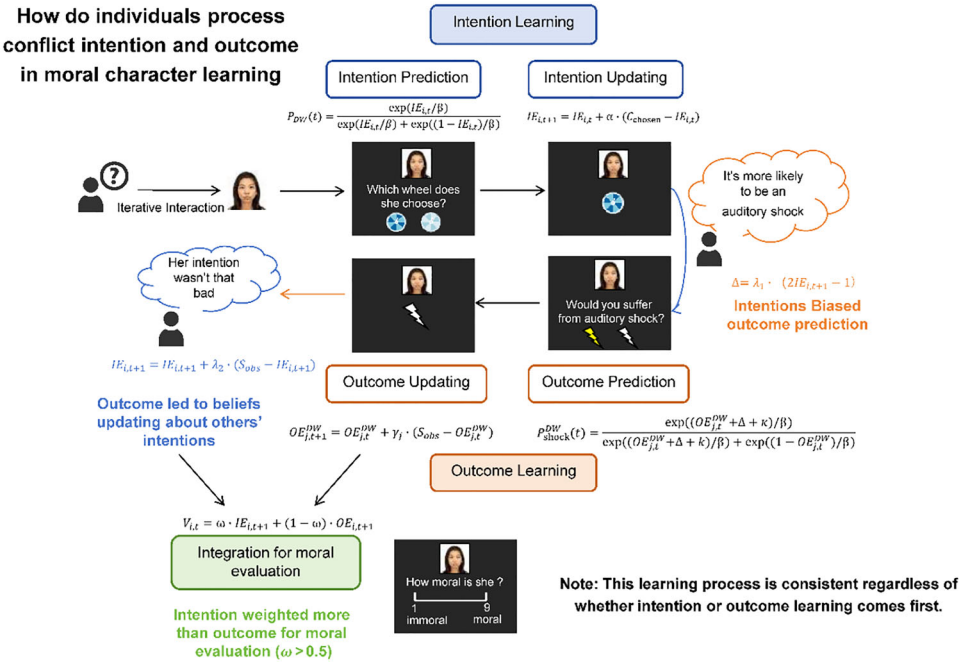
Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting Information: nyas70233-sup-0001-SuppMat.docx

Graphical Abstract

Please note that Graphical Abstracts only appear online as part of a table of contents and are not part of the main article (therefore, they do not appear in the article HTML or PDF files).



How do individuals process conflicting intention and outcome when learning about another's moral character? Our findings reveal that intentions and outcomes mutually influence each other prior to integrating them for moral evaluation. Specifically, intentions biased predictions about outcomes, while observed outcomes retroactively updated beliefs about intentions. This bidirectional influence occurred regardless of whether intention or outcome was presented first. Furthermore, intentions consistently carried more weight than outcomes in moral evaluations.