Running head: ALIGNMENT BIAS IN ADVISING

**Do we advise as one likes? The alignment bias in social advice giving**

Xitong Luo1, Lei Zhang2,3,4\*, Yafeng Pan1,5\*

*1 Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China*

*2 Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK*

*3 Institute for Mental Health, School of Psychology, University of Birmingham, Birmingham, UK*

*4 Centre for Developmental Science, School of Psychology, University of Birmingham, Birmingham, UK*

*5 The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China*

**\* Corresponding authors:**

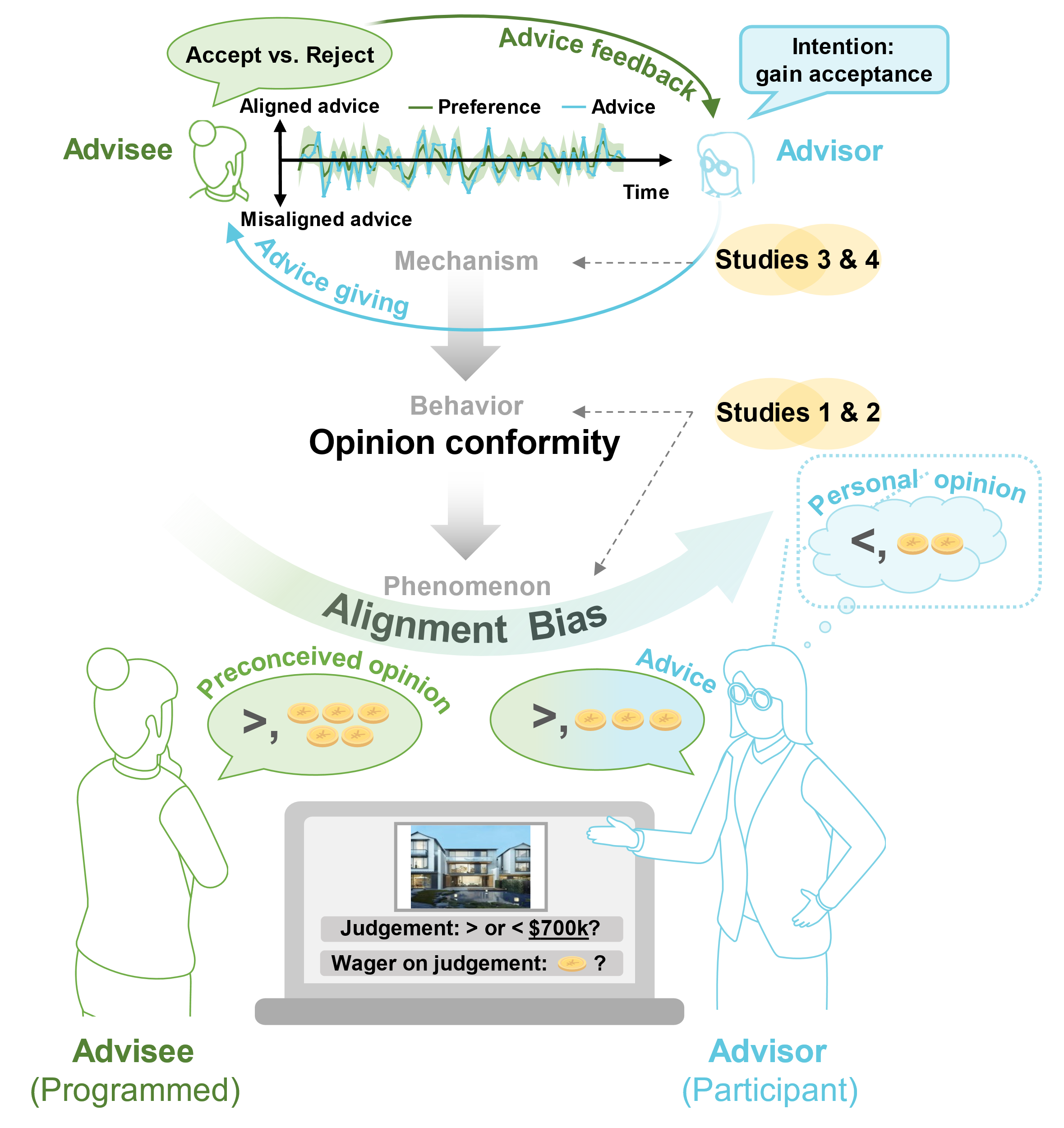
Yafeng Pan (yafeng.pan@zju.edu.cn) or Lei Zhang (l.zhang.13@bham.ac.uk)

**Abstract**

We give advice to influence the advisees, but is our advice, in turn, also influenced by the advisees? This question is largely neglected in the field of social cognition. Here, we systematically examine how advisees’ opinions delicately shape the advice-giving process. Across four studies (*n* = 346, student samples), we reveal the robust existence of the alignment bias—advisors tend to align their advice with advisees’ opinions—primarily yielded by the conformity to advisees’ opinions (Studies 1 & 2). This biased tendency evolved into adaptive strategies tailoring advisees’ fluctuating preferences, as which became disclosed by their feedback (accepting or rejecting advice) (Studies 3 & 4). Reinforcement learning models, featuring advisors’ vulnerability to feedback, best captured advisors’ strategy adaptations and supported that the acceptance-directed intention exclusively underlies the alignment bias. Our research sheds light on a nuanced interplay between advisees and advisors: advice interactions, commonly seen as informative, surprisingly foster false beliefs and opinion homogeneity and polarization.

**Keywords**: advice giving; advisee-advisor interaction; advice taking; social influence; reinforcement learning; computational modelling

**Graphical abstract**

**Research Transparency Statement**

**General Disclosures**

* Conflicts of interest: All authors declare no conflicts of interest.
* Funding: This work was supported by the National Natural Science Foundation of China (Nos. 62207025 and 62337001), the Humanities and Social Sciences Research Project from the Ministry of Education of China (No. 22YJC190017), and the Fundamental Research Funds for the Central Universities to Y.P.
* Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article.
* Ethics: All experiments in this paper received approval from the Department of Psychology and Behavioral Sciences, Zhejiang University (No. 2023015), and were conducted in accordance with the Helsinki Declaration.
* Computational reproducibility: All results in this paper are computationally reproducible via the Open Science Framework <https://osf.io/aehxr/>.

**Study 1-4**

Preregistration: No aspects of these studies were preregistered. Data: All primary data are publicly available via the Open Science Framework (<https://osf.io/aehxr/>). Analysis scripts: All analysis scripts (including the custom code written in R and MATLAB) are publicly available via the Open Science Framework (<https://osf.io/aehxr/>).

**Introduction**

Advice facilitates the transfers of information and social influence, functioning as an essential gear in human society (Hertz et al., 2017; Morin et al., 2021; Yaniv, 2004). From learning cooking tips with a friend to seeking career guidance from mentors, the beliefs and decisions of individuals (advisees, those who receive advice) are greatly influenced by advisors’ opinions (Bonaccio & Dalal, 2006; Kämmer et al., 2023; Meshi et al., 2012). In turn, from recommending goods to customers to advocating policies to the public, advisors’ success in exerting influence on others and developing reputations highly relies on advisees’ attitudes and reactions to their advice (Belkin & Kong, 2018; Blunden et al., 2019; Hertz et al., 2017; Kurvers et al., 2021).

The fact that advisors value how their advice is perceived by advisees prompts an unheeded influence flowing in the advice interaction: advice-giving behaviors may also be delicately shaped by advisees. It is common that individuals preconceive their own opinions before seeking advice (Soll et al., 2022; Yaniv et al., 2009; Yaniv & Choshen-Hillel, 2012). For instance, an individual may be approached by a friend deciding between a red tie and a black tie, with asking ‘I think the red one suits my shirt better. What do you think?’. Even if the individual may personally prefer black, the individual is likely to align toward the friend’s opinions, and advise on the red tie eventually, which we refer to as the “*alignment bias*” in this paper. This subtle yet pervasive bias potentially contributes to opinion homogenization and misinformation dissemination. Despite extensive investigations into biases in the utilization of information from others (Kappes et al., 2020; Yaniv et al., 2009; Yaniv & Kleinberger, 2000), the biases in information sharing, a fundamental way we influence others, are largely unexplored.

On one hand, the phenomenon of alignment bias may be driven by the informational intention, which involves advisors’ optimal utilization of advisees’ opinions to provide accurate advice (Jonas et al., 2005). Early research has indicated that individuals’ sense of responsibility enhances when they are entrusted with providing advice to others (Kray & Gonzalez, 1999), prompting them to engage in more comprehensive and objective information assessments before offering advice compared to making decisions for oneself (Jonas & Frey, 2003; Kray & Gonzalez, 1999; Kray, 2000). This intention also consistent with advisors’ interests, as high-quality advice enhances professional reputation and leading to increased approvals (Van Swol, 2011). By this assumption, advisors may exhibit an exclusive stronger inclination to aligning with accurate opinions, thereby enhancing the accuracy of their advice ultimately.

On the other hand, the alignment towards advisees’ opinions may also stem from the intention to mitigate normative pressures (Asch, 1955, 1956; Cialdini & Goldstein, 2004). Misaligning with others’ opinions is inherently associated with social tensions and conflicts (Klucharev et al., 2009; Shamay-Tsoory et al., 2019), as it implies disagreeing others’ opinions and indicating others’ fallacies. To resolve misalignments, individuals incline to conform their opinions to those of others (Asch, 1956), even when others’ opinions are evidently incorrect (Asch, 1955). Additionally, aligning with others can also be a means of evading decision-making responsibility (Mahmoodi et al., 2015; Zein et al., 2019), a type of pressure related with misleading. In advice-giving contexts, individuals have also been observed to conform with their competitors’ opinions, to avoid the risk of being misadvising (Zaatri et al., 2022). Under this assumption, advisors could pursue conformity with advisees’ opinions invariably, even if it is wrong.

Advances in the nascent field of advice giving suggest a strategic-intention account: alignment can be a strategy to gain acceptance and avoid rejection from advisees. A wide spectrum of literature has presented that people prefer to use advice or information that aligns with or resemble their pre-existing opinions (Kappes et al., 2020; Soll et al., 2022; Yaniv & Milyavsky, 2007; Zhang & Gläscher, 2020). This general preference, serving as a feasible shortcut to the acceptance of advice, can be widely identified based on our abundant experiences of advising. In support of this, humans are experts in identifying attributes linked with latent rewards through iterative learning (Hackel & Kalkstein, 2023). For advisors, having advice accepted is specifically rewarding, as it signifies success in influencing others and gaining prestige (Hertz et al., 2020; Mobbs et al., 2015); while rejection is aversive, representing overt disbelief (Belkin & Kong, 2018; Blunden et al., 2019).To compete for acceptance, advisors have been observed to tailor their advising strategies to advisees’ heuristics deliberately (Hertz et al., 2017; Kurvers et al., 2021). By this account, advisors would adapt acceptance-directed strategies upon disclosure of advisees’ feedback on advice, while opting for opinion conformity when feedback is concealed. In summary, the intentional mechanisms underlying the alignment bias remains unclear and necessitate direct investigation.

In this work, we systematically examined the above-mentioned hypotheses on the potential intentions contributing to the alignment bias. In Study 1(*n* = 73) and Study 2 (*n* = 62), participants were tasked with providing advice in isolation and after observing an alleged advisee’s opinions in two separated sessions sequentially. Advisees’ opinions were manipulated to be incongruent with half of participants’ independently provided advice (either correct or incorrect), to test whether the opinion conformity with advisees or advice accuracy was prioritized and yielded the alignment bias. In Study 3 (*n* = 111) and Study 4 (*n* = 100), immediate feedback on advice from advisees (acceptance or rejection) were disclosed to the participants. Administered within probabilistic reversal structures (Gläscher et al., 2009), advisees’ acceptance and rejection on advice characterized their fluctuating preferences for either aligned or misaligned advice. Leveraging computational models, we examined whether the alignment bias was driven by individuals’ intention to mitigate normative pressures, positing that advisors would align with advisees’ opinions invariably; or to optimize advising outcomes, positing that advisors would adapt acceptance-directed strategies through iterative learning of the feedback. By unveiling the intrinsic nature of the alignment bias, our studies offered a novel insight into the complexity of the advisee-advisor interactions.

**Study 1**

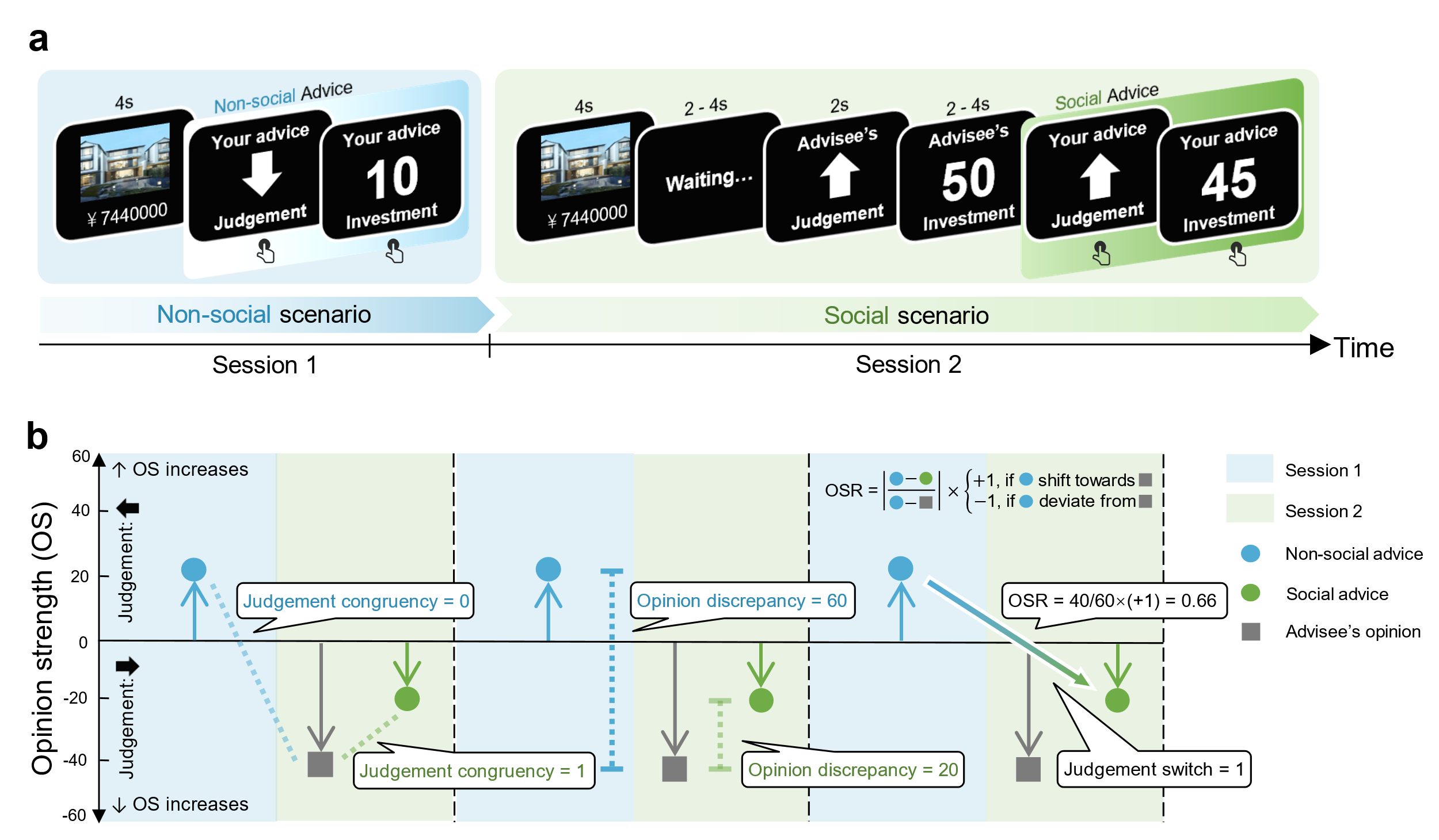
***Methods***

***Participants.*** We determined the sample size through a power analysis conducted using G\*Power 3.1. To detect a small to medium true effect (Cohen’s *d* = 0.4) for the alignment bias (i.e., mean differences in the advice provided in social vs. non-social scenario), with a Type I error of 0.05 and a power level of 0.85, a sample of 59 participants were required. We recruited 80 participants to ensure an adequate sample size after potential participant exclusions. After excluding participants (*n* = 7) not passing attention checks, 73 participants (mean age: 22.23 years, range: 18-29 years; 42 females) were included in the final analyses.

***Experimental Task.*** The advice-giving task was adapted from a risk investment game (Kappes et al., 2020), where players make judgements on the prices of estates properties, and place wager on judgements to maximize their incentives (players win their wagers on accurate judgements as incentive). Before the experiment, participants were informed that they would play the role of advisors throughout the task. Their responses were conveyed as advice to an online participant played the role of advisee, who decided whether to take advice and made final decisions. An evaluation-based incentive structure was employed to simulate typical advisory circumstances (e.g., advisors’ benefits closely linked with advisees’ satisfaction or approval). Participants were informed that their payments were partially determined by the advisees given the subjective evaluations of their advice. For each trial, participants initially provided a binary judgement about the price of an estate property (e.g., ‘Is the asking price for this property higher or lower than 5,000,000 RMB?’) and subsequently invested in the judgement (i.e., post-decision wagering, a measurement of opinion strength, ranging from 1 to 60 tokens).

Participants provided advice in both non-social and social scenarios sequentially (**Fig. 1a**). In the non-social scenario (session 1), participants gave advice in isolation right after each stimulus (an estate and its possible price) was displayed (i.e., advising without observing the opinion from the advisee). In the social scenario (session 2), participants observed all the stimuli again and provided advice after the advisee proposed their opinions (both judgements and investments) on the stimuli, except that for 20% of trials advisees’ opinions were masked (identical to the non-social scenario). Note that, we fixed the order of session 1 and 2 (instead of counterbalancing) for two reasons: (1) participant data from session 1 were required to generate the alleged advisees’ opinions in session 2; (2) having seen advisees’ opinions in session 2 could potentially jeopardize the independence of non-social advice given in session 1. Both would hamper the experimental design and clear interpretation of the results.

Unbeknownst to participants, advisees’ opinions were all preprogrammed. Advisees’ judgements were generated according to participant’s non-social judgements in session 1, ensuring that in half of the trials wherein advisees’ opinions were presented, advisees held incongruent judgements with the advice given by participants in session 1 (the other half of the trials were hence congruent). Advisees’ investments were generated from normal distributions, *Normal*(10,10) or *Normal*(50,10), distinguish their risk-avoiding or risk-seeking investment preferences serving as a between-subject variable.



**Fig. 1. Experimental task in Study 1 and conceptual schematics of key behavioral measurements.** (a) Task design in Study 1. (b) Conceptualschematics of key behavioral measurements on the alignment bias: judgement congruency, opinion discrepancy, opinion shift rate (OSR), and judgement switch.

***Stimuli.*** To create the stimuli, 66 real estate properties were selected from a domestic estate trading website. The stimulus price was either 20% higher or lower than the real price of a property. These stimuli were randomly divided into 3 subsets and assigned to different conditions (advisee’s judgement was congruent, incongruent with participants’ non-social judgement, or was not presented) with counterbalance.

A pretest was conducted on the stimuli subsets (see **Supplementary Text 1** for details). No significant differences were detected between these stimuli subsets in both judgement confidence (*F*(2,28) = 0.81, *p* = .45) and familiarity (*F*(2,28) = 0.88, *p* = .42).

***Procedures.***After consent to participation, participants were required to read the task instructions and underwent a brief quiz to ensure they comprehensively understood the task requirements. Following this, participants engaged in a two-session advice-giving task. Upon completion of the task, participants received monetary compensation based on the alleged evaluation on their advice and were debriefed about the artificial manipulations of the alleged online advisees.

***Design and variables.*** We adopted a within-subject design in this study to examine the alignment bias, by assessing the inter-session differences between the advice given in non-social (session 1) and social (session 2) scenarios.

To comprehensively depict alignment bias, we computed multiple key measurements in the trials where advisees’ opinions were presented—judgment congruency and opinion discrepancy (contrasting advice with respect to advisees’ opinions presented in the social scenario)—using the non-social scenario as a baseline (**Fig. 1b**). To illustrate to what extent participants exhibited alignment bias on a fine-grained scale, opinion shift rate (OSR) was computed to quantify the ratio extent to which advice shifts towards (OSR > 0) or deviates from (OSR < 0) the advisees’ opinions (**Fig. 1b**).

To investigate what pursuit (opinion conformity vs. advice accuracy) primarily leads to alignment bias, we examined the impact of judgment congruency (in the non-social scenario) and advisees’ judgment accuracy on judgment switch from session 1 to session 2 (**Fig. 1b**). If opinion conformity was prioritized, participants would globally demonstrate an increased judgement switch as they found their initial opinions (i.e., non-social advice in session 1) incongruent with advisees’ judgments regardless of its accuracy. However, if participants prioritized advice accuracy, they would demonstrate a selective propensity to switch to advisees’ accurate but incongruent opinions, yet refrain from being congruent with advisees’ inaccurate opinions.

***Data exclusions and statistical analyses.*** Before conducting data analyses, we excluded trials where the value of the examined variable was missing, or where the value exceeded the constraint scales (e.g., investments value outside the range of 1-60). Under the consideration of the fractional property, we incorporated additional exclusion criteria for OSR (**Supplementary Text 2**).

After data exclusions, we constructed linear mixed-effects model to conduct statistical analysis using the *lme4* package (Bates et al., 2015) in R 4.3.1. Post-hoc comparisons were performed using the *emmeans* package (Lenth, 2019). Participants and stimuli were considered as random factors for all main effects and interactions. For each analysis, we constructed 3 candidate models with different assumptions on the random effects: (1) random intercepts; (2) random slopes; and (3) random intercepts and random slopes. The best-fitting model was chosen to present the results of effect estimations. All variables were mean-centered. For statistical analysis involving only one observation per participant (e.g., individual-level computational parameters), we employed conventional statistical approaches (e.g., ANOVA) using Jamovi computer software (The jamovi project, 2024).

**Results**

***Advisors inclined to align their advice towards advisees’ opinions***

Advisors demonstrated alignment bias towards advisees’ opinions (**Fig. 2a**), as evidenced by the increased judgement congruency with the advisees’ opinions in the social compared to non-social scenario (*b* = 0.64, *z* = 8.14, 95% CI = [0.48, 0.79], *p* < .001; **Fig. 2b**). The alignment bias was also evident on a fine-grained scale, as indicated by the reduced opinion discrepancy with the advisees (*b* = -12.61, *t* = -15.00, 95% CI = [-14.26, -10.96], *p* < .001; **Fig. 2c**). The calculation of Opinion Shift Rate (OSR) allowed us to parse the to what extent did participants shift towards advisees’ opinion that contributed to the decreased opinion discrepancy. The analysis of OSR revealed that advisors shifted towards (OSR > 0) advisees’ opinions in most of the trials (**Fig. 2d**). Notably, advisors frequently over-weighed advisees’ opinions than their own (OSR > 0.5) and even excessively agreed with advisees’ opinions (OSR > 1). Moreover, the investment magnitude of advisors was also shown to be gravitated towards advisees’ investment preferences: In trials where advisees’ opinions were masked, the investment magnitude of advisors decreased or increased when they were interacting with risk-avoiding advisee (*b* = -10.27, *t* = -11.64, 95% CI = [-12.00, -8.54], *p* < .001) or risk-seeking advisee (*b* = 3.38, *t* = 3.78, 95% CI = [1.62, 5.14], *p* < .001), respectively.

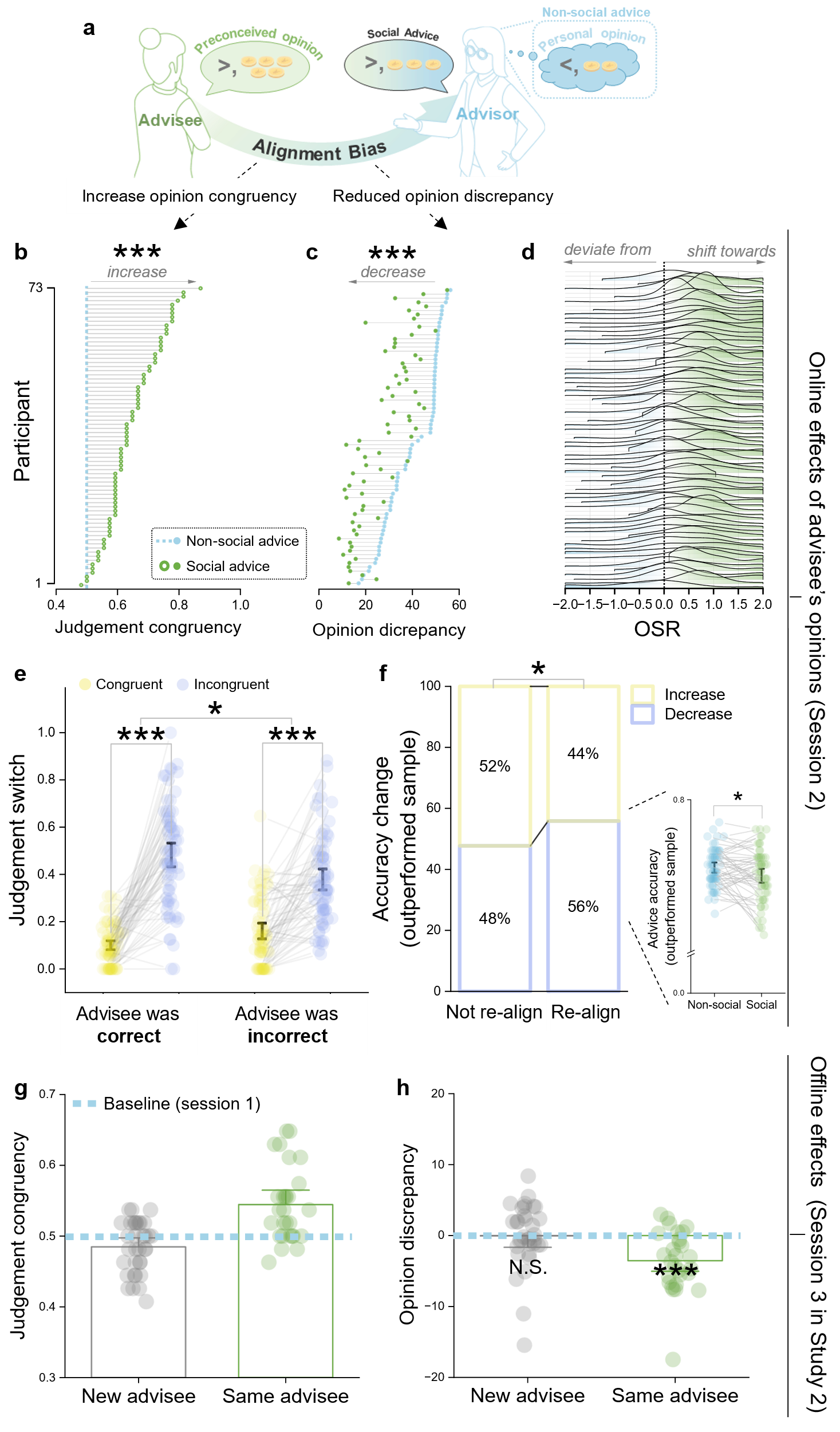
***Advisors prioritized opinion conformity with advisees over advice accuracy***

Next, we investigated whether the alignment bias primarily stemmed from advisors’ conformity to, or optimal utilizations of advisees’ opinions to improve advice accuracy (**Fig. 2e**). More specifically, by analyzing the impacts of judgment congruency and advisees’ accuracy on participants’ judgment switch in session 2, we tested whether participants prioritized resolving misalignment with their advisees (i.e., to switch) or maintaining their advice accuracy (i.e., not to switch) when these two aims were in conflict (i.e., advisees hold an inaccurate opinion incongruent with participants’).

The results showed that advisors switched their judgments more frequently when their original judgments (in the non-social scenario) were incongruent, as opposed to congruent, with the advisee’s judgement (*b* = 1.97, *z* = 13.52, 95% CI = [1.69, 2.26], *p* < .001), indicating an inclination to realign with advisees’ judgements. The results from interaction (*b* = 1.28, *z* = 2.24, 95% CI = [0.16, 2.39], *p* = .03) revealed that, participants re-aligned their judgements more frequently when advisees’ judgements were correct (*b* = 2.61, *z* = 8.40, 95% CI = [2.00, 3.22], *p* < .001), indicating participants’ awareness of the accuracy of judgements. However, when advisees’ judgements were incorrect, the inclination to realign with advisees’ judgements persisted (*b* = 1.33, *z* = 4.05, 95% CI = [0.69, 1.98], *p* < .001), echoing with the hypothesis on the prioritization for the pursuit of opinion conformity over advice accuracy (see ***Design and variables.*** for details).

Despite the observed propensity of re-aligning with advisees’ inaccurate opinions, it could be argued that participants were merely persuaded by advisees, thereby largely switched their initial judgments. If so, according to the confidence–persuasiveness effect (Pulford et al., 2018; Sah et al., 2013), advisees’ confidence in their opinions that are incongruent with participants’ could be indicative of participants’ judgment re-alignment. To test this conjecture, we conducted mixed-effect logistic regressions to examine the impact of advisees’ confidence (measured by their investment magnitude), controlling for advisors’ confidence in their initial judgments. Advisees’ confidence merely had marginal influence on advisors’ judgement re-alignment (*b* = -7.4 × 10-3, *z* = -1.78, 95% CI = [-0.02, 0.0007], *p* = .08). This was true both for trials where advisees’ judgements were correct (*b* = -8.0× 10-3, *z* = -1.48, 95% CI = [-0.02, 0.003], *p* = .14) and incorrect (*b* = -9.3× 10-3, *z* = -1.75, 95% CI = [-0.02, 0.001], *p* = .08). The limited impact of advisees’ confidence was further confirmed by comparing participants’ overall judgement re-alignment rare when they were interacting with risk-avoiding (advisees invested conservatively, i.e., with low confidence) vs. risk-seeking (advisees invested aggressively, i.e., with high confidence) advisees (*b* = -0.30, *z* = -1.39, 95% CI = [-0.12, 0.72], *p* = .17).

The prioritization of opinion conformity over advice accuracy was also evident on the decrease of advice accuracy (**Fig. 2f**). Though we did not found significant declines in advice accuracy across the whole sample (*b* = -0.02, *z* = -0.34, 95% CI = [-0.12, 0.08], *p* = .74), participants who initially (i.e., in the non-social scenario) demonstrated higher accuracy in judgement than the advisee (*n* = 57, out of 73, referred to the outperformed sample) showed a significant decrease in advice accuracy in session 2 (*b* = -0.13, *z* = -2.07, 95% CI = [-0.25, -0.01], *p* = .04). Additionally, a positive relationship between accuracy decreases and judgements re-alignments in the outperformed sample was observed (*b* = 0.18, *z* = 2.09, 95% CI = [0.01, 0.36], *p* = .04), further suggesting that advisors prioritized opinion conformity with advisees, even at cost of sacrificing advice accuracy.



**Fig. 2. Advisors tended to align their advice with advisee’s opinions.** (a)-(d) The alignment bias in social advice giving. (a) The schematic of alignment bias in social advice giving. (b) Judgement congruency and, (c) opinion discrepancy with advisee’s opinion in non-social vs. social advice-giving scenario of each participant. For illustration purpose, we sorted the data according to the judgement congruency in social scenario and opinion discrepancy in non-social scenario, respectively. (d) The frequency of opinion shift rate (OSR) among all trials of each participant. The OSRs exceeding the scale were consolidated into the terminal columns. (e)-(f) Advisors prioritized opinion conformity with advisees. (e) Advisors globally tended to conform with advisees’ opinions—switched their initial judgements (in session 1) when advisees’ judgement was incongruent (compared to congruent) with theirs—regardless of its accuracy. (f) The pursuit of opinion conformity counteracted advice accuracy in the outperformed sample (right panel). Advice decreases positively correlated with judgement re-alignment in the outperformed sample (left panel). (g)-(h) Advisors retained on alignment only when interacting with the same advisees. (g) Judgement congruency and, (h) opinion discrepancy with the opinions of the advisee in session 2. Dots represent individual participants’ data; error bars represent the 95% confidence intervals (CIs). Online effects of advisees’ opinions refer to all the effects in session 2 (i.e., with observations on these opinions); offline effects of advisees’ opinions refer to all the effects in session 3 (i.e., without but after observations on these opinions).

**Study 2**

Evidence from Study 1 provided pioneering insights into the alignment bias and showed that this bias is primarily yielded by opinion conformity. However, several important questions remained to be investigated. First, one might argue that, the observation of opinion conformity might not be a deliberate pursuit or an outcome of persuasion, but rather could be a mere result of opinion contagion (Suzuki et al., 2016) or the anchoring effect (Tversky & Kahneman, 1974): Advisees’ opinions act as anchors or the source of contagion, leading advisors to unconsciously assimilate their advice towards them. Second, will advisors persist to pursue opinion conformity even when external rewards are explicitly introduced to encourage advice accuracy?

To address these questions, we introduced: (1) a performance-based incentive structure. This structure linked participants’ bonus to their advice quality, sufficiently motivating them to prioritize advice accuracy; (2) an additional session (session 3) after session 2, where participants gave advice in isolation, but were either informed they would give advice to a new advisee, or the same advisee as in session 2. This allowed us to examine the deliberativeness of the pursuit of opinion conformity, assuming that the offline effects of advisees’ opinions would be contingent upon interaction with the same advisee.

***Methods***

***Participants.*** We recruited 64 participants, based on the same power analysis from Study 1, which required 59 participants. After excluding participants (*n* = 2) not passing attention checks, 62 participants (mean age: 23.03 years, range: 18-32 years; 39 females) were included in the final analysis.

***Experimental Task.*** In Study 2, the task comprised 3 sessions. Session 1 and 2 were consistent with Study 1. In session 3, participants again provided advice in isolation (i.e., without observation of advisees’ opinions). In this session, we introduced a between-subject variable: participants were either informed that their advice would be delivered to a new advisee vs. the same advisee from session 2 that they had interacted with. The purpose of this manipulation was to examine the deliberativeness of opinion conformity, by testing whether advisors persisted in aligning with previous advisees’ opinions exclusively in the context that they were aware it would be effective.

Different from Study 1, participants’ payments were mostly determined by the actual performance of their advice (= accuracy × investment magnitude, where accuracy was 1 for accurate judgement and -1 for in accurate judgement).

***Post-task survey***

Participants were asked to report on their belief of playing the role of advisors by rating on their level of agreement with the statements below on a scale of 1 (= “strongly disagree”) to 9 (= “strongly agree”):

1. ‘I am aware that I played the role of an advisor.’
2. ‘My advice was provided independently.’
3. ‘My advice benefited my advisee.’
4. ‘I made my advice after careful consideration.’
5. ‘I resorted to my advisee’s opinions to provide advice.’

Participants who rated lower than 5 on the first four statements or higher than 5 on the fifth statements were excluded from the final analysis, as this indicated a miscomprehension of the instructions.

Finally, upon completion of the experiments, participants were instructed to share their subjective experiences during the experiments. Those who indicated that they perceived the advisees as fictitious were also excluded.

**Results of Study 2**

***Replication of the dominant role of opinion conformity in alignment bias.***

Consistent with Study 1, participants exhibited alignment bias in social advice-giving scenario, as evidenced by increased judgment congruency (*b* = 0.49, *z* = 6.20, 95% CI = [0.34, 0.65], *p* < .001), decreased opinion discrepancy (*b* = -8.15, *t* = -9.36, 95% CI = [-9.85, -6.44], *p* < .001), and investment magnitude gravitating towards advisees’ investment preferences, though we only observed a trend when interacted with risk-seeking advisees (risk-avoiding advisee:*b* = -5.51, *t* = -6.04, 95% CI = [-7.30, -3.72], *p* < .001; risk-seeking advisee: *b* = 1.44, *t* = 1.48, 95% CI = [-3.36, 0.48], *p* = .14). The generalizability of alignment bias was further demonstrated by advisors shifting towards (OSR > 0) advisees’ opinions in most trials (**Supplementary Fig. 1**).

Consistent with Study 1, participants exhibited an inclination to re-align with advisees’ opinions (*b* = 1.46, *z* = 8.89, 95% CI = [1.14, 1.78], *p* < .001). However, a larger interaction effect of judgement congruency × advisees’ accuracy (*b* = 1.59, *z* = 2.81, 95% CI = [0.48, 2.69], *p* = .005) was found compared to Study 1, suggesting participants’ enhanced concern on advice accuracy. Notably, the inclination of re-alignment towards advisees’ judgements persisted both when it was correct (*b* = 2.25, *z* = 7.07, 95% CI = [0.01, 1.32], *p* < .001) and when it was incorrect (*b* = 0.67, *z* = 2.00, 95% CI = [0.01, 1.32], *p* = .046). The enhanced concern for accuracy was also evident on the diminished positive association between accuracy decrease and judgement re-alignment in the outperformed sample (*n* = 49, out of 62) (*b* = -0.16, *z* = -1.77, 95% CI = [-0.33, 0.02], *p* = .08), however, it did not revert into a negative association (i.e., re-alignment resulted in accuracy increase). Additionally, we also verified that the manifestation of opinion conformity was not an outcome of being persuaded by advisees by testing the impact of advisees’ confidence on judgement re-alignment (*b* = -8.0 × 10-3, *z* = -1.67, 95% CI = [-0.02, 0.003], *p* = .10).

***Advisors deliberatively conformed to but not being permeated by advisees’ opinions.***

Next, we investigated whether opinion conformity was deliberatively pursued. By this account, participants were hypothesized to become impervious to alignment bias (towards the opinions from advisees in session 2) when they recognized that their advice would be provided to a new advisee (new advisee condition); whereas continued to exhibit alignment bias when participants believed their advice would be provided to the same advisees from session 2 (same advisee condition). In contrast, if the alignment bias was merely a result of being influenced by advisees’ opinions, there would not be any differences between these conditions.

Revealed by the offline effects of advisees’ opinions on advice, participants in the new advisee condition no longer showed alignment towards the advisee (who expressed their opinion in session 2), as evidenced by their judgment congruency (*b* = -0.07, *z* = -0.79, 95% CI = [-0.23, 0.10], *p* = .43; **Fig. 2g**) and opinion discrepancy (*b* = -0.06, *t* = -0.06, 95% CI = [-2.00, 1.89], *p* = .95; **Fig. 2h**) not being significantly different from the baseline level (session 1). However, participants in the same advisee condition demonstrate a persistent alignment bias, manifested by higher judgment congruency (*b* = 0.18, *z* = 1.97, 95% CI = [0.001, 0.36], *p* = .049; **Fig. 2g**) and lower opinion discrepancy (*b* = -3.56, *t* = -3.95, 95% CI = [-5.33, -1.80], *p* < .001; **Fig. 2h**) in comparison to the baseline level (i.e., session 1). Consequently, our data directly evidenced the deliberate nature of the alignment bias, which could not be rather explained by opinion persuasion or permeation effects.

**Study 3**

In Study 2, the deliberateness and robustness of opinion conformity with advisees’ opinions were illustrated. However, the intention behind this conformity were still to be investigated: Is the conformity a strategic accommodation to advisees’ preference for aligned advice? Or it is driven by the normative intentions to reconcile coherence or avoid responsibility?

To address these hypotheses, immediate feedback on advice (acceptance/rejection) from advisees was incorporated into session 2. The feedback was administered within a probabilistic reversal learning structure (Gläscher et al., 2009). In this setup, the likelihood of aligned or misaligned advice being accepted alternated periodically during the task, reflecting as advisees’ fluctuating preferences for either advice types. Based on the preference-accommodation hypothesis, participants would exhibit adaptive advising tendencies tailoring advisees’ preferences for a distinct type of advice (aligned/misaligned with their opinions). On the contrary, based on the normative-intention hypothesis, participants were hypothesized to pursue invariable alignment with advisees’ opinions.

***Methods***

***Participants.***

In Study 3, advisor’s adaptation to the preference of advisees were examined. To detect a small to medium true effect (Cohen’s *d* = 0.3) for the mean difference in judgement alignment between the conditions that advisee preferred aligned vs. misaligned advice, with a Type I error of 0.05 and a power level of 0.85, we determined that 102 participants were necessary. Thus, we recruited 120 participants to ensure sufficient statistical power to detect true effects after potential exclusions. The criteria for data exclusion were consistent with Study 2, with participants failing attention checks (*n* = 5) and demonstrated misunderstandings of the task in the post-task survey (*n* = 4) being excluded. As a result, the final sample consisted of 111 participants (mean age: 21.90 years, range: 18-30 years; 69 females).

***Experimental Task.***

The experiment settings were mainly consistent with previous studies, except that feedback on advice was immediately provided after advisors submitted their advice in each trial during session 2. Session 2 was truncated into three phases, throughout which advisees’ preferences fluctuated between phases (**Fig. 3a**). In each phase, advisees’ preferences for a specific type of advice (aligned/misaligned advice, we referred this to ‘phased preference’) was presented by a higher probability of acceptance on that, *P*(accept | advice type), compared to the other (**Fig. 3b**). For example, as advisees exhibited phased preference for aligned advice, aligned advice would be accepted at a higher rate (70%) than misaligned advice (30%), and vice versa. To emulate typical everyday advice-interaction, mild advisees were simulated, who moderately accepted advice throughout the task, with an overall acceptance rate of 50%. Therefore, the acceptance rate for each type of advice was distributed across three levels: 30%, 50%, and 70% (**Fig. 3b**).

***Computational modeling.***

We employed a computational modeling approach to parse the trial-by-trial dynamics of the advice-giving behaviors within a feedback-presented context. Four categories of computational models were constructed, to quantificationally hypothesize the intentions that potentially involved in social advice giving:

The ‘baseline’ category (M1) posited that, advisors solely depended on their independent opinions to provide advice and did not consider any social information. The ‘opinion alignment’ category (M2, M3, M4) proposed that, advisors gravitated towards advisees’ opinions invariably when giving advice regardless of the feedback from advisees, indicating a major intention to maintain coherence with advisees or avoid responsibility for misadvising. The ‘preference accommodation’ category (M5, M6, and M7) assumed that, advisors sought to gain acceptance from advisees, consequently exhibiting a learning-based adaptation of their advising tendencies to advisees’ phased preferences. The ‘mixed intention’ category (M8, M9, and M10) assumed advisors accommodated to advisees’ phased preferences, while attempting to align with advisees’ opinions simultaneously.

Parameter estimation was conducted by the *Stan* package (Carpenter et al., 2017) in R 4.0.4 within the hierarchical Bayesian framework (Ahn et al., 2017; Gelman et al., 2013). The candidate models were fitted separately to the data of two distinct sub-samples (participants with risk-avoiding/risk-seeking advisee, referred to R-A/R-S, respectively) taking data homogeneity into account, which is essential for accurate parameter estimations (see **Supplementary Text 3** for details on model estimation and model selection). Model comparisons were performed by the *loo* package (Vehtari et al., 2017) to compute leave-one-out information criteria (LOOIC) for each candidate model. This approach allowed for a robust evaluation of model fit and facilitated comparison between different models (Vehtari et al., 2017).

*Baseline category (M1)*

In M1, we assumed that advisors generated judgements for advice solely from their independent opinions (measured by their non-social advice in session 1), and the independent opinions was accounted for by the value , a two-element vector of the judgement options, specifying the value of the options ‘higher’ and ‘lower’ in trial *t*:

The non-social value of the selected option in the non-social scenario was proportional to (measured by the wager on the selected option), while the value of unselected option remained 0.

where was a normalization function () scaling the magnitude of opinions strength from the wagering scale (i.e., 1-60) to 0-1.

The probability () of choosing an option was calculated using SoftMax function:

Finally, the advice judgement () was modeled by the categorical distribution:

*‘Opinion alignment’ category (M2-M4)*

This model category assumed that advisors considered their own independent opinions, while inclined to align with advisee’s trial-by-trial opinion. In the models, advisors considered both the non-social value and the social value of each advice type (i.e., aligned/misaligned with advisees’ opinions).

M2 hypothesized that advisors conformed with advisees’ opinions at a constant level (noted that, in all the subsequent models, in the trials that advisees’ opinions were not presented, the model assumptions were identical to M1), captured by a positive constant () assigned to the value of aligned advice. Thereby, the value of ‘aligned’ and ‘misaligned’ in trial was defined as below:

M3 aimed to further address the possibility that the phenomenon of opinion conformity is merely judgement switch in response to uncertainty arising from opinion incongruence. This model hypothesized that the tendency of aligning with advisee’s judgements increased with the magnitude of advisees’ opinion strength () in their judgements:

M4 was a hybrid of M2 and M3, hypothesizing that advisors inherently inclined to align with advisees’ opinions, with this alignment intensified with advisees’ confidence:

where () denoted advisor’s relative sensitivity to advisees’ confidence.

Across M2 to M4, advisors determined their judgements based on the integration of non-social and social value:

Due to was the value-vector of option ‘higher’ and ‘lower’, we mapped the into these values based on the advisee’s judgement in trial *t* (), to enable the combination:

*‘Preference accommodation’ category (M5-M7)*

The pursuit of acceptance required advisors to actively adapt their advice-giving tendencies in accordance with the varying preferences of advisees, which largely determined whether a particular type of advice would be accepted or rejected. This process required reward learning of each advice type (aligned/misaligned advice), where acceptance was perceived as reward and rejection was perceived as punishment; and the utilization of these learned values to adjust subsequent advice-giving behaviors to maximize future acceptance. Therefore, we describe this adaptive dynamic using reinforcement learning (RL) models (Ahn et al., 2017; Hackel et al., 2020; Sutton & Barto, 2018; Zhang et al., 2020), a classic and powerful framework to parse the reward-driven learning processes, speculating that individuals incorporate prediction errors (PE; difference between expected value and actual outcome of an action (Sutton & Barto, 2018)) to update expected value of an action, and posited that individuals tend to behave based on the averaged values previously associated with each option.

In M5, the updating of was defined according to the classical Rescorla-Wagner model (Sutton & Barto, 2018):

where denotes the prediction error between the outcome (1= being accepted; -1 = being rejected) and the value of the chosen advice type, while the value of advice type unchosen remained unchanged. () denoted the learning rate that captured the weight of in value updating.

Note that, the current model category hypothesized that advisors demonstrated an inclination to align with advisees’ opinions at the very beginning of the interaction (i.e., pre-existing alignment bias), described by a positive initial value that could be updated with learning (rather than being constant throughout the interaction):

M6 is constructed to serve as a contrast to M7, to test if the optimistic learning bias in aligned advice could be defined in a simpler way. Specifically, the persistence of giving aligned advice could be rather manifested as the overweight on the prior beliefs (i.e., *β*) compared to the subsequent outcomes of giving aligned advice. This could be indicated by a smaller learning rate. Thus, we differentiated the learning rate of aligned and misaligned advice:

M7 hypothesized that the optimistic learning bias could be attributed to advisors’ tendency to reinforce a belief that aligned advice leads to acceptance. In this case, larger learning rate on acceptance and the smaller learning rate on rejection when giving aligned advice (compared to misaligned advice) could be expected. Therefore, we distinguished the learning rate of each association of advice type × outcome as follows:

*‘Mixed intention’ category (M8-M10)*

This model category assumed that advisor was inclined to pursue both preference accommodation and opinions alignment. Therefore, to represent these two distinct intentions, and were constructed and integrated into the social value :

where () denoted advisor’s weight of acceptance-seeking intention. in M8, M9, and M10 were equal to in M5, M6, and M7, respectively. Given that M2 outperformed among the ‘opinion alignment’ category (M2–M4), and evidence from Study 1 and Study 2 showed that advisors pursued opinion alignment irrespective of advisees’ confidence, was based on M2’s specifications on .

***Results.***

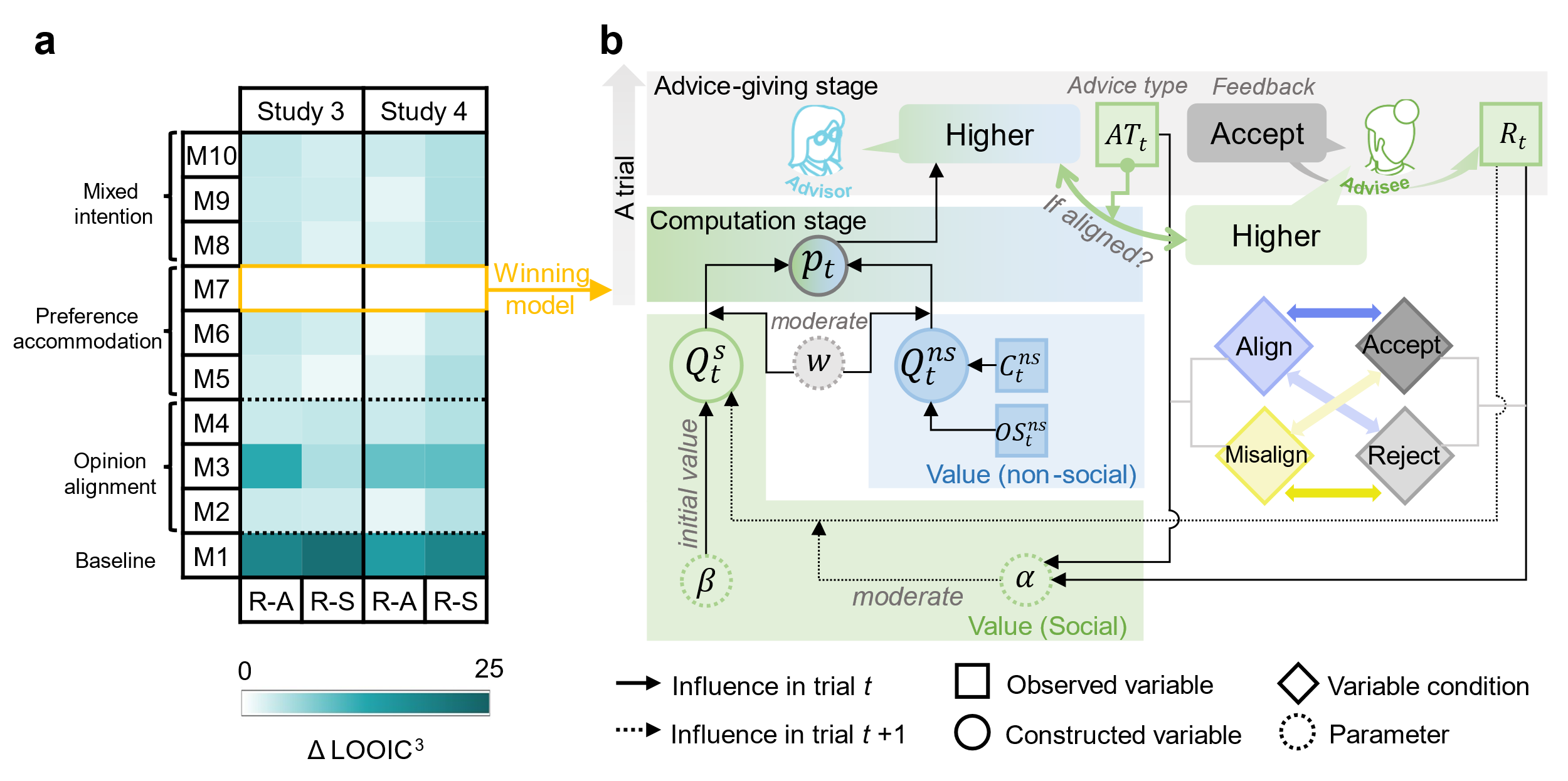
***Advisors adapted advice-giving tendencies with advisees’ phased preferences.***



**Fig. 3. Experimental tasks in Study 3 and Study 4 and advising tendencies adapted with advisees’ preferences.** (a)Experimental designs in Study 3 (*n* = 111) and Study 4 (*n* = 100). (b) Judgement alignment rate with advisee’s opinions in the trial-by-trial (left panel) and across trials (right panel) format. In the left panel, the bars represent the judgement alignment rate at latter half of each phase, while the dash lines represent the acceptance rate for aligned (in purple) or misaligned (in yellow) advice in each phase. The line charts present both true data (green lines) and the predicted data of the winning model (dark lines: the predicted data averaged across participants; shaded areas: 95% highest density intervals of the predicted data). In the right panel, the bars represent participants’ overall judgement alignment rate when advisee preferred aligned vs. misaligned advice (pooling 4 conditions in the left panel). Dots represent individual participants’ data; error bars represent the 95% CIs.

We examine whether advising tendencies adapted with the phased preferences of advisees, by comparing the extent to which participants align with advisees’ opinions between phases (in the latter stages, to ensure sufficient feedback acquisition) where advisees preferred aligned vs. phases advisees preferred misaligned advice. As shown in the results, advisors were more inclined to provide advice contradicting advisees’ opinions when advisees had phased preference for misaligned advice, as opposed to when they had phased preference for aligned advice (*b* = 0.22, *z* = 2.25, 95% CI = [0.03, 0.40], *p* = .03; **Fig. 3b**).

Although advising tendency being explicitly adjusted with advisees’ preferences, participants persisted in aligning with advisees’ opinions when they favored misaligned advice. Inspired by extensive literature on confirmation bias in belief updating leading to biased outcomes (Bellucci & Park, 2020; Biele et al., 2009), we proposed that this persistence can be ascribed to a confirmation bias on in learning from the feedback on advice—advisors overweigh the instances of acceptance and downplay the instances of rejection after aligned advice. To test this conjecture and validate the intention to gain acceptance contributing to the advising adjustments, computational models were constructed based on a range of sophisticated hypotheses regarding the advice-giving mechanisms (see **Methods.** for details).



**Fig. 4. Modeling advice-giving behaviors in contexts where advisees’ feedback was provided.** (a) Model comparisons. Model comparisonshows model evidence, measured by the Leave-One-Out Information Criteria (LOOIC). Lower LOOIC indicates better model fit after balancing model complexity. The M7 model highlighted in yellow frame had the lowest LOOIC, indicating it best and most parsimoniously described the data. R-A represents the sub-sample of participants interacted with risk-avoiding advisees; R-S represents the sub-sample of participants interacted with risk-seeking advisees. We transformed the Δ LOOIC values onto a well-proportioned scale by computing Δ LOOIC3, ensuring the legibility of the illustration. (b) The conceptual schematics of the winning model.

Evidenced by model comparisons (**Fig. 4a**): M7 (**Fig. 4b**) in the ‘preference accommodation’ category outcompeted candidate models in other categories (including ‘baseline’, ‘opinion alignment’ and ‘mixed intention’ category), supporting that the manifestation of persistence in giving aligned advice could not be attributed to a mere pursuit of opinion alignment *per se*. Furthermore, M7 outcompeted the other models in the ‘preference accommodation’ category, confirming that advisors exhibited learning biases towards different feedback received from different advice types. Posterior predictive checks (Zhang et al., 2020) on the winning model were conducted to test the congruence between true behavioral data and the simulated data at different scales: trial-wise scale (**Supplementary Table 1)**, individual-wise scale (**Supplementary Table 2**), and grand average scale across trials and individuals (**Supplementary Fig. 2**). The validity of the winning model M7 was further confirmed by linking computational parameters and the associated behavioral measurements (**Table 1**). Parameter *w* (in equation (15)) characterized the weight of social value relative to non-social value in the process of advice giving, capturing advice’s susceptibility to the feedback from advisees. It demonstrated a positive correlation with the overall judgement switch (from session 1 to session 2). Additionally, Parameter *β* (in equation (18)) captured individuals’ pre-existing alignment bias, exhibiting a positive correlation with the judgement alignment rate at the beginning of session 2.

Based on the established model validity, we further confirmed the relationship between the acceptance-seeking intention and the alignment bias. Positive correlations were observed between parameter *β* and *w* (**Table 2**), indicating a close connection between advisors’ pre-existing alignment bias and their susceptibilities for acceptance, which manifested advisors’ acceptance-intended inclinations. Similar results were observed between parameter *w* andparticipants’ initial inclination to align with advisees’ opinions (**Supplementary Table 3**).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1 |** Linking computational parameters with behavioral measurements | | | | | | | | |
| *β* ~ Judgement alignment at the beginning of interaction | | | | | | | | |
|  |  | Study 3 | | |  | Study 4 | | |
| sub-sample | ***r* (s.e.)** | | ***df*** | ***P*** |  | ***r* (s.e.)** | ***df*** | ***P*** |
| RA | 0.73 (0.06) | | 55 | < .001 \*\*\* |  | 0.80 (0.05) | 46 | < .001 \*\*\* |
| RS | 0.58 (0.09) | | 52 | < .001 \*\*\* |  | 0.59 (0.09) | 50 | < .001 \*\*\* |
| *w* ~ Judgement switch throughout the interaction | | | | | | | | |
|  |  | Study 3 | | |  | Study 4 | | |
| sub-sample | ***r* (s.e.)** | | ***df*** | ***P*** |  | ***r* (s.e.)** | ***df*** | ***P*** |
| RA | 0.82 (0.05) | | 55 | < .001 \*\*\* |  | 0.52 (0.11) | 46 | < .001 \*\*\* |
| RS | 0.76 (0.06) | | 52 | < .001 \*\*\* |  | 0.78 (0.06) | 50 | < .001 \*\*\* |

*Note*. \*\*\**P* < 0.001.The first 50% trials in 1st phase were selected to calculate the judgement alignment at the beginning of interaction. R-A represents the sub-sample of participants interacted with risk-avoiding advisees; R-S represents the sub-sample of participants interacted with risk-seeking advisees.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2 |** Linking individuals’ pre-existing alignment bias (*β*) and the acceptance-seeking intention (*w*) | | | | | | | |
|  | Study 3 | | |  | Study 4 | | |
| sub-sample | ***r* (s.e.)** | ***df*** | ***P*** |  | ***r* (s.e.)** | ***df*** | ***P*** |
| RA | 0.50 (0.10) | 55 | < .001 \*\*\* |  | 0.64 (0.09) | 46 | < .001 \*\*\* |
| RS | 0.47 (0.11) | 52 | < .001\*\*\* |  | 0.38 (0.12) | 50 | .006 \*\* |

*Note*. \*\**P* < 0.01; \*\*\**P* < 0.001. R-A represents the sub-sample of participants interacted with risk-avoiding advisees; R-S represents the sub-sample of participants interacted with risk-seeking advisees.

As predicted, the optimistic learning biases of giving aligned advice were observed (**Table 3**): the learning rate of acceptance from giving aligned advice () was higher than that of misaligned advice (). Conversely, the learning rate of rejection from giving aligned advice () was lower than that of misaligned advice (). Consequently, the optimistic learning bias () could be gauged by the amalgamation of learning rate disparities in acceptance () and rejection (), thereby yielding . Subsequently, the relationship between optimistic learning bias () and the rigidity of providing aligned advice conflicting advisees’ preference (for misaligned advice) were validated (**Table 4**). The rigidity was measured by comparing the judgment alignment rate of latter trials in phases where advisees demonstrated a preference for misaligned advice to the initial trials of session 2 that captured participants’ pre-existing alignment bias.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 3 |** The optimistic learning biases of giving aligned advice | | | | | |
| Study 3 | | | | | |
| sub-sample | > | |  | < | |
| ***MD* (s.e.)** | ***P*** |  | ***MD* (s.e.)** | ***P*** |
| RA | 0.49 (0.01) | < .001\*\*\* |  | -0.19 (0.01) | < .001 \*\*\* |
| RS | 0.55 (0.01) | < .001 \*\*\* |  | -0.39 (0.02) | < .001 \*\*\* |
| Study 4 | | | | | |
| sub-sample | > | |  | < | |
| ***MD* (s.e.)** | ***P*** |  | ***MD* (s.e.)** | ***P*** |
| RA | 0.48 (0.01) | < .001 \*\*\* |  | -0.52 (0.01) | < .001 \*\*\* |
| RS | 0.43 (0.01) | < .001 \*\*\* |  | -0.69 (0.01) | < .001 \*\*\* |

*Note*. \*\*\**P* < 0.001. R-A represents the sub-sample of participants interacted with risk-avoiding advisees; R-S represents the sub-sample of participants interacted with risk-seeking advisees.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4 |** The relationship between optimistic learning bias () and the persistence of providing aligned advice | | | | | | | |
|  | Study 3 | | |  | Study 4 | | |
| sub-sample | ***r* (s.e.)** | ***df*** | ***P*** |  | ***r* (s.e.)** | ***df*** | ***P*** |
| RA | 0.44 (0.11) | 55 | < .001 \*\*\* |  | 0.28 (0.13) | 46 | .056✝ |
| RS | 0.53 (0.10) | 52 | < .001\*\*\* |  | 0.28 (0.13) | 50 | .047 \* |

*Note*. ✝*P* < 0.1; \**P* < 0.05; \*\*\**P* < 0.001. R-A represents the sub-sample of participants interacted with risk-avoiding advisees; R-S represents the sub-sample of participants interacted with risk-seeking advisees.

**Study 4**

Extensive research has demonstrated that advisees largely adhere to their initial opinions and disregard advice (i.e., egocentric bias, Yaniv, 2004; Yaniv & Kleinberger, 2000). Validating the vigilance to the feedback from those who are insusceptible to advice is crucial for understanding how individuals gain acceptance in a more generalized context. Therefore, in Study 4, we continued to examine whether the intention of gaining acceptance persisted even when acceptance was scant. Furthermore, by comparing the tendency of advice alignment in contexts where advice being more responsible for advisees’ final decision (i.e., advice being accepted more, Study 3) to contexts where advice being less responsible (Study 4), we could further elucidate whether aligning towards advisees’ opinions was employed to evade responsibility of giving advice.

***Methods***

***Participants.***

We determined a sample size (*n* =112) that resembled that of Study 3. Consistent with Study 3, participants who failed attention checks (*n* = 7) or demonstrated misunderstandings of the task in the post-task survey were excluded (*n* = 5). Consequently, 100 participants (mean age: 20.49 years, range: 18-27 years; 71 females) contributed to usable dataset included in the subsequent analysis.

***Experimental Task.***

The experiment settings were largely consistent with Study 3 (**Fig. 3a**), except that insusceptible advisees were simulated in session 2, who displayed a high resistance to taking advice, rejecting advice in the majority of time (overall acceptance rate = 30%, the acceptance rate for each type of advice was distributed across three levels: 10%, 30%, and 50%, **Fig. 3c**).

***Results.***

***Advisors persisted in preference accommodation even when being frequently rejected.***

Similar to Study 3, the winning model (M7) outperformed all other candidate models in both sub-samples (**Fig. 4a**), and accurately captured the true behavioral data in Study 4 (**Fig. 3c**, **Supplementary Table 1-2**, **Supplementary Fig. 3**, and **Table 1**). Participants demonstrated a marginal adjustment in their advising tendencies responding to advisees’ various phased preferences (*b* = 0.30, *z* = 1.75, 95% CI = [-0.04, 0.64], *p* = .08, **Fig. 3c**), which could be attributed to a generally larger optimistic learning bias in Study 4 (compared to Study 3) (**Table 3**). Furthermore, positive correlations between the optimistic learning bias () and the rigidity of providing aligned advice were observed (**Table 4**). These results further support the notion that advisors persisted to strive for acceptance by adapting to advisees’ phased preferences, even when the acceptance of advice was scant.

***Alignment bias was not advisors’ attempt to evade responsibility.***

Despite evidence from computational modeling already claimed the pursuit of acceptance rather than opinion alignment *per se.* (to achieve coherence or avoid responsibility) serving as the dominant intention behind alignment bias, we further leveraged behavioral evidence to reinforce this notion. Specifically, we addressed whether the intention to avoid responsibility contributed to the alignment bias by comparing the data from Study 3 and Study 4: If advisors inclined to avoid responsibility for advising, they would exhibit a higher degree of alignment with advisees in the circumstances where advisees generally more acceptive of advice.

Comparing data from Study 4, where advisors were less responsible of advisees’ judgements, participants did not exhibit a greater degree of judgement alignment with advisees’ opinions in Study 3, where advisors were more responsible of advisees’ judgements (*b* = 0.02, *z* = 0.65,95% CI = [-0.05, 0.10], *p* = .52). In line with this, OSR, quantifying the ratio extent to which advice aligned towards advisees’ opinions, also did not differ between Study 3 and Study 4 (*b* = 0.05, 95% CI = [-0.02, 0.13], *t* = 1.48, *p* = .14). Moreover, to mitigate negative outcomes caused by misadvising, conservative investments could be expected when advisees showed greater acceptance. However, advisors invested more aggressively in Study 3 compared to Study 4 during session 2 (*b* = 2.86, 95% CI = [0.04, 5.68], *t* = 1.99, *p* = .048), contradicting with this posit.

**General Discussion**

Advice interaction is commonly perceived as a uni-direction transmission of social influence from advisors to advisees (Bonaccio & Dalal, 2006; Kämmer et al., 2023). The bi-directional information exchanges within this interactive process, particularly the transmission from advisees to advisors, remains largely neglected. We addressed this gap by unveiling how and why advice giving is influenced by advisees’ opinions. The alignment bias, charactering advisors’ inclination to align their advice with advisees’ initial opinions, is identified as a strategic adaptation to advisees’ preference in taking advice, with the intention of enhancing the likelihood of advice acceptance. These results offer novel insights into the reciprocal nature of social influence (Mahmoodi et al., 2018)—the individuals who influence and perceive successful influence as a form of reward may, therefore, be susceptible to the reactions of those whom they seek to influence—within the widespread informational interactions.

In a series of four experiments, participants were tasked with providing advice to advisees to optimize their decision outcomes. Our findings demonstrated that, advisors have a robust propensity to align their advice with the opinions of advisees, regardless of they were within an evaluation-based incentive context embedded in everyday advice interactions, or a performance-based incentive context in which advice accuracy was explicitly encouraged. Deciphering this phenomenon, we found that advisors re-aligned with advisees’ opinions persistently even when they are incorrect (thus resulted in decreased advice accuracy), indicating advisors’ prioritization of opinion conformity to advisee over advice accuracy. Crucially, our data evidenced that the inclination of re-aligning with advisees’ opinions could not be explained as a consequence of persuasion (Pulford et al., 2018), or permeation (e.g., behavioral contagion, Suzuki et al., 2016; or anchoring effect, Kahneman, 1992). Instead, our data demonstrated the strategic deliberativeness of opinion conformity by illustrating the context-dependent nature of alignment bias, as it persisted only when interacting with the same advisee, while vanished once the interaction with whom ended.

In the following studies, acceptance and rejection from advisees on advice were provided, enabling participants to engage in acceptance maximization, if the alignment bias was essentially stemmed from advisors intention to gain acceptance by capitalizing on individuals’ preference for aligned advice (Kappes et al., 2020; Yaniv & Milyavsky, 2007). As predicted, advisors demonstrated a strategic pursuit of acceptance by implementing adjustments in advising tendencies (either to give aligned or misaligned advice). This was reflected on the decreased behavioral alignment towards advisees’ opinions as advisees exhibited an exclusive preference for misaligned advice, contradicting the normative pressure-avoiding hypothesis, which posited that opinion conformity would be consistently pursued disregarding the feedback from advisees. Unraveling its computational mechanisms, the advising adaptations were best captured by a pure Reinforcement Learning model positing that advisors have a singular intention to gain acceptance, but not hybrid models positing that dual intentions (to pursue both advisees’ acceptance and the opinion conformity) simultaneously underpinned these adaptations. Computational evidence further showed robust associations between advisors’ intention to gain acceptance and their pre-existing alignment bias, in line with our prediction that the alignment bias emerges from strategic adaptations to advisees’ preference.

Zooming in, our findings closely align with the research demonstrating advisors perceive acceptance from advisees as reward (Belkin & Kong, 2018; Hertz et al., 2017, 2020), and their dislike of being reject or ignored by advisees (Belkin & Kong, 2018; Blunden et al., 2019). Furthermore, our findings also echo with previous investigation on advisors’ proactive strategy utilizations to extract advisees’ acceptance and enhance personal influences (Hertz et al., 2017; Kurvers et al., 2021; Zaatri et al., 2022). Our studies suggest that advisors are learned-to-be experts on the behavioral tendencies of advisees. As shown in Hertz et al. (2017), advisors generally utilize competitive strategies to secure acceptance, i.e., advising over-confidently when ignored by the advisees and diffidently when trusted by the client. This context-dependent strategy resonates with previous research showing that advisees (van Swol & Sniezek, 2005), as well as information recipients (Pulford et al., 2018), are more attentive to advice or information conveyed with higher confidence; however, as the advisors’ confidence is found to be unindicative of accuracy, it is perceived as arrogance, and acceptance drops accordingly (Sah et al., 2013). Our studies suggest a mechanistic account on why these two veins of findings are mystically linked: the observed advising tendencies or biases could be adaptive strategies tailoring advisees’ behavioral heuristics or preferences.

Zooming out, advisors’ adaptation towards advisees’ volatile preferences provides direct evidence for the learning-based account on the emergence of the alignment bias, which is in line with the broader realm of social cognition. Our observation of advisors’ flexibility in navigating acceptance through feedback learning well corresponds with the widespread reward learning inherent in the social world. Shown by a large body of literature, individuals are adept at identifying the attributes mostly associated with reward within a specific context and engage in goal-directed learning on these attributes rapidly (i.e., model-based learning)(Behrens et al., 2008; Boorman et al., 2013). Even when the generative structures of reward are latent, individuals can trial-by-error infer which object is rewarding quickly through model-based learning (Hackel et al., 2020; Hackel & Kalkstein, 2023). Consequently, these latent learning processes proficiently direct and optimize our behaviors within interactions where potential reward embedded. Additionally, we observed that advisors exhibit a rigidity in giving aligned advice. This appears to contradict the intention to maximize acceptance. In fact, this rigidity was shown to be a result of learning biases in the advice interaction: advisors incline to learn more of acceptance and learn less of rejection following aligned advice. This observation echoes with the confirmation bias (Nickerson, 1998; Plous, 1993)—a ubiquitous phenomenon where individuals tend to disregard information contradicting their existing beliefs, which has also been observed in various interaction-based learning processes (Bellucci & Park, 2023; Biele et al., 2009). Thus, the ingrained association between giving aligned advice and gaining acceptance, can selectively amplify sensitivity to evidence that supports them and dampening sensitivity to evidence that violates them.

In summary, this study presents a comprehensive investigation into the alignment bias. At the core of this phenomenon, we depicted advisors’ acceptance-seeking intentions—advisors deliberately adapt their advice-giving behaviors to align with a general advice-taking preference for aligned advice. Our findings highlight a striking but reasonable reality: advisor, who are typically perceived as impartial and in control of influencing others, can also be swayed by social interests and thereby influenced by advisees, who are conventionally seen as the ones being influenced.

**Limitations**

The current study has several limitations. First, both advisees’ opinions and preferences were artificially manipulated. In real-life scenarios, individuals can disclose their ideas when they anticipate validation (Schultze et al., 2015) or need help with decisions. Advice-taking preference can also be swayed by individual differences and task difficulty, rather than fluctuates periodically. Second, the advice-taking behaviors, serving as the feedback on advice, were limited to binary responses (i.e., acceptance and rejection). It should be acknowledged that advice taking can also exhibit a proportional nature (i.e., weighing advice in the final decisions)(Schultze et al., 2015). However, this could entail substantial computational demands in decoding the weight of advice into advisees’ feedback on advice. Future studies should expand upon the current findings in more ecological designs embedding realistic advisee-advisor interactions where advisees provide natural responses. Third, in the circumstances that advisors are experts or highly proficient in their specialties, they may exhibit a reduced susceptibility to external influences owing to their superior knowledge. Consequently, deeper investigations into the potential impact of advisor’ expertise on the alignment bias will also be crucial for understanding whether our findings can be extended to a broader spectrum of advice-giving contexts. Furthermore, future studies can benefit from fMRI (Mobbs et al., 2015) and hyperscanning (Pan et al., 2023; Xie et al., 2023) approaches to decipher the single- (e.g., reward-related activities induced by the feedback on advice) and dual-brain mechanisms underlying the bidirectional dynamics of the advisor-advisee interactions.

**Author contributions**

X.L., and Y.P. conceived of the project and designed the experiments; X.L. implemented the experiments and collected data; X.L. analysed the data and interpreted the results under the supervision of Y.P. and L.Z.; X.L. wrote the original manuscript. Y.P., L.Z. provided critical revisions. All authors approved the final version of the manuscript for submission.

**Acknowledgements**

We thank Dr. Feng Sheng for providing valuable comments and suggestions on an earlier version of the manuscript.

**References**

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 24–57. https://doi.org/10.1162/CPSY\_a\_00002

Asch, S. E. (1955). Opinions and Social Pressure. *Scientific American*, *193*(5), 31–35. https://doi.org/10.1038/scientificamerican1155-31

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*, 1–70. https://doi.org/10.1037/h0093718

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package ‘lme4.’ *Convergence*, *12*(1), 2.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), Article 7219. https://doi.org/10.1038/nature07538

Belkin, L. Y., & Kong, D. T. (2018). Implications of advice rejection in repeated exchanges: Advisor responses and advisee gratitude expression as a buffer. *Journal of Experimental Social Psychology*, *78*, 181–194. https://doi.org/10.1016/j.jesp.2018.03.012

Bellucci, G., & Park, S. Q. (2020). Honesty biases trustworthiness impressions. *Journal of Experimental Psychology: General*, *149*, 1567–1586. https://doi.org/10.1037/xge0000730

Bellucci, G., & Park, S. Q. (2023). Neurocomputational mechanisms of biased impression formation in lonely individuals. *Communications Biology*, *6*(1), 1–13. https://doi.org/10.1038/s42003-023-05429-2

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational Models for the Combination of Advice and Individual Learning. *Cognitive Science*, *33*(2), 206–242. https://doi.org/10.1111/j.1551-6709.2009.01010.x

Blunden, H., Logg, J. M., Brooks, A. W., John, L. K., & Gino, F. (2019). Seeker beware: The interpersonal costs of ignoring advice. *Organizational Behavior and Human Decision Processes*, *150*, 83–100. https://doi.org/10.1016/j.obhdp.2018.12.002

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Boorman, E. D., O’Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The Behavioral and Neural Mechanisms Underlying the Tracking of Expertise. *Neuron*, *80*(6), 1558–1571. https://doi.org/10.1016/j.neuron.2013.10.024

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1. https://doi.org/10.18637/jss.v076.i01

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(Volume 55, 2004), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). Chapman & Hall/CRC texts in statistical science. *Bayesian Data Analysis*.

Gläscher, J., Hampton, A. N., & O’Doherty, J. P. (2009). Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. *Cerebral Cortex*, *19*(2), 483–495. https://doi.org/10.1093/cercor/bhn098

Hackel, L. M., & Kalkstein, D. A. (2023). Social Concepts Simplify Complex Reinforcement Learning. *Psychological Science*, *34*(9), 968–983. https://doi.org/10.1177/09567976231180587

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, 103948. https://doi.org/10.1016/j.jesp.2019.103948

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-02314-5

Hertz, U., Tyropoulou, E., Traberg, C., & Bahrami, B. (2020). Self-competence increases the willingness to pay for social influence. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-74857-5

Jonas, E., & Frey, D. (2003). Information search and presentation in advisor–client interactions. *Organizational Behavior and Human Decision Processes*, *91*(2), 154–168. https://doi.org/10.1016/S0749-5978(03)00059-1

Jonas, E., Schulz-Hardt, S., & Frey, D. (2005). Giving Advice or Making Decisions in Someone Else’s Place: The Influence of Impression, Defense, and Accuracy Motivation on the Search for New Information. *Personality and Social Psychology Bulletin*, *31*(7), 977–990. https://doi.org/10.1177/0146167204274095

Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes*, *51*(2), 296–312. https://doi.org/10.1016/0749-5978(92)90015-Y

Kämmer, J. E., Choshen-Hillel, S., Müller-Trede, J., Black, S. L., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*, *10*(2). https://doi.org/10.1037/dec0000199

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others’ opinion strength. *Nature Neuroscience*, *23*(1), Article 1. https://doi.org/10.1038/s41593-019-0549-2

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Kray, L., & Gonzalez, R. (1999). Differential Weighting in Choice versus Advice: I’ll do this, you do that. *Journal of Behavioral Decision Making*, *12*(3). https://doi.org/10.1002/(SICI)1099-0771(199909)12:3<207::AID-BDM322>3.0.CO;2-P

Kray, L. J. (2000). Contingent Weighting in Self-Other Decision Making. *Organizational Behavior and Human Decision Processes*, *83*(1), 82–106. https://doi.org/10.1006/obhd.2000.2903

Kurvers, R. H. J. M., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., & Bahrami, B. (2021). Strategic disinformation outperforms honesty in competition for social influence. *iScience*, *24*(12), 103505. https://doi.org/10.1016/j.isci.2021.103505

Lenth, R. (2019). *Emmeans: Estimated marginal means, aka least-squares means. R package version 1.4. 3.01*.

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-04925-y

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., Roepstorff, A., Rees, G., & Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840. https://doi.org/10.1073/pnas.1421692112

Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How Expert Advice Influences Decision Making. *PLOS ONE*, *7*(11), e49748. https://doi.org/10.1371/journal.pone.0049748

Mobbs, D., Hagan, C. C., Yu, R., Takahashi, H., FeldmanHall, O., Calder, A. J., & Dalgleish, T. (2015). Reflected glory and failure: The role of the medial prefrontal cortex and ventral striatum in self vs other relevance during advice-giving outcomes. *Social Cognitive and Affective Neuroscience*, *10*(10), 1323–1328. https://doi.org/10.1093/scan/nsv020

Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828), 20200052. https://doi.org/10.1098/rstb.2020.0052

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Pan, Y., Vinding, M. C., Zhang, L., Lundqvist, D., & Olsson, A. (2023). A Brain-To-Brain Mechanism for Social Transmission of Threat Learning. *Advanced Science*, *10*(28), 2304037. https://doi.org/10.1002/advs.202304037

Plous, S. (1993). *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.

Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, *147*(10), 1431–1444. https://doi.org/10.1037/xge0000471

Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, *121*(2), 246–255. https://doi.org/10.1016/j.obhdp.2013.02.001

Schultze, T., Rakotoarisoa, A.-F., & Stefan, S.-H. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making*, *10*(2), 144–171. https://doi.org/10.1017/S1930297500003922

Shamay-Tsoory, S. G., Saporta, N., Marton-Alper, I. Z., & Gvirts, H. Z. (2019). Herding Brains: A Core Neural Mechanism for Social Alignment. *Trends in Cognitive Sciences*, *23*(3), 174–186. https://doi.org/10.1016/j.tics.2019.01.002

Soll, J. B., Palley, A. B., & Rader, C. A. (2022). The Bad Thing About Good Advice: Understanding When and How Advice Exacerbates Overconfidence. *Management Science*, *68*(4), 2949–2969. https://doi.org/10.1287/mnsc.2021.3987

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning, second edition: An Introduction*. MIT Press.

Suzuki, S., Jensen, E. L. S., Bossaerts, P., & O’Doherty, J. P. (2016). Behavioral contagion during learning about another agent’s risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences*, *113*(14), 3755–3760. https://doi.org/10.1073/pnas.1600092113

The jamovi project. (2024). *jamovi (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org*.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Van Swol, L. M. (2011). Forecasting another’s enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120. https://doi.org/10.1016/j.ijforecast.2010.03.002

van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*(3), 443–461. https://doi.org/10.1348/014466604X17092

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Xie, E., Liu, M., Li, K., Nastase, S. A., Gao, X., & Li, X. (2023). The single- and dual-brain mechanisms underlying the adviser’s confidence expression strategy switching during influence management. *NeuroImage*, *270*, 119957. https://doi.org/10.1016/j.neuroimage.2023.119957

Yaniv, I. (2004). Receiving other people’s advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., & Choshen-Hillel, S. (2012). Exploiting the Wisdom of Others to Make Better Decisions: Suspending Judgment Reduces Egocentrism and Increases Accuracy: Exploiting the Wisdom of Others. *Journal of Behavioral Decision Making*, *25*(5), 427–434. https://doi.org/10.1002/bdm.740

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 558–563. https://doi.org/10.1037/a0014589

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. https://doi.org/10.1016/j.obhdp.2006.05.006

Zaatri, S., Aderka, I. M., & Hertz, U. (2022). Blend in or stand out: Social anxiety levels shape information-sharing strategies. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1975), 20220476. https://doi.org/10.1098/rspb.2022.0476

Zein, M. E., Bahrami, B., & Hertwig, R. (2019). Shared Responsibility in Collective Decisions. *Nature Human Behaviour*, *3*(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4

Zhang, L., & Gläscher, J. (2020). A brain network supporting social influences in human decision-making. *Science Advances*, *6*(34), eabb4159. https://doi.org/10.1126/sciadv.abb4159

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089