American Psychological Association

Journal of Experimental Psychology: General


ISSN: 0096-3445

2020, Vol. 149, No. 8, 1567–1586
http://dx.doi.org/10.1037/xge0000730

# Honesty Biases Trustworthiness Impressions

Gabriele Bellucci and Soyoung Q. Park
University of Lübeck and German Institute of Human Nutrition (DIfE), Nuthetal, Germany

Honesty is central to trust and trustworthiness. However, how a good reputation as honest person is learned and induces trustworthiness impressions is still unexplored. Developing a novel paradigm, we show in 3 consecutive experiments that individuals prefer trusting honest others who share truthful information, especially if honest behavior is consistent over time. Trust in honest others was independent of proximal benefits, and honest individuals were repaid for their honesty with higher trust in a subsequent interaction. Crucially, signs of dishonesty decreased trust but only in those who had not previously built a good reputation as honest partners. On the contrary, those who could establish a good reputation were trusted even when they were no longer trustworthy, suggesting that participants could not successfully track changes in trustworthiness of those with an established good reputation. These findings suggest that a good reputation biases the ability to learn the momentary trustworthiness of another person and impairs the updating of one's beliefs about the other's character for behavior revision. Computational modeling analyses indicate an asymmetry in information integration when interacting with honest individuals that likely underlies such learning impairment. By showing how a good reputation influences learning processes in trust-based interactions, our results provide a mechanistic account for biases in social learning and social interactions, advancing our understanding of social behaviors in particular and human cognition in general.

*Keywords:* good reputation, honesty, social decision-making, trust, trust game

*Supplemental materials:* http://dx.doi.org/10.1037/xge0000730.supp



In many circumstances in life, individuals seek advice before making a decision. Seeking advice is a form of information gathering to make more informative decisions and thereby improve the accuracy of the decision-making process (Yaniv, 2016). Better decisions are of pivotal importance to the individual because suboptimal choices may jeopardize her survival chances. Previous work has provided evidence on how individuals take advice from others. However, the effects of an adviser's reputation on an advisee's trust in the adviser's advice remain unexplored.

## Taking Advice

Previous work has shown that individuals switch from individual to social learning and integrate advice from others in different contexts and for different reasons. On the one hand, individuals are sensitive to the quality of the advice and tend to discount poor advice, suggesting that an accuracy maximization strategy plays a role in advice taking (Budescu & Rantilla, 2000; Yaniv & Kleinberger, 2000). On the other, individuals also appear to rely more on individual knowledge—a phenomenon known as *egocentric advice discounting*, and more-knowledgeable individuals (e.g., those with a better performance in an estimation task) take less advice than less-knowledgeable individuals (Yaniv, 2004, 2016; Yaniv & Kleinberger, 2000). Further, access to information about others' choices for social learning occurs especially in environments that are perceived as highly unpredictable, and advice-taking behaviors are sensitive to the character of the other, as individuals take less advice from those who do not reciprocate in advice taking (Mahmoodi, Bahrami, & Mehring, 2018; McElreath et al., 2005). These findings suggest that individuals more strongly rely on their own knowledge when uncertainty (about one's decisions or the environment) is low (Festinger, 1954) and they value the social qualities of the adviser when deciding whether to take or discount pieces of advice.

So far, previous work has focused on how and when advice from others affects individual learning. For instance, advice on the best

This article was published Online First January 9, 2020.

Gabriele Bellucci and Soyoung Q. Park, Department of Psychology I, University of Lübeck, and Department of Decision Neuroscience and Nutrition, German Institute of Human Nutrition (DIfE), Nuthetal, Germany.

All data associated with this work are publicly available (DOI: http://dx.doi.org/10.17605/OSF.IO/5KVWC).

This work was supported by German Research Foundation Grants INST 392/125-1, PA 2682/1-1, and PA 2682/2-1 and by a grant from the German Ministry of Education and Research (BMBF) and the State of Brandenburg (DZD Grant FKZ 82DZD00302). Gabriele Bellucci conceived of the idea. Gabriele Bellucci and Soyoung Q. Park designed the experiments. Gabriele Bellucci collected the data and carried out the data analysis. Gabriele Bellucci and Soyoung Q. Park wrote the manuscript. All authors approved the final version of the manuscript.

The authors declare no competing interests.

Correspondence concerning this article should be addressed to Gabriele Bellucci, who is now at Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Germany. E-mail: gbellucc@gmail.com

option in a task biases participants' initial choice patterns (Behrens, Hunt, Woolrich, & Rushworth, 2008; Pilditch & Custers, 2018; Staudinger & Buechel, 2013). However, little is known on how character traits of the adviser impact the advisee's impressions and beliefs about the adviser and the consequential advice-taking behavior of the advisee. Of particular relevance are impressions about the adviser's trustworthiness, because the other's trustworthiness plays a central role in the advisee's decision to ultimately follow or discount the advice. The trustworthiness of those who give advice appears to be central to many individual decisions and, ultimately, individual lives. For instance, impressions of a doctor's trustworthiness have been recently suggested as playing a central role in a patient's health and life expectancy (Baker et al., 2016; Pereira Gray, Sidaway-Lee, White, Thorne, & Evans, 2018). It is, for instance, likely that higher trust in the doctor may make an individual more willing to take particular actions, even when their positive outcomes are not readily and clearly predictable, resulting in the long-term benefit of the patient. However, to our knowledge, no study has experimentally tested whether a good reputation induces higher trust in an adviser and whether it does so independently of proximal benefits.

## Trust and Trustworthiness

Across disciplines, trust is defined as the willingness to accept vulnerability based on positive expectations of the intentions and behavior of another person (Rousseau, Sitkin, Burt, & Camerer, 1998). Complementary, trustworthiness can be defined as the likelihood that the other intends to provide truthful information and can be distinguished from expertise, which is the capacity to do so (Harris, Hahn, Madsen, & Hsu, 2016; Madsen, 2016). Behaviorally, trust and trustworthiness have mainly been investigated using the trust game (TG; Berg, Dickhaut, & McCabe, 1995). In this game, an investor is endowed with a monetary amount and can decide whether to share (trust) some of it with the trustee. If money is shared, the trustee receives the generally tripled amount of what invested and can decide to send back part of it (reciprocate).

Using the TG, previous studies have shown that trust in others hinges on both the character and intentions of the social partner. For instance, individuals have been shown to trust those who have a morally good character (Delgado, Frank, & Phelps, 2005). Further, having good intentions or being a good cooperator increases others' willingness to trust (Falk, Fehr, & Fischbacher, 2008; McCabe, Rigdon, & Smith, 2003; Nelson, 2002; van 't Wout & Sanfey, 2008), whereas signs of self-related motives in a social interaction (for instance, threating sanctions if not enough money is shared) decrease willingness to trust (Li, Xiao, Houser, & Montague, 2009).

Research in social psychology has identified the warmth-trustworthiness dimension as one universal dimension of social cognition that includes, among others, traits like generosity, honesty, righteousness, and sincerity (Cuddy, Glick, & Beninger, 2011; Fiske, Cuddy, & Glick, 2007). An influential model of trustworthiness identifies three determining factors of trustworthiness impressions: ability, benevolence, and integrity (Mayer, Davis, & Schoorman, 1995). Different social cues may signal these traits, for instance, whether someone has previously broken a promise, or has second aims, or does not apologize for a transgression (Kim, Ferrin, Cooper, & Dirks, 2004; Levine & Schweitzer, 2015; Sah, Loewenstein, & Cain, 2013; Schweitzer, Hershey, & Bradlow, 2006). In particular, the tendency to act with integrity reveals someone's honesty—a trait that implies social qualities pivotal to cooperation, like sincerity, fairness and truth-telling (Ashton et al., 2004; Levine, Bitterly, Cohen, & Schweitzer, 2018). Honesty appears to foster prosocial behavior (Ashton & Lee, 2007; Lee & Ashton, 2004), because it is associated with altruistic behavior, unconditional kindness and reciprocity (Ashraf, Bohnet, & Piankov, 2006; Baumert, Schlösser, & Schmitt, 2014; Hilbig, Thielmann, Hepp, Klein, & Zettler, 2015; Thielmann & Hilbig, 2015).

Honesty might thus represent one important antecedent of trustworthiness impressions about advisers. On the one hand, honest group members are believed to be a source of truthful information (Gordon & Spears, 2012). Because sharing of truthful information increases the accuracy of both individual and group decisions (Becker, Brackbill, & Centola, 2017; Mellers et al., 2014) and information sharing is associated with trust (Burt & Knez, 1995), individuals may form trustworthiness impressions of an honest adviser that make them more willing to take the adviser's advice. On the other hand, honest social partners signal good intentions and a good character, which makes them desirable candidates for future cooperation (Thielmann & Hilbig, 2015; Zhao & Smillie, 2015). Individuals may hence be more willing to trust an honest partner also in future interactions and across contexts. However, how people integrate information from others' honest behavior to form trustworthiness beliefs that guide their trust decisions in social situations is still unknown. Moreover, given that information from reliable sources is considered as more convincing than information from unreliable sources (Bovens & Hartmann, 2003; Hahn, Harris, & Corner, 2009), reputation for being honest as a sign of source credibility might impact how these trustworthiness beliefs are formed and revised.

## Reputation Learning

Previous studies have suggested that cues about others' behavior during repeated interactions are integrated via reinforcement learning processes. Reinforcement learning is proposed to allow the formation of beliefs about another person based on feedback following one's decisions—beliefs that ultimately inform trusting behaviors (Delgado et al., 2005; Fouragnan et al., 2013; King-Casas et al., 2005). For instance, trustworthiness beliefs are dynamically updated based on feedback about the other's reciprocity over multiple interactions with the partner in a TG (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010). Similarly, in an advice-taking paradigm, participants integrate advice by weighting the different outcomes of the recommended and not recommended options (Biele, Rieskamp, & Gonzalez, 2009).

Moreover, reinforcement learning relies on prediction errors that signal the discrepancy between actual and expected rewards and are encoded, among others, in the orbitofrontal cortex (Rudebeck, Saunders, Prescott, Chau, & Murray, 2013; Tsuchida, Doll, & Fellows, 2010). Crucially, in a neuroimaging study with functional MRI, activity in the orbitofrontal cortex was associated with valuation of expert advice before a decision with the advice was made, suggesting that on the neural level, a reinforcement learning mechanism is involved in advice utilization (Meshi, Biele, Korn, &

Heekeren, 2012). These results provide behavioral and neural evidence that reinforcement-learning models may be best suited to capture the mechanistic dynamics of honest reputation learning for trust.

## The Current Research

The extant literature has mainly operationalized the quality of the advice through the reward magnitudes associated with the advice (Behrens et al., 2008; Biele et al., 2009; Biele, Rieskamp, Krugel, & Heekeren, 2011; Diaconescu et al., 2014, 2017; Meshi et al., 2012; Rodriguez Buritica, Heekeren, & van den Bos, 2019; Yaniv & Kleinberger, 2000). That is, good, informative advice is generally operationalized as the best or one of the best choice options available, which is in turn associated with the highest or one of the highest reward outcomes. On the contrary, poor, uninformative advice is in general operationalized as the worst or one of the worst choice options in the task, which is in turn associated with the lowest or one of the lowest reward outcomes. However, advice can be informative without being an advice about the best or near-best option (e.g., advice not to do something or how to make a decision; Dalal & Bonaccio, 2010). This information-reward confound may have reduced social information processing to reward processing in previous studies. That is, learning and cognitive processes associated with estimations of reward outcome contingencies are difficult to disentangle from evaluations about the partner's character (e.g., their competence, honesty and generosity) in previous work.

Further, in many advice-taking paradigms in which participants are supposed to interact with their advisers, advisers have incentives to send truthful advice (Bonaccio & Dalal, 2006). Although on the one hand, this may have preserved the face-validity of the task, on the other, it may have forced participants to keep track of the reasons behind the partner's behavior to detect when she would be motivated to do otherwise (Behrens et al., 2008; Diaconescu et al., 2014), shifting the focus away from learning the partner's character. A similar issue has been raised for the TG, in which trusting might be associated with other behaviors that are equally likely to arise from other causes such as one's own benefits associated with the act of trust (Dirks & Ferrin, 2002; Kramer, 1999). Previous studies have confirmed that when external incentives cease or trusting is associated with monetary losses, trust rates drop drastically (Aimone & Houser, 2012; Rode, 2010), suggesting that some sort of strategic, reward-driven thinking is intertwined with trust decisions in this context (Camerer, 2003).

In many instances in life, giving good advice, which represents a form of prosocial behavior similar to helping, being generous, or supporting charities, has no clear incentives for the adviser and presupposes no clear commitment to reciprocate for the advisee. Indeed, giving good advice as well as keeping a secret represent acts of trust that rely on the impressions of the partner's trustworthiness without the requirement of an initial generous act and thus beyond a mere calculative reaction (Levine et al., 2018). For these reasons, we here developed a novel and more ecological paradigm (e.g., the take advice game [TAG]) in which (a) an adviser has no incentives to give good advice or deterrents to give bad advice so that an adviser's behavior can largely be traced back to her good or bad character and in which (b) the adviser's advice and the reward feedback are decorrelated.

In three lab-based experiments, we investigated the relationships between honesty and trust. In particular, we examined how honesty in advice-giving induces trustworthiness impressions that inform trusting behaviors across social contexts. In Experiment 1, we investigated whether individuals would take informative, truthful advice independently of the positive outcomes associated with the choice to take the advice. Further, we investigated whether a good reputation as honest person would generalize to a new trust-based social situation. In Experiment 2, we replicated the results of Experiment 1 and tested whether advice-taking behavior follows from first impressions of the other's honesty without feedback about current honesty. Moreover, we investigated whether inconsistency in honest behavior would be perceived more negatively than having no reputation at all. In Experiment 3, we applied reinforcement-learning models to mathematically formalize how social information is processed and integrated to form and update beliefs about the other's reputation.

## Experiment 1

In Experiment 1, we investigated how advice utilization changes as a function of the adviser's good and bad reputation established through honest and dishonest behavior. In particular, we were interested in how individuals use advice in cases in which they cannot rely on self-acquired knowledge but only on information provided by others. Participants were invited to the lab and told they were going to play two games with other participants, that is, the TAG (Figure 1 and Figure S1 in the online supplemental materials) and the TG. Participants believed that role assignment was randomly determined by drawing a ball from a box. In fact, every participant was assigned the role of advisee in the TAG and investor in the TG.

In the TAG, participants received advice from different advisers about one of two cards they otherwise knew nothing about. Thus, in the TAG, our participants were completely uninformed and had to make decisions based exclusively on the information provided by the advisers. In particular, they were required to pick the highest card to win money (€1). This scenario creates an interdependency between advisers and participants necessary for trust. Moreover, participants also knew that the advisers could see only one of the two cards. Hence, even though the advisers knew more about the cards than participants, they were not provided with complete knowledge. This means that the advisers could provide participants with informative advice (e.g., the number they saw on one of the two cards) but could not advise them about the best decision to make (i.e., which card was the winning one). This makes our design highly ecological, as it resembles many advice-giving scenarios in real life where advisers often do not even know (or cannot know) what the best action to take is in a particular context. For instance, school counselors can provide students with valuable information about different career paths but cannot know which of these career paths may be most suitable for a particular student.

Further, our participants knew that the advisers were asked to help them but there were no external incentives for the advisers to do so in the TAG. However, both partners knew that after the TAG, they were going to interact again in a new social interaction, namely, in the TG. In the TG, our participants, now in the role of investors, could decide to share money (€3) with the advisers who

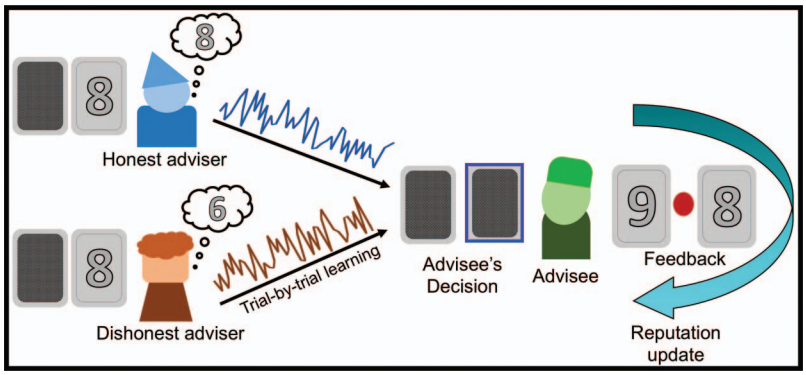


*Figure 1.* Take advice game. Schematic representation of the paradigm. Advisers could see one of the cards and send an advice to the advisee who had to decide which card to pick to win money. Advisees won money if they had picked the card with the highest number. To do that, however, they had to rely exclusively on the adviser's advice. Over trials, the advisee could learn, on the basis of the information provided in the feedback phase, who of the advisers was honest. Thereby, advisees updated their beliefs about advisers' reputation, which in turn guided their future trust decisions. See the online article for the color version of this figure.

advised them previously. This design closely resembles real-life scenarios in which giving advice has often no proximal benefits for the adviser but may contribute to improve her reputation for future cooperation (a distal benefit).

Finally, we manipulated the informativeness of the adviser's advice. In particular, some advisers shared informative advice (honest advisers), while others shared uninformative advice (dishonest advisers). Importantly, informative advice was not equal to the best or rewarding option in the task. For instance, as shown in Figure 1, the honest adviser might send advice informative of the actual number on the right card, which, however, might still turn out to be the losing card. Thereby, we were able to disentangle social information about the adviser's honesty from reward information related to the advisee's decisions. This allowed us to separately investigate the different cognitive processes underlying social information and reward information processing. Finally, although the advisees did not commit to repay the adviser's honesty, they had the opportunity to do so in the TG. The TG was, thus, for the advisee a way to repay the adviser's honesty, but for the adviser, it was a weak and uncertain benefit that did not represent a strong incentive to behave in a particular fashion.

## Method

**Subjects.** We invited 28 participants (18 females; 21.43 ± 3.47, mean age ± *SD*) to our lab to participate in Experiment 1. We used G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) to calculate the desired sample size. We based our power analysis on previous research on the effects of others' moral character on trustworthiness impressions (Delgado et al., 2005), which found an effect size of ~1.5. Because we reasoned that the effect of honesty might be less strong than that of moral character and given that even replication effects are half the magnitude of the mean effect size of the original effects (Open Science Collaboration, 2015), we estimated an effect size of 0.7 for our study. To achieve a power of 0.90, we needed a sample of 24 participants. Hence, we aimed for 30 participants to ensure the desired final sample size in case ineligible participants had to be excluded.

For this and the next two experiments, exclusion criteria were present or past neurological and psychiatric disorders, pharmacological medication up to 2 weeks prior to the study (including participation in other experiments with pharmacological intervention), current physical or mental stress, and other severe health complications. All participants had normal or corrected-to-normal vision. The study was approved by the local Ethics Committee of the University of Lübeck (15–185: "Psychological and Neural Correlates of Social Cognition and Regulation") and conducted in accordance to the Declaration of Helsinki. All participants provided written consent for participation. The experiment took around 1 hr, and participants received an hourly wage of €8/hr for participation.

**Procedure.** In the TAG, participants in the role of advisee interacted with other coplayers in the role of adviser (Figure 1 and Figure S1A and S1B in the online supplemental materials). Participants were invited to the lab all together. They were told that they were going to be assigned different cubicles and play with each other two games. In these games, different roles were possible, but they could not know who received which role. Roles were assigned by drawing a ball from a box. Each participant drew a ball with a different letter on it. All letters represented two possible roles in the game, but participants did not know which letter represented which role. After being assigned one cubicle, participants played the games on one of the lab computers. At the very beginning of the task, participants were prompted to insert the letter they drew. Afterward, instructions to their role were presented. This procedure was implemented to increase participants' credibility. In reality, each participant received the role of advisee in the TAG and investor in the TG, and the advisers were computer-programmed.

After instructions, participants were told that each participant was required to choose their avatar, which would have guaranteed anonymity in the games. Participants could use the mouse to pick one out of 10 avatars (i.e., "Greeble"; Gauthier & Tarr, 1997) presented on the screen in a circle. Participants knew that it was not possible to choose the same avatar as another, as different

subsets of avatars were randomly presented to each participant. Greebles were chosen because we assumed that participants were not familiar with this particular stimulus material. This enabled us to further increase the credibility in the task and to present different recognizable avatars that were supposed to represent different advisers. This, in turn, allowed participants to trace each adviser's behavior and change their decisional strategy accordingly. Finally, the use of avatars made it possible to create a social context in which participants could learn the partner's trustworthiness only based on the partner's behavior without implementing commonly used stimulus material, such as faces, that might have introduced confounding factors (e.g., facial trustworthiness).

After avatar selection, participants were connected to the other coplayers. A fake connection procedure was presented, which failed twice and took around one minute to be successful. Participants saw a text saying: "another participant is not ready yet." Once all participants were supposedly connected, a countdown of 5 s was presented with the text: "make yourself ready to start." To improve the credibility of the connection, breaks between runs were also timed. Participants believed that everybody had a 15-s break after each run. Before the subsequent run started, a countdown of 10 s was presented asking participants to get ready.

Participants played in a randomized order 12 trials with each of four advisers (priming block). Two of them were highly honest (whose advice was truthful 100% of the time), and the other two highly dishonest (whose advice was truthful 0% of the time). Whereas the truthful advice was always the exact card number on one of the two cards, untruthful advice could have been either a card number advised on the wrong card or a wrong card number. Thus, untruthful advice from dishonest advisers was not informative of the actual card numbers or the likely winning card in a given trial. After that, participants played a total of 192 trials with all advisers ($N = 8$) including inconsistently honest advisers whose advice was truthful 50% of the time. The study design was within-subjects, thus every participant played with each of these different advisers.

Participants could win €1 by choosing the highest card. Advisers received one of the cards disclosed and could share this information with the advisees. Advisers always gave advice at the beginning of each trial. The trial consisted of four phases. First, after a short interstimulus interval (ISI) of 0.5 s, participants saw the avatar of the adviser they were matched with in the trial at the center of the screen for 2 s (adviser phase). Second, the advice appeared either to the right or to the left for 1 s (advice phase). Third, after a variable ISI (range between 2 and 8 s, mean: 4 s), the two cards appeared again and participants had 1 s to choose one of them by pressing either the "e" (for left) or the "i" (for right) key on a standard QWERTY computer keyboard (decision phase). Feedback was finally presented for 1 s (feedback phase) before the intertrial interval (ITI; 2–8 s, mean: 4 s). In the feedback phase, participants received two types of information: (a) the actual card numbers based on which they could learn the partner's honesty (social information); (b) a positive or negative feedback informing whether the participant won or lost in the trial (nonsocial information).

To disentangle honesty from reward information, we made truthful advice unpredictive of the winning card. By manipulating the card number sampling procedure, we ensured that each truthful advice was associated with the winning card only 50% of the time (chance level) and that no higher gains were associated with truthful advice as opposed to untruthful advice. However, to avoid that additional cognitive processes related to how participants processed the advice from the different advisers might have confounded our results, we optimized the pseudorandom sampling procedure to have a realized probability of card drawing that approximates chance for each adviser's advice. Feedback analyses confirmed that participants received the same proportion of positive and negative feedback irrespective of the honesty of the adviser (see below) and Kolmogorov–Smirnov tests confirm that the distributions of advised card numbers did not change across advisers (see below). Finally, for each adviser, the advised number was presented half of the time to the left and half of the time to the right card in a randomized order.

After the TAG, we tested the generalizability of honesty-based trustworthiness impressions to a different context, that is, the TG. Here, participants played as investor a one-shot TG with each adviser in the role of trustee. For each adviser, they received an initial endowment, that is, 10 monetary units (MUs) and decided whether to share it with the adviser. They could share nothing or any portion of this endowment on an 11-point Likert-scale ranging from 0 to 10. Participants knew that the shared amount would be tripled by the experimenter and that the advisers could then reciprocate by sending back any portion of this tripled amount. Participants did a single investment decision for each adviser before they could see how much the advisers reciprocated to avoid that an adviser's reciprocity could affect participants' subsequent investment decisions with the other advisers. Moreover, we asked participants to rate the advisers' behavior during the TAG (i.e., trustworthiness ratings) while the advisers were making their back-transfer decisions (i.e., before seeing the adviser's reciprocity). Trustworthiness ratings were made on a 7-point Likert scale ranging from *very untrustworthy* to *very trustworthy*. Participants' payoffs in MUs were then converted into Euros. Participants knew that each MU corresponded to 30 cents. No time constraints were given for either investment decisions in the TG or trustworthiness ratings.

Stimuli in this and the following experiments were presented using Psychtoolbox 3 (http://psychtoolbox.org) on MALTAB 2016b (https://www.mathworks.com).

**Analyses.** Trusting behavior in the TAG was assessed as advice-taking behavior, which was operationalized as the probability of choosing a card given the truthfulness of the advice received. Thus, as participants knew that card numbers ranged from 1 to 9 (except for 5), they were considered to take an adviser's advice, if their card-choice behavior was consistent with a "$>5$" strategy according to the advice's truthfulness. In particular, had the advice been truthful (i.e., it was provided by honest advisers) and had the advised number on the advised card been greater than five, participants should have been more likely to choose the advised card over the other. Similarly, they should have been more likely to choose the other card if the truthful advice on the advised card had been a number smaller than five. On the contrary, we expected a more random card-choice behavior for untruthful advice (i.e., by dishonest advisers), as we hypothesized that participants would discount uninformative advice from dishonest advisers. An 8 (card number) $\times$ 3 (adviser) repeated-measures analysis of variance (ANOVA) confirmed our hypotheses revealing an interaction effect between advisers and card

numbers, $F_{(14,378)} = 5.28$ ($p < .001$; $\eta_p^2 = 0.16$), indicating that participants were more likely to show a card-choice behavior following a >5 strategy for honest advisers, while discounting the uninformative advice of dishonest advisers with a more random pattern of card-choice behavior (Figure S2 in the online supplemental materials). A Kolmogorov–Smirnov test further confirmed that the distributions of advised card numbers did not change across advisers ($K\text{-}S = 0.08$; $p = .523$). Finally, trusting behavior in the TG (Figure S3 in the online supplemental materials) corresponded as usual to the amount of money invested by the participants (Berg et al., 1995).

Overall difference in advice-taking behaviors in the TAG, investment behavior in the TG, and trustworthiness judgments was assessed by computing a one-way ANOVA with one factor (adviser) and three levels (honest, dishonest and inconsistently honest advisers). Differences between the levels were tested by one-sample *t* tests (two-tailed). Mixed-effects logistic regression analyses were computed to test which variable could better explain trial-by-trial advice-taking behavior in the TAG, implementing the following formula:

$$AT_t = \beta_0 + \beta_1 HA_t + \beta_2 DA_t + \beta_3 AN_t + \beta_4 AC_t + \beta_5 FB_{t-1}, \quad (1)$$

where $AT_t$ is advice-taking behavior (1 = advice taken; 0 = advice not taken) on trial $t$, $\beta_0$ is the intercept, $HA_t$ codes for the honest adviser on trial $t$ (1 = honest adviser; 0 = otherwise), $DA_t$ codes for the dishonest adviser on trial $t$ (1 = dishonest adviser; 0 = otherwise), $AN_t$ is the advised number on trial $t$, $AC_t$ is the advised card on trial $t$ (1 = right; −1 = left), $FB_{t-1}$ is the feedback received on trial $t-1$ after the current adviser's advice (1 = positive; 0 = negative). To examine the robustness of our results, we further ran all analyses using Bayesian statistics, which confirmed the results based on frequentist statistics. To investigate the relationships between advice-taking behavior in the TAG, investment decisions in the TG and trustworthiness ratings, Pearson correlations were computed.

Frequentist and computational modeling analyses (see Experiment 3) were performed in MATLAB 2016b (https://www.mathworks.com). Bayesian mixed-effects model regressions were performed in R (v. 3.6.1) by R Core Team (2019; https://www.R-project.org/) using *brms* from Stan (https://mc-stan.org; Carpenter et al., 2017). All other Bayesian analyses were performed in JASP (v. 0.10.2) by JASP Team (2019; https://jasp-stats.org).

## Results

An ANOVA with advisers as within-subject factor and three levels (honest, dishonest, and inconsistently honest advisers) revealed that participants' advice-taking behavior was significantly different across advisers, $F_{(2,54)} = 15.51$ ($p < .001$; $\eta_p^2 = .365$; Bayes Factor for the alternative over the null hypothesis [$BF_{10}$] = 3,520.80). Participants took on average the advice of honest advisers significantly more than that of dishonest, $t_{(27)} = 4.65$ ($p < .001$; 95% CI [0.08, 0.22]; Cohen's $d = 0.88$; $BF_{10} = 326.61$) or inconsistently honest ones, $t_{(27)} = 4.27$ ($p < .001$; 95% CI [0.06, 0.17]; $d = 0.81$; $BF_{10} = 130.56$; Figure 2A and 2B). Advice from inconsistently honest advisers was taken descriptively more than advice from dishonest advisers, but not significantly so, $t_{(27)} = 1.53$ ($p = .139$; 95% CI [−0.09, 0.01]; $d = 0.29$; $BF_{10} = 0.56$; Figure 2B).

By implementing a mixed-effects logistic regression (equation 1), we then tested whether trial-by-trial advice-taking behavior could be predicted by the good reputation of the honest adviser irrespective of reward information (Table 1 and Table S1 in the online supplemental materials). Our results revealed that participants integrated information about the advised card number into their decisions ($\beta = 0.05$; $p < .001$; $SE = 0.01$; 95% CI [0.02, 0.07]). Moreover, they were more likely to take honest advisers' advice ($\beta = 0.53$; $p < .001$; $SE = 0.08$; 95% CI [0.38, 0.69]), and less likely to take dishonest advisers' advice ($\beta = -0.29$; $p < .001$; $SE = 0.08$; 95% CI [−0.44, −0.15]), suggesting that participants' trusting behavior was based on the integration of information about the received advice and the source of that advice.

On the contrary, reward information from the previous trial was not significantly associated with trial-by-trial advice-taking behavior ($\beta = 0.005$; $p = .934$; $SE = 0.06$; 95% CI [−0.12, 0.13]). Moreover, comparisons of types of feedback (positive and negative) received across advisers indicate that the ratios of positive and negative feedback did not differ across advisers, $F_{(2,54)} = 1.34$ ($p = .271$; $\eta_p^2 = .047$; $BF_{10} = 0.34$). For completeness, we check that neither a comparison between honest and dishonest advisers, $t_{(27)} = 1.47$ ($p = .154$; 95% CI [−0.02, 0.10]; $d = 0.27$; $BF_{10} = 0.49$) nor between inconsistently honest advisers and honest, $t_{(27)} = 0.154$ ($p = .872$; 95% CI [−0.04, 0.05]; $d = 0.03$; $BF_{10} = 0.20$) or dishonest, $t_{(27)} = -1.25$ ($p = .212$; 95% CI [−0.10, 0.02]; $d = 0.24$; $BF_{10} = 0.41$) advisers reached significance. Hence, these results suggest that even though informative advice was not useful for the advisee to gain more benefits, participants still preferred taking the informative advice of honest advisers.

Moreover, investment behavior in the TG, $F_{(2,54)} = 7.44$ ($p < .002$; $\eta_p^2 = .216$; $BF_{10} = 23.37$) and trustworthiness ratings, $F_{(2,54)} = 11.65$ ($p < .001$; $\eta_p^2 = .301$; $BF_{10} = 3{,}120.38$) were influenced by the reputation of the advisers established in the TAG (Figure 2C and 2D). In particular, in the TG, participants shared significantly more money with honest than dishonest ($M_{honest} = 6.04$, $M_{dishonest} = 4.32$; $t_{(27)} = 2.87$; $p = .008$; 95% CI [0.49, 2.94]; $d = 0.54$; $BF_{10} = 5.64$) or inconsistently honest advisers ($M_{honest} = 6.04$, $M_{inconsistent} = 4.88$; $t_{(27)} = 2.65$; $p = .013$; 95% CI [0.26, 2.06]; $d = 0.50$; $BF_{10} = 3.65$). Also, inconsistently honest advisers tendentially received more than dishonest advisers, $t_{(27)} = -2.11$ ($p = .044$; 95% CI [−1.09, −0.02]; $d = 0.40$; $BF_{10} = 1.35$), suggesting that being honest, even if inconsistently so, might pay more than dishonesty. Moreover, honest advisers were also judged as more trustworthy than dishonest ($M_{honest} = 4.86$, $M_{dishonest} = 3.29$; $t_{(27)} = 4.25$; $p < .001$; 95% CI [0.81, 2.33]; $d = 0.80$; $BF_{10} = 125.59$) and inconsistently honest advisers ($M_{honest} = 4.86$, $M_{inconsistent} = 3.88$; $t_{(27)} = 2.71$; $p = .012$; 95% CI [0.24, 1.71]; $d = 0.51$; $BF_{10} = 4.11$), whereas inconsistently honest advisers induced slightly higher trustworthiness impressions than dishonest advisers, $t_{(27)} = -2.47$ ($p = .020$; 95% CI [−1.10, −0.10]; $d = 0.47$; $BF_{10} = 2.58$).

Finally, advice-taking behavior in the TAG significantly correlated with the amount of money shared in the TG (honest advisers: $r_{(26)} = 0.53$, $p = .004$; 95% CI [0.20, 0.76]; $BF_{10} = 13.52$; dishonest advisers: $r_{(26)} = 0.53$, $p = .004$; 95% CI [0.19, 0.75]; $BF_{10} = 11.81$) and with trustworthiness judgments (honest advisers: $r_{(26)} = 0.57$, $p < .002$; 95% CI [0.25, 0.78]; $BF_{10} = 25.78$;

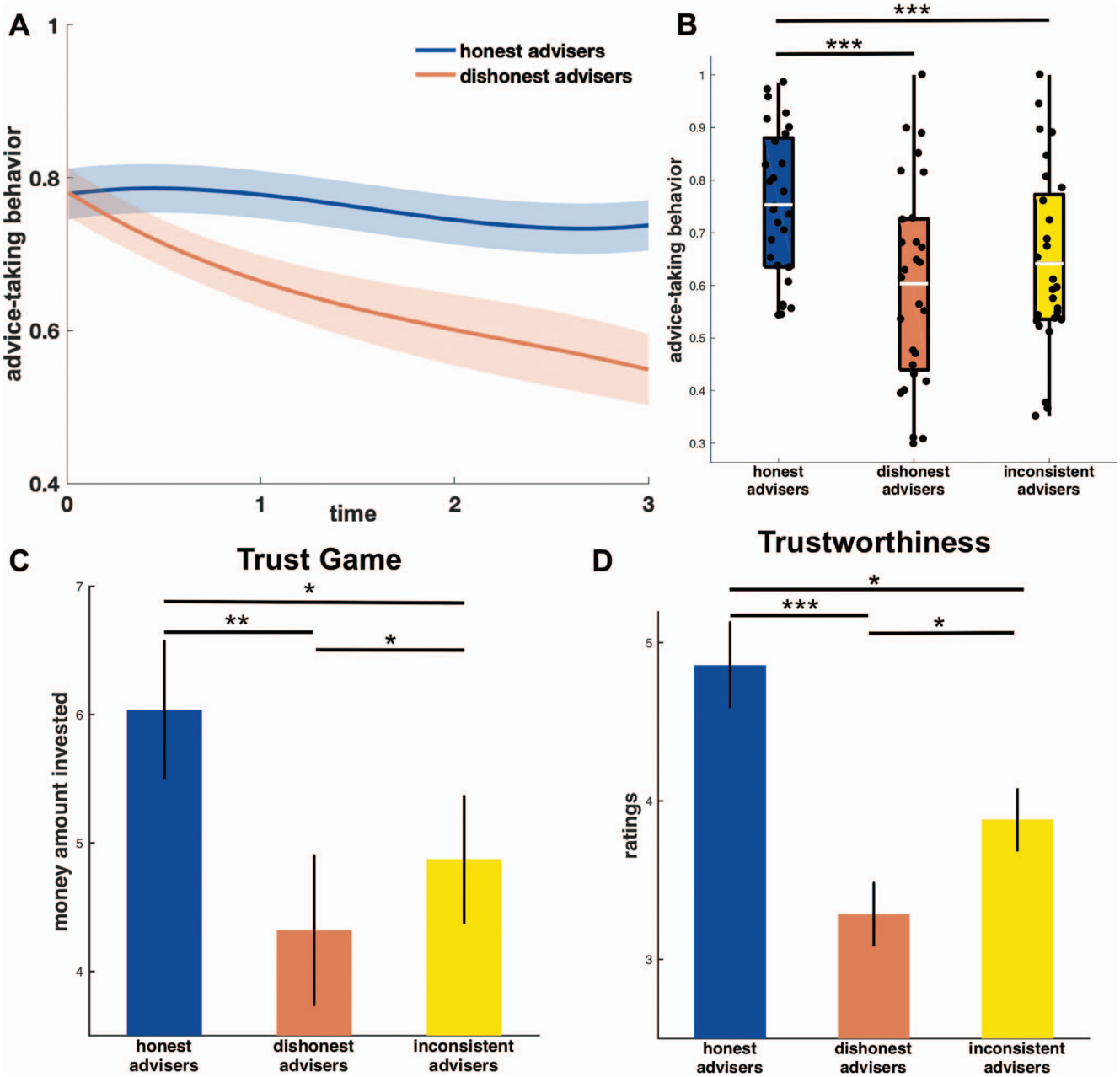


*Figure 2.* Experiment 1. (A) Average advice-taking behavior in the take advice game (TAG) over time (i.e., blocks) toward honest and dishonest advisers (± *SEM* in shadowed area). Data points were interpolated for visualization purposes. (B) Average individual advice-taking behavior toward honest, dishonest and inconsistently honest advisers. (C) Average amount of money entrusted in the trust game. (D) Average trustworthiness ratings. $^{*}$ $p < .05$. $^{**}$ $p < .01$. $^{***}$ $p < .001$. See the online article for the color version of this figure.

dishonest advisers: $r_{(26)} = 0.44$, $p = .019$; 95% CI [0.08, 0.70]; $BF_{10} = 3.17$; Figure 3A) for both honest and dishonest advisers. Importantly, however, winnings in the TAG did not correlate with the amount of money shared in the TG (honest advisers: $r_{(26)} = -0.07$, $p = .717$; 95% CI [−0.43, 0.31]; $BF_{10} = 0.25$; dishonest advisers: $r_{(26)} = 0.23$, $p = .243$; 95% CI [−0.16, 0.55]; $BF_{10} = 0.45$), suggesting that participants' willingness to trust in the TG specifically represented an act of reciprocity for the adviser's honesty in the previous interaction.

## Discussion

Results of Experiment 1 suggest that our task was successful in inducing honesty-based trustworthiness impressions in our participants. Participants trusted honest others more than dishonest ones and repaid them for their honesty in a subsequent social interaction. This is consistent with previous findings showing a decrease in trust and cooperation after a defection (Baumgartner, Fischbacher, Feierabend, Lutz, & Fehr, 2009; Hula, Vilares, Lohrenz,

Table 1
*Mixed-Effects Regression Analyses of Advice-Taking Behavior*

| | Estimates | | | |
|---|---|---|---|---|
| | Study 1 | Study 2 | | Study 3 |
| Regressor | | Feedback block | No-feedback block | |
| Intercept | 0.44 (0.20)* | 0.35 (0.22) | 1.29 (0.25)*** | 1.45 (0.20)*** |
| Honest adviser | 0.53 (0.08)*** | 0.57 (0.14)*** | 0.26 (0.09)** | 0.32 (0.06)*** |
| Dishonest adviser | −0.29 (0.08)*** | | −0.40 (0.08)*** | |
| Advised number | 0.05 (0.01)*** | 0.06 (0.03)* | 0.01 (0.01) | −0.01 (0.01) |
| Advised card | 0.05 (0.06) | 0.09 (0.13) | 0.03 (0.07) | −0.03 (0.06) |
| Feedback previous trial | 0.01 (0.06) | 0.25 (0.14) | | 0.05 (0.06) |
| $R^2$ | 0.21 | 0.09 | 0.24 | 0.14 |

*Note*. β coefficients (standard errors) from mixed-effects logistic regression model with random participant-level intercept predicting advice-taking behavior (1 = advice taken; 0 = advice not taken).
* $p < .05$. ** $p < .01$. *** $p < .001$.

Dayan, & Montague, 2018). Interestingly, even though the behavioral unpredictability of inconsistently honest partners disincentivized participants from taking advice from them, inconsistently honest partners were still trusted more and perceived as more trustworthy than dishonest partners in a subsequent social interaction. These results suggest that even small signs of honesty make people willing to reciprocate with more trust. Finally, even though the advice was not helpful for the participants to improve their performance, participants consistently integrate truthful, informative advice of honest partners into their decisions. This result suggests that truthful advice is associated with an information bonus that is integrated into the decision-making process by uninformed decision-makers. It is still an open question, however, whether participants would take the advice of an honest partner even if they were not able to estimate the current honesty of the partner and whether they would be more willing to take advice from those with an established reputation than from those whose reputation is still unknown.

## Experiment 2

In everyday life, we often need to seek out to others for advice about the most disparate reasons. However, we do not always have the opportunity to check the quality of the other's advice before we decide whether to take it. Thus, the decision to follow one or many pieces of advice often needs to be made before the quality of each advice is clear or determinable. In Experiment 2, we investigated how individuals integrate information from others in the absence of proximal feedback about the truthfulness of the other's advice. Moreover, we also investigated the extent to which individuals are willing to take advice from social partners without a known reputation as opposed to those with an established reputation as honest or dishonest partners. A better understanding of advice-taking behaviors toward these unknown others is of pivotal importance because we often need to decide whether to take advice from strangers (e.g., taking advice from a new doctor with respect to a certain therapy or from a new boss with respect to a particular work assignment).

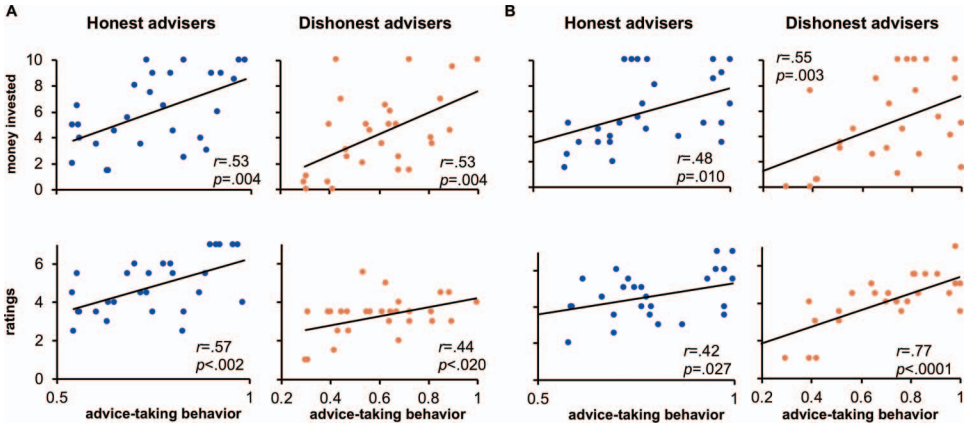


*Figure 3.* Correlations between the different measures of trust. (A) Pearson correlations between advice-taking behavior in the TAG and both investment behavior in the TG and trustworthiness ratings for honest and dishonest advisers in Experiment 1. (B) Pearson correlations between advice-taking behavior in the no-feedback block of the TAG and both investment behavior in the TG and trustworthiness ratings for honest and dishonest advisers in Experiment 2. TAG = take advice game; TG = trust game. See the online article for the color version of this figure.

To do so, we implemented a new version of the TAG with two main blocks. In the first block (feedback block), participants learnt the advisers' honesty through social information provided in the feedback phase, analogous to Experiment 1. In the second block (no-feedback block), participants did not receive any feedback about the truthfulness of the adviser's advice anymore (Figure S1B in the online supplemental materials). Moreover, in the no-feedback block, participants played with new advisers they did not meet in the previous (feedback) block and whose reputation was hence unknown to them (i.e., advisers without reputation). After the TAG and similarly to Experiment 1, participants in the role of investor interacted again with all advisers in the TG.

## Method

**Subjects.** We invited 28 participants (18 females; 24.54 ± 4.0, mean age ± *SD*) to our lab to participate in the experiment. The power analysis for Experiment 2 was based on the smallest effect size from Experiment 1, namely, the differences in investment behaviors in the TG ($\eta_p^2 = .216$). Based on this effect size, we estimated a needed sample size of 26 participants to achieve a power of 0.90. Hence, we aimed for a similar sample size as in Experiment 1.

The same exclusion criteria as Experiment 1 were applied. All participants had normal or corrected-to-normal vision. The study was approved by the local Ethics Committee of the University of Lübeck and conducted in accordance to the Declaration of Helsinki. All participants provided written consent for participation. Study length (approximately 1 hr) and hourly wage (€8/hr) were analogous to Experiment 1 to avoid that external factors such as fatigue or economic compensation for participation might have confounded the results making comparisons between the experiments difficult.

**Procedure.** In Experiment 2, participants were invited to the lab and were assigned their role like in the previous experiment and were told that they were going to play with each other two games. In Experiment 2, we implemented a new version of the TAG with two main blocks. The feedback block consisted of four phases: (a) adviser phase, (b) advice phase, (c) decision phase, and (d) feedback phase, analogous to Experiment 1. The no-feedback block consisted of only three phases, because no feedback phase was presented in this part of the task (Figure S1B in the online supplemental materials). Moreover, in the no-feedback block, new advisers were introduced whose reputation was unknown to the participants. As in the priming block of Experiment 1, participants played in a randomized order 12 trials with each honest (who shared truthful information 100% of the time) and dishonest advisers (who shared truthful information 0% of the time). Further, in the no-feedback block, participants played a total of 192 trials with all advisers ($N = 8$), analogous to Experiment 1. Importantly, participants were informed about these two blocks, their structure, and the introduction of new advisers (i.e., advisers without reputation) in the second, no-feedback block. Finally, the study design was within-subjects, thus every participant played with each of these different advisers.

As in Experiment 1, after the TAG, participants played the TG and rated the trustworthiness of the advisers in the TAG. In this study, we also introduced a catch and unrelated question on the attractiveness of the advisers. In particular, we asked our participants to report on a 7-point Likert-scale from very unattractive (0) to very attractive (7) how attractive they thought the other participants in the role of adviser were. As expected, attractiveness ratings did not differ across advisers (honest vs. dishonest advisers: $t_{(27)} = 0.88$; $p = .386$; 95% CI [−0.29, 0.71]; $d = 0.17$; $BF_{10} = 0.29$; honest advisers vs. advisers without reputation: $t_{(27)} = 1.18$; $p = .250$; 95% CI [−0.23, 0.83]; $d = 0.22$; $BF_{10} = 0.38$; dishonest advisers versus advisers without reputation: $t_{(27)} = 0.39$; $p = .701$; 95% CI [−0.38, 0.56]; $d = 0.07$; $BF_{10} = 0.22$). At the end of the task, we also asked participants to report (with binary response option: Yes/No) whether they used a strategy in the TAG and whether they thought that their strategy was successful for winning more money in the game. In particular, we were interested to know whether they were aware that the information received from the advisers was not useful to improve their performance. Timing of the tasks was as the previous experiment.

**Analyses.** The same analyses were implemented as in Experiment 1 to investigate advice-taking behavior in the TAG, investment behavior in the TG, trustworthiness ratings, and their relationships. We checked also in Experiment 2 that participants followed a > 5 strategy and an 8 (card number) × 3 (adviser) repeated-measures ANOVA confirmed this, $F_{(14,378)} = 2.41$ ($p = .003$; $\eta_p^2 = 0.08$). A Kolmogorov–Smirnov test further confirmed that the distributions of advised card numbers did not change across advisers ($K\text{-}S = 0.04$; $p = .993$). Moreover, in the feedback block in which only two types of adviser (honest and dishonest) were presented, the mixed-effects logistic regression analysis testing variables that could explain trial-by-trial advice-taking behavior in the TAG was slightly different from equation (1), namely:

$$AT_t = \beta_0 + \beta_1 HA_t + \beta_2 AN_t + \beta_3 AC_t + \beta_4 FB_{t-1}, \qquad (2)$$

where $AT_t$ is advice-taking behavior (1 = advice taken; 0 = advice not taken) on trial $t$, $\beta_0$ is the intercept, $HA_t$ codes for the honest adviser on trial $t$ (1 = honest adviser; 0 = dishonest adviser), $AN_t$ is the advised number on trial $t$, $AC_t$ is the advised card on trial $t$ (1 = right; −1 = left), $FB_{t\text{-}1}$ is the feedback received on trial $t$-1 after the current adviser's advice (1 = positive; 0 = negative). In the no-feedback block, the logistic regression analysis follows equation (1).

Finally, to compare advice-taking behavior toward inconsistently honest advisers in Experiment 1 with advice-taking behavior toward advisers without reputation in Experiment 2, between-subjects comparisons were performed using a two-sample $t$ test (two-tailed).

## Results

We first replicated our findings from Experiment 1. Results from the feedback block showed that participants took honest advisers' advice significantly more than dishonest advisers' advice, $t_{(27)} = 3.32$ ($p < .003$; 95% CI [0.04, 0.17]; $d = 0.63$; $BF_{10} = 14.64$). Regression analyses using equation (2) revealed that trial-by-trial advice-taking behavior was predicted by both the adviser's honesty ($\beta = 0.57$; $p < .001$; $SE = 0.14$; 95% CI [0.31, 0.84]) and the advised card number ($\beta = 0.06$; $p = .036$; $SE = 0.03$; 95% CI [0.004, 0.11]) but not by the reward outcome received on the previous trial ($\beta = 0.25$; $p = .072$; $SE = 0.14$; 95% CI [−0.02, 0.51]), suggesting again that participants' decisions were based on information about the received advice and the

honesty of its source, but not on reward information (Table 1 and Table S1 in the online supplemental materials). Moreover, we also checked whether participants received the same proportion of positive and negative feedback in this phase. Replicating results from Experiment 1, feedback information did not differ between honest and dishonest advisers, $t_{(27)} = 0.62$ ($p = .538$; 95% CI [−0.05, 0.09]; $d = 0.12$; $BF_{10} = 0.24$).

As shown in Figure 4A, in the no-feedback blocks, participants maintained their trust levels toward honest and dishonest others constant and accordingly to the adviser's reputation learned in the previous feedback block. An ANOVA on advice-taking behavior in the no-feedback blocks revealed a significant main effect of adviser, $F_{(2,54)} = 4.76$ ($p = .013$; $\eta_p^2 = .15$; $BF_{10} = 3.72$), with higher trust in advisers who demonstrated an honest behavior in the feedback block (honest vs. dishonest advisers: $t_{(27)} = 2.31$; $p = .029$; 95% CI [0.01, 0.21]; $d = 0.44$; $BF_{10} = 1.92$; Figure 4A). Further, the mixed-effects logistic regression described in equation (1) and fitted to participants' behavior in the no-feedback blocks (Table 1 and Table S1 in the online supplemental materials) revealed that good reputation as honest partner positively predicted trial-by-trial advice-taking behavior ($\beta = 0.26$; $p = .002$; $SE = 0.09$; 95% CI [0.09, 0.43]), whereas bad reputation as dishonest partner was associated with reduced advice-taking behavior ($\beta = -0.40$; $p < .001$; $SE = 0.08$; 95% CI [−0.56, −0.24]). Moreover, because participants did not receive any social information about the advice's truthfulness in the no-feedback blocks, the actual advice received (i.e., the card number) had no significant impact on advice-taking behavior, unlike the previous analyses ($\beta = 0.006$; $p = .650$; $SE = 0.01$; 95% CI [−0.02, 0.03]). These results suggest that participants decided whether to take or discount advice exclusively on the basis of the reputation of the adviser who shared it. These findings demonstrate that when the quality of the adviser's advice cannot be proven, the past reputation of the adviser is decisive in the question as to whether an individual would follow the present advice.

Although these results show the importance of a good reputation as honest partner for others to be willing to take one's advice, it is still an open question how willing an individual will be to take advice from those whose reputation is unknown in contexts in which an immediate control of their behavior is not possible. Interestingly, our results show that participants tended to be more willing to take advice from advisers without reputation than from dishonest advisers, $t_{(27)} = 2.12$ ($p = .044$; 95% CI [0.002, 0.14]; $d = 0.40$; $BF_{10} = 1.36$) but took as much advice from advisers without reputation as from honest advisers, $t_{(27)} = -1.71$ ($p = .099$; 95% CI [−0.09, 0.01]; $d = 0.32$; $BF_{10} = 0.73$; Figure 4B). These findings suggest that participants may have assumed that advisers without reputation might be as honest and trustworthy as advisers with an established good reputation. Crucially, however, participants took significantly more advice from advisers without reputation in Experiment 2 than from inconsistently honest advisers in Experiment 1, $t_{(2,54)} = 2.09$ ($p = .042$; 95% CI [0.004, 0.19]; $d = 0.56$; $BF_{10} = 1.60$; Figure 4C), suggesting that even though inconsistently honest behavior pays more than dishonesty, it still corrupts the initial positive expectation that others are trustworthy partners.

TG results in Experiment 2 replicated and extended results from Experiment 1. Participants shared their initial endowments depending on the honesty of the adviser ($F_{(2,54)} = 3.95$; $p = .025$; $\eta_p^2 = .13$; $BF_{10} = 2.03$) with honest advisers receiving tendentially more money than dishonest advisers ($M_{honest} = 5.89$, $M_{dishonest} = 4.71$; $t_{(27)} = 1.97$; $p = .059$; 95% CI [−0.05, 2.41]; $d = 0.37$; $BF_{10} = 1.08$) and advisers without reputation ($M_{honest} = 5.89$, $M_{without\text{-}reputation} = 4.89$; $t_{(27)} = 2.70$; $p = .012$; 95% CI [0.24, 1.76]; $d = 0.51$; $BF_{10} = 4.00$). No differences were found, however, between dishonest advisers and advisers without reputation despite descriptively more money entrusted to advisers without reputation ($t_{(27)} = -0.52$; $p = .606$; 95% CI [−0.88, 0.52]; $d = 0.10$; $BF_{10} = 0.23$). Although trustworthiness ratings in Experiment 2 did not reach significance ($F_{(2,54)} = 2.42$; $p = .099$; $\eta_p^2 = .08$; $BF_{10} = 0.76$), they show similar rating patterns as in Experiment 1. In particular, honest advisers ($M_{honest} = 4.55$) received higher trustworthiness ratings than dishonest advisers ($M_{dishonest} = 3.93$) and advisers without reputation ($M_{without\text{-}reputation} = 4.07$). Moreover, advice-taking behavior in the no-feedback block correlated with investment behavior in the TG (honest advisers: $r_{(26)} = 0.48$, $p = .010$; 95% CI [0.13, 0.72]; $BF_{10} = 5.56$; dishonest advisers: $r_{(26)} = 0.55$, $p = .003$; 95% CI [0.22, 0.76]; $BF_{10} = 17.27$; advisers without reputation: $r_{(26)} = 0.43$, $p = .021$; 95% CI [0.07, 0.69]; $BF_{10} = 2.93$) and trustworthiness ratings (honest advisers: $r_{(26)} = 0.42$; $p = .027$; 95% CI [0.05, 0.69]; $BF_{10} = 2.44$; dishonest advisers: $r_{(26)} = 0.77$; $p < .001$; 95% CI [0.56, 0.89]; $BF_{10} = 14{,}166$; advisers without reputation: $r_{(26)} = 0.45$, $p = .017$; 95% CI [0.09, 0.70]; $BF_{10} = 3.53$; Figure 3B).

Finally, trust in the TG did not correlate with winnings in the TAG either for honest, $r_{(26)} = -0.04$ ($p = .854$; 95% CI [−0.40, 0.34]; $BF_{10} = 0.24$) or dishonest advisers, $r_{(26)} = -0.02$ ($p = .917$; 95% CI [−0.39, 0.36]; $BF_{10} = 0.24$), suggesting again that trust in the TG specifically represented an act of reciprocity for the adviser's honesty in the previous interaction. Moreover, the majority of our participants (24 of 28) reported they used a strategy for their decisions, $\chi^2_{(1,\ N\ =\ 28)} = 19.98$ ($p < .001$); however, only four participants believed that their card-choice strategy helped them win more money, $\chi^2_{(1,\ N\ =\ 28)} = 9.82$ ($p = .002$), suggesting that participants were aware that the advice was not useful to increase their payoffs despite its informativeness.

## Discussion

Overall, these results point to a likely common decision-making pattern that individuals adopt when taking advice in everyday life. As our findings show, because checking the quality of one's advice prior to the decision to take the advice is often impossible, an uninformed decision-maker likely grounds her advice-taking strategy exclusively on the adviser's reputation. These results importantly extend previous work showing that less-knowledgeable individuals are more likely to take advice from others (Yaniv & Kleinberger, 2000), as they further demonstrate that uninformed individuals use other criteria, such as the other's character or reputation, to guide their advice-taking and advice-discounting strategies. In particular, participants may have assumed that those they knew had a good reputation would keep behaving consistently with their character, whereas they were reluctant to trust those who previously manifested signs of dishonesty, especially when the other's dishonesty could not be detected anymore. Moreover, that individuals were as likely to take advice from those with an unknown reputation as they were to take advice from honest others

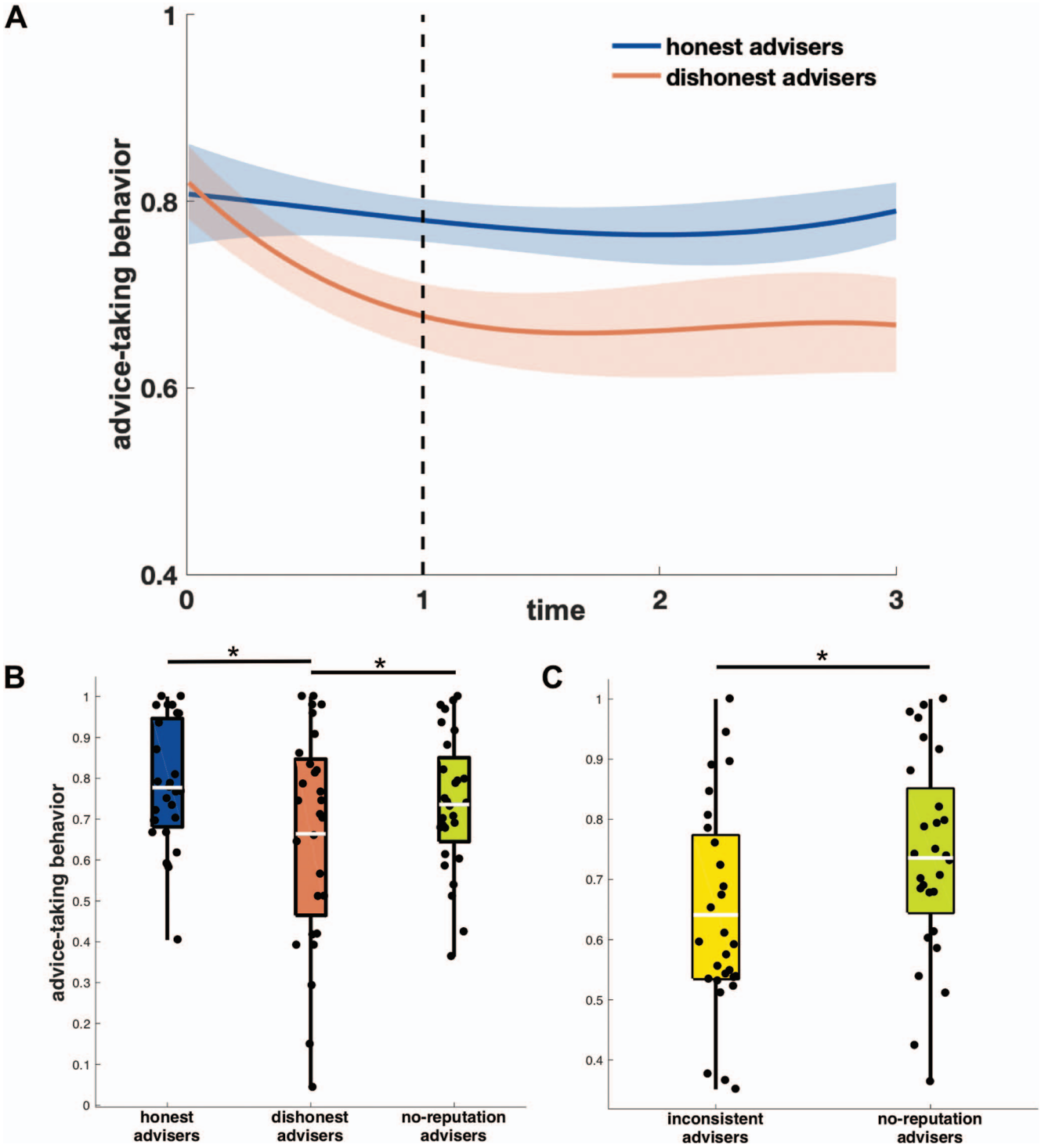


*Figure 4.* Experiment 2. (A) Average advice-taking behavior in the feedback block (until dotted line) and no-feedback blocks (after dotted line) in the TAG (± *SEM* in shadowed area). Data points were interpolated for visualization purposes. (B) Average individual advice-taking behavior in the no-feedback block in the TAG toward honest, dishonest, and advisers without reputation. (C) Differences in advice-taking behaviors toward advisers without reputation in Experiment 2 and inconsistently honest advisers in Experiment 1. $^{*}$ $p < .05$. See the online article for the color version of this figure.

suggests that individuals have positive, initial expectations of others. If such initial expectations of the partner's character are confirmed, no behavioral adaptation will be required. On the contrary, when the default behavior in a social interaction (i.e., trust) is recognized as inappropriate on the basis of what is learnt about the partner, an individual will revise her behavior and slowly start to distrust the partner. This is in line with the descriptive observation that in the very first trials of both Experiment 1 and Experiment 2, advice-taking behavior toward dishonest advisers was as likely as advice-taking behavior toward honest advisers (see Figures 2A and 4A). That is, participants seem to have by default trusted the adviser they played with, revising their trusting behavior only once they realized their trust in the partner was misplaced. These behavioral patterns raise the question as to how individuals dynamically update and revise their beliefs about another's honest character.

## Experiment 3

Being able to readily revise one's beliefs about how likely it is that something would or would not occur is crucial to successfully navigate in a dynamically changing world. In interpersonal interactions, individuals have an even more pressing requirement to constantly update their expectations of others, adapting them to the current partner. In particular, when interacting with a new partner, individuals quickly gather evidence to form beliefs about the partner's character. Accurate character estimations allow good-enough predictions about another person's most likely behavior in new social situations. As shown in the previous experiments, knowing that someone is an honest person allows for a largely truthful expectation that she will behave fairly in a future encounter. However, individuals may change their behavioral attitudes across contexts and time. When this happens, reputational priors about the other person might lose their accuracy and might thus require revision. In Experiment 3, we investigated how reputational beliefs about another person's honesty is revised and updated across time. To this end, we implemented a modified version of the TAG, in which the honesty of the adviser changes over the course of the interaction. In particular, we were interested in testing whether a reputation as an honest partner is more quickly revised than a reputation as a dishonest partner.

### Method

**Participants.** In Experiment 3, we aimed at a sample size similar to Experiment 1 and Experiment 2. The final sample consisted of 33 participants (23 females; 22.27 $\pm >$ 3.13, mean age $\pm$ *SD*). The same exclusion criteria as the previous experiments applied. All participants had normal or corrected-to-normal vision. The study was approved by the local Ethics Committee of the University of Lübeck and conducted in accordance of the Declaration of Helsinki. All participants provided informed consent and received economic compensation for participation. Study length (approximately 1 hr) and hourly wage (€8/hr) were analogous to Experiment 1 and 2 to avoid that external factors such as fatigue or economic compensation for participation might have confounded the results making comparisons between the experiments difficult.

**Procedure.** Here, we developed a new version of the TAG. Participants were invited to the lab and were assigned their role like in the previous two experiments. The role-assignment procedure, however, was a bit different, as participants drew the ball with their role in front of a camera on top of a screen. They were told that each participant was going to draw a ball with their role in front of a camera, which was transferring the videos to the computers in the other participants' rooms. Transparency of the role-assignment procedure was adduced as justification of the video recording. To ensure participant's anonymity, only participants' torsos (up to the chin) were recorded. Through the videos, participants could see the experimenter entering the other rooms with the box and witnessed their coplayers drawing a ball. Participants were also asked not to interact with each other during this procedure. In reality, videos of the other participants were prerecorded. To improve participants' credibility, special attention was given to the experimenter's look that was the same over the entire data collection and matched the overall appearance of the experimenter in the videos.

In this new version of the TAG, participants played with two advisers over the course of three blocks for a total of 132 trials (within-study design) and sharing of truthful information followed a probabilistic rule. In the first block (48 trials), one adviser was predominantly honest sharing truthful information with 75% probability (initially honest adviser), whereas the other was predominantly dishonest sharing truthful information with 25% probability (initially dishonest adviser). In the second block (48 trials), the truthfulness of the advisers' advice reversed with the initially honest adviser being predominantly dishonest (the advice was truthful 25% of the time) and the initially dishonest adviser being predominantly honest (the advice was truthful 75% of the time). Finally, in the last block (36 trials), both advisers behaved in an equally honest fashion sharing truthful information with ~83% probability (30 trials of 36).

As in the previous two experiments, after the TAG, participants played the TG and rated the trustworthiness of the advisers in the TAG. Attractiveness ratings were also introduced as a catch question, analogously to Experiment 2. As expected, attractiveness ratings did not differ between advisers (initially honest vs. initially dishonest advisers: $t_{(32)} = 0.39$; $p = .702$; 95% CI [−0.39, 0.57]; $d = 0.07$; $BF_{10} = 0.20$). At the end of the task, as in Experiment 2, participants reported (with binary response option: Yes/No) whether they used a strategy in the TAG and whether they thought that their strategy was successful to win more money in the game. Timing of the tasks was the same as in the previous two experiments.

**Analyses.** A 3 (block) × 2 (adviser) repeated-measures ANOVA was computed in Experiment 3 to test differences in advice-taking behaviors toward advisers across blocks. *T* tests were employed for post hoc tests and comparisons between advisers in the TG and trustworthiness ratings. Mixed-effects logistic regression analyses followed equation (2). We checked also in Experiment 3 that participants followed a $>$ 5 strategy and an 8 (card number) × 2 (adviser) repeated-measures ANOVA confirmed this ($F_{(7,224)} = 5.86$; $p < .001$; $\eta_p^2 = 0.16$). Finally, a Kolmogorov–Smirnov test confirmed that the distributions of advised card numbers did not change across advisers ($K\text{-}S = 0.09$; $p = .257$).

To mathematically formalize individual learning of others' honest reputation, we fitted 12 different computational models to our data that updated individual learning based on nonsocial informa-

tion, social information or both (*Supplementary methods. Computational models*). Thereby, we maximized the likelihood of the model given participants' advice-taking behavior. Bayesian model comparison was used to compare models (Burnham & Anderson, 2002; Schwarz, 1978). For each participant and type of adviser (honest/dishonest), the individual Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) scores were computed, which are based on the model evidence and penalize models with greater complexity (i.e., greater number of parameters). These scores were then summed across participants to obtain a group BIC and AIC score of the model's goodness of fit (Table S2 in the online supplemental materials). The best model is the one with the lowest group BIC/AIC score. Bayesian model comparison indicated that the winning model was a reinforcement-learning model with two learning rates weighting truthful and untruthful information separately and independently of reward information:

$$\begin{aligned} V_{(t)} &= V_{t-1} + \tau(I_t - V_{t-1})I_t + \delta(I_t - V_{t-1})(1 - I_t) \\ \tau &= \begin{cases} \tau_{honest} & \text{if advice from honest adviser} \\ \tau_{dishonest} & \text{if advice from dishonest adviser} \end{cases} \\ \delta &= \begin{cases} \delta_{honest} & \text{if advice from honest adviser} \\ \delta_{dishonest} & \text{if advice from dishonest adviser} \end{cases} \\ I_t &= \begin{cases} 1 & \text{if truthful information} \\ 0 & \text{if untruthful information} \end{cases}, \end{aligned} \tag{3}$$

where $V_t$ is the subjective value of trusting the adviser's advice on trial $t$, $I_t$ is the type of social information (truthful or untruthful advice) received on trial $t$, $\tau$ is the honesty learning parameter, and $\delta$ is the dishonesty learning parameter. Trial-by-trial subjective values were transformed into trust probabilities using the following stochastic decision rule (i.e., softmax function):

$$p_{trust} = \frac{1}{1 + e^{-\beta(V_{trust} - V_{distrust})}} \tag{4}$$

where $p_{trust}$ is the probability of choosing to trust and $\beta$ is the participant-specific inverse temperature—a free parameter indicating the extent to which participants' choices are deterministic with respect to the values of the available action options, namely, $V_{trust}$ and $V_{distrust}$, which represent the value of choosing to trust (i.e., take the advice) and the value of choosing to distrust (i.e., disregard the advice), respectively.

## Results

A 3 (block) × 2 (adviser) repeated-measures ANOVA on advice-taking behavior revealed an interaction effect between adviser and block, $F_{(2,64)} = 4.27$ ($p = .018$; $\eta_p^2 = 0.12$; $BF_{10} = 10.06$; BF of interaction over main effects = 3.63) and a main effect of adviser, $F_{(1,32)} = 8.72$ ($p < .006$; $\eta_p^2 = 0.21$; $BF_{10} = 25.41$). In the first block, participants took more advice from the initially honest adviser than from the initially dishonest adviser ($M_{honest} = 0.83$, $M_{dishonest} = 0.73$; $t_{(32)} = 3.28$; $p < .003$; 95% CI [0.04, 0.17]; $d = 0.57$; $BF_{10} = 14.39$), whereas in the second block advice-taking behavior did not differ between the two advisers ($M_{honest} = 0.81$, $M_{dishonest} = 0.79$; $t_{(32)} = 1.37$; $p = .181$; 95% CI [−0.01, 0.05]; $d = 0.24$; $BF_{10} = 0.44$; Figure 5A and 5B). This was attributable to a significant increase in trust in the initially dishonest adviser in the second block ($M_{block1} = 0.73$, $M_{block2} = 0.79$; $t_{(32)} = -2.98$; $p < .006$; 95% CI [−0.11, −0.02]; $d = 0.52$; $BF_{10} = 7.29$). On the contrary, advice-taking behavior toward the initially honest adviser did not change significantly across the first and second blocks ($M_{block1} = 0.83$, $M_{block2} = 0.81$; $t_{(32)} = 0.95$; $p = .347$; 95% CI [−0.02, 0.06]; $d = 0.17$; $BF_{10} = 0.28$). Similar advice-taking behavior toward the two advisers was observed in the third block, although participants descriptively took slightly more advice from the initially honest adviser ($M_{honest} = 0.80$, $M_{dishonest} = 0.78$; $t_{(32)} = 0.668$; $p = .509$; 95% CI [−0.04, 0.07]; $d = 0.12$; $BF_{10} = 0.23$).

Across all blocks (Table 1 and Table S1 in the online supplemental materials), participants took significantly more advice from the initially honest adviser ($\beta = 0.32$; $p < .001$; $SE = 0.06$; 95% CI [0.21, 0.43]), even if doing so was not associated with more gains ($\beta = 0.05$; $p = .356$; $SE = 0.06$; 95% CI [−0.06, 0.16]). Further, although the majority of our participants (26 out of 33) reported to have used a strategy for their choice behavior, $\chi^2_{(1, N=33)} = 8.17$ ($p = .004$), only seven reported to believe that their strategy was successful for them to win more money in the game, $\chi^2_{(1, N=33)} = 14.27$ ($p < .001$). These results suggest that a good reputation as honest partner, but not a bad reputation as dishonest partner, impairs belief updating about an other person's character. This was further confirmed by the fact that the initially honest adviser was entrusted with significantly more money in the TG ($M_{honest} = 6.06$, $M_{dishonest} = 5.52$; $t_{(32)} = 2.21$; $p = .034$; 95% CI [0.04, 1.05]; $d = 0.39$; $BF_{10} = 1.57$) and was judged tendentially as more trustworthy ($M_{honest} = 4.27$, $M_{dishonest} = 3.70$; $t_{(32)} = 1.91$; $p = .065$; 95% CI [−0.04, 1.19]; $d = 0.33$; $BF_{10} = 0.93$) than the initially dishonest adviser.

These results suggest a peculiar behavioral pattern (as also shown by Figure 5A). That is, participants flexibly adapted their trusting behavior to the changing trustworthiness of the initially dishonest adviser, whereas trust in the initially honest adviser did not change much across time (although it slightly decreased toward trust levels similar to those observed in Experiment 1 and 2). These behavioral patterns suggest an impairment in learning for the partner with an initial good reputation. To gain insights into the underlying mechanisms that gave rise to this learning asymmetry, we built different computational models that captured the learning dynamics related to the integration of social information from the advisers. The best model was the one described in Equation (3). This model explained individual advice-taking behavior well, as confirmed by a mixed-effects logistic regression using estimated choice probability to predict trial-by-trial choices ($\beta = 4.53$; $p < .001$; $SE = 0.21$; 95% CI [4.12, 4.95]), and could closely mimic the observed behavioral patterns of advice taking across blocks (Interaction Effect Adviser × Block based on model-based estimations of advice-taking behavior: $F_{(2,64)} = 5.88$; $p < .005$; $\eta_p^2 = 0.16$; BF of interaction effect over main effects = 2.24).

A 2 (learning rate) × 2 (adviser) ANOVA on model parameters revealed a significant interaction effect, $F_{(1,32)} = 8.67$ ($p < .006$; $\eta_p^2 = 0.21$; BF of interaction effect over main effects = 3.60; Figure 5C). In particular, when the advice had been given by the initially honest adviser, truthful advice was valued significantly more than untruthful advice ($M_\tau = 0.25$; $M_\delta = 0.10$; $t_{(32)} = 2.99$; $p < .006$; 95% CI [0.05, 0.25]; $d = 0.52$; $BF_{10} = 7.43$). On the contrary, when the advice had been given by the initially dishonest adviser, truthful advice was valued tendentially less than untruthful advice although not significantly so ($M_\tau = 0.08$; $M_\delta = 0.13$; $t_{(32)} = -0.80$; $p =$

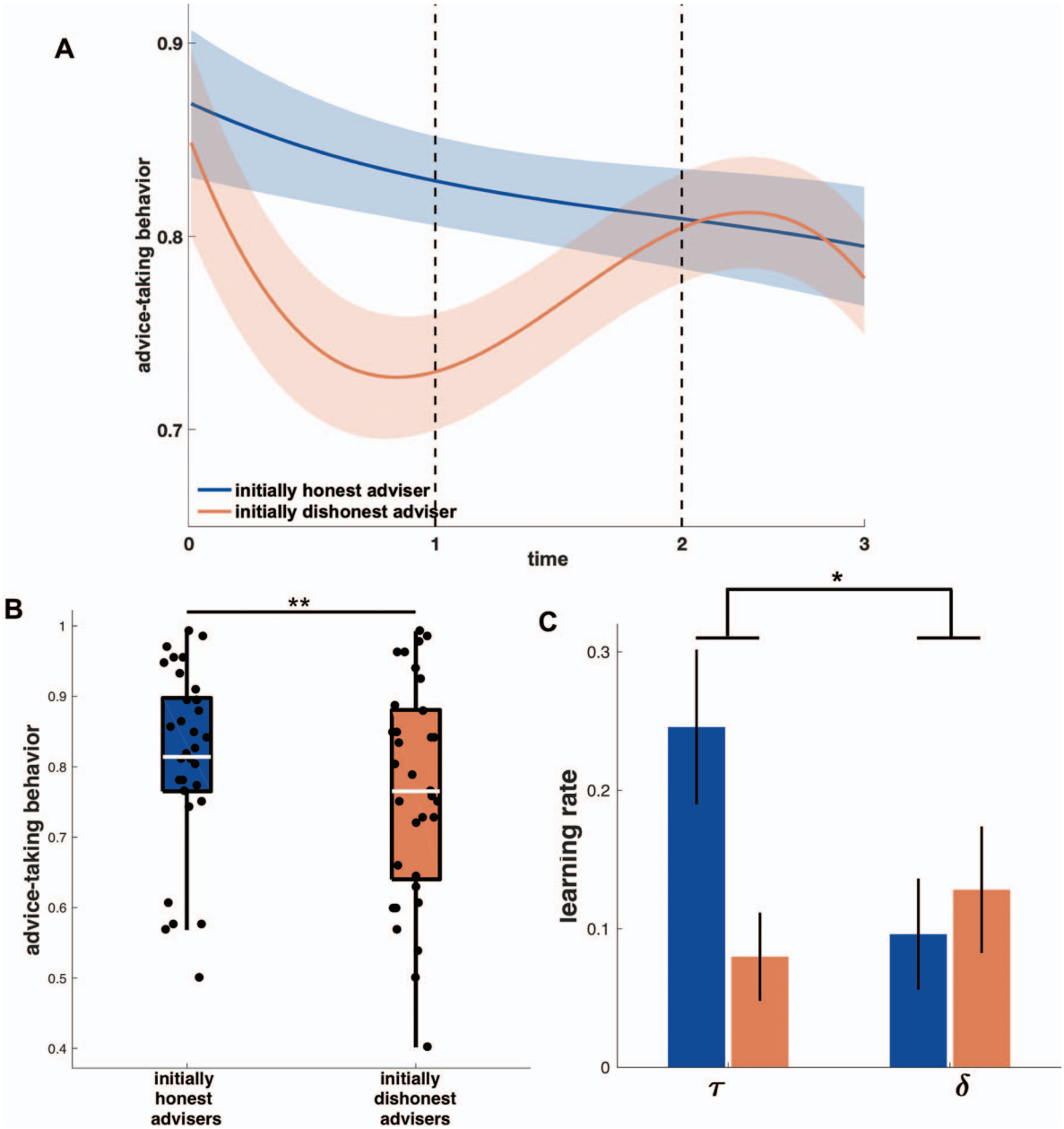


*Figure 5.* Experiment 3. (A) Average advice-taking behavior in the TAG over time (i.e., blocks) toward initially honest and initially dishonest advisers (± *SEM* in shadowed area). Data points were interpolated for visualization purposes. Dotted lines represent the three different blocks in the task. In the first block, the initially honest adviser was predominantly trustworthy, whereas the initially dishonest adviser was predominantly untrustworthy. In the second block (between Time 1 and Time 2), the advisers' trustworthiness reversed. Although participants nicely tracked this change in trustworthiness in the initially dishonest adviser (with steadily increasing trust in their advice), they did not do so for the initially honest adviser. In the last block, advisers' trustworthiness was comparable. (B) Average individual advice-taking behavior toward honest and dishonest advisers across all blocks. (C) Honesty (τ) and dishonesty (δ) learning rates for initially honest and initially dishonest advisers. $^{*}$ $p < .05$. $^{**}$ $p < .01$. See the online article for the color version of this figure.

.432; 95% CI [−0.17, 0.08]; $d = 0.14$; $BF_{10} = 0.25$). In addition, truthful advice from the initially honest adviser was valued significantly more than truthful advice from the initially dishonest adviser ($t_{(32)} = 2.84$; $p < .008$; 95% CI [0.05, 0.29]; $d = 0.49$; $BF_{10} = 5.38$). These results indicate that honesty impairs belief updating about another person's reputation due to a reputation-dependent, asymmetric weight on the information provided by honest advisers.

## Discussion

Results from Experiment 3 replicated and extended the behavioral patterns observed in Experiment 1 and 2. First, participants were seen to initially trust both advisers. However, after a couple of trials, when they realized that their trust in the dishonest adviser was misplaced, they immediately adapted their behavior, discounting advice from the dishonest partner. However, in Experiment 3 we further observed that the same did not happen for the adviser that could establish a reputation as an honest partner over the course of the first block. This suggests that participants were integrating information from the two advisers differently over the course of the social interaction. In particular, at the beginning of the social interaction, when no reputational knowledge about the other was yet available, all information was likely integrated in a similar fashion for both advisers. However, once the initial positive expectation that the partner would be trustworthy was confirmed over the first period of exchanges, participants might have relaxed the integration of new incoming information about the current reputation of the initially honest adviser. Thus, like participants' behavior in the no-feedback blocks of Experiment 2, participants in Experiment 3 kept trusting the initially honest adviser over the course of the entire social interaction as if they did not receive feedback about the other's honesty any longer. On the contrary, when the initial positive expectation that the partner would behave in a trustworthy manner was not confirmed in the first interactions, participants might have readily updated their belief about the character of the dishonest partner and more closely tracked her trial-by-trial decisions, allowing for optimal behavior revision over the course of the social exchange.

This is further supported by our results on how individuals valued the advisers' advice. In fact, even though untruthful advice was valued observationally more when it came from the initially dishonest adviser, the difference between honesty and dishonesty learning parameters was not significant for the initially dishonest adviser. The fact that truthful and untruthful advice from the initially dishonest adviser was valued in a similar fashion explains why participants could readily update their belief about the initially dishonest adviser in the second block when the adviser changed behavior and became trustworthy. Similarly, the weighting asymmetry in information integration observed for the initially honest adviser explains why our participants did not revise their belief about the character of the initially honest adviser. In particular, stronger valuation of truthful information confirming the initial impression that the other is trustworthy might have impaired a flexible behavior revision. These findings thus suggest that an asymmetry in how social information is valued biases integration of information received by others, thereby impairing a successful revision of one's beliefs and behavior.

# General Discussion

In the current study, we showed how individuals form beliefs about others' reputation and how these beliefs inform behaviors across contexts. Our results reveal that a good reputation predicts trusting behavior across contexts and even in the absence of feedback about the other's current behavior. In particular, consistently and inconsistently honest social partners were trusted more and perceived as more trustworthy than dishonest partners. However, those whose reputation was unknown were trusted more than inconsistently honest others, suggesting that signs of dishonesty negatively impact one's initial positive expectations of others' trustworthiness. Further, confirmation of these initial positive expectations produced strong trustworthiness impressions of honest others that impaired social learning and adaptive, flexible behavior. This learning impairment was explained by a weighting asymmetry for information from partners with an established reputation as honest partners.

## The Importance of Informative Advice

Information gathering is central to form better models of the world. Building better models of the world reduces surprise about future events and improves prediction accuracy, boosting an animal's survival chances (Badcock, Friston, & Ramstead, 2019; Bang & Frith, 2017). In social interactions, most of the information derives from other social partners, for instance, in the form of advice. It has been noted that people show a social bias in information gathering, preferring social information over nonsocial information (Biele et al., 2009, 2011). However, to improve one's decisional accuracy, it is of pivotal importance that individuals take advice from the right person. One way individuals have to make rough estimations about the quality of a piece of advice is through the reputation of the adviser. To our knowledge, no study had so far investigated how an adviser's reputation is learnt and valued for advice utilization.

Our results suggest that when individuals have no reasons to believe the other to be untrustworthy or dishonest, they prefer taking the other's advice. These results are in line with previous studies showing that individuals prefer options that are either preferred or suggested by others (Biele et al., 2009; Mahmoodi et al., 2018). This preference may turn out extremely advantageous to the individual, as exploiting others' knowledge prevents from applying more costly and riskier exploratory strategies (Kendal et al., 2018). As a good reputation of being trustworthy and honest is built on truthful information shared by others, having a good reputation may work as a proxy for the quality of the other's information (Gordon & Spears, 2012). Thus, preferring the advice of honest others may disclose the attempt to improve decisional accuracy and reduce uncertainty by using likely truthful information. This is consistent with people's tendency to rely on others' advice in highly uncertain situations, for instance, when they are not completely informed or when they perceive the environment as unpredictable (McElreath et al., 2005; Sniezek & Buckley, 1995; Sniezek, May, & Sawyer, 1990; Van Swol & Sniezek, 2005).

Moreover, even though the advice was not useful to gain more benefits, our participants still showed a preference for truthful advice. Experimental paradigms in previous studies did not carefully control for the confound between reward information about one's gains and social information about another's character, as an advice had often had the form of a piece of advice on the best option in the task, generally associated with higher rewards (Behrens et al., 2008; Biele et al., 2009; Diaconescu et al., 2014; Rodriguez Buritica et al., 2019). Thus, it was yet unclear whether an advice is taken because a decision-maker has learnt that that advice is associated with higher rewards or because of other reasons. Our results suggest that uninformed decision-makers prefer taking informative advice even though doing so does not bring them proximal benefits. Because our participants were aware that

their strategies were not successful in the game, this preference may reflect a form of reciprocity to the honest other for sharing informative advice. This is further supported by the fact that the degree to which participants took the other's advice correlated with participants' direct reciprocity in a subsequent interaction, that is, with their willingness to entrust money with the adviser. These results provide evidence that in a social exchange, an individual's behavior is motivated by concerns related to compliance with social norms (such as the social norms of fairness and reciprocity). This is line with previous work showing that individuals tend not to reject freely offered advice or prefer discounting advice from those who do not comply with a norm of reciprocity regardless of the actual quality of the advice (Mahmoodi et al., 2018; Sniezek & Buckley, 1995).

## Signs of Dishonesty Corrupt Positive Expectations of Others

In cases in which individuals do not have information about the other's reputation or the quality of the other's advice (e.g., when interacting with strangers), individuals may use social norms to infer the likely future behavior of the interacting partner (Bicchieri, 2005). They may thus assume that the other person complies with social norms of fairness and equity, making a broad array of social behaviors possible, such as helping, generosity, and trust. For instance, although trust has been associated with a strong betrayal aversion, because the trustee may behave selfishly and betray the trustor (Aimone & Houser, 2011, 2013; Bohnet, Greig, Herrmann, & Zeckhauser, 2008; Bohnet & Zeckhauser, 2004), expectations of a compliant behavior help the trustor overcome the aversive feelings associated with the possibility of being betrayed, facilitating trust in the partner (Baumgartner et al., 2009; Masuda & Nakamura, 2012; van 't Wout & Sanfey, 2008). In our study, when interacting with advisers of unknown reputation, participants likely assumed that the partner had good intentions when giving advice and trusted the received information from the unknown partner.

On the contrary, when participants had the opportunity to interact with the other over the course of multiple exchanges, they could form beliefs about the partner's reputation based on feedback about the quality of the partner's advice. Reputation can be thought of as a prior that allows good-enough estimations of another's behavior (Fehr & Fischbacher, 2003). Reputation is especially important for inferences on others' behaviors and intentions in new social interactions, as individuals believe that someone is likely to behave in the present situation as she did in the past (Milinski, Semmann, & Krambeck, 2002; Semmann, Krambeck, & Milinski, 2004; Wedekind & Braithwaite, 2002). Consistently with that, our results show that a good reputation as honest partner in advice giving does not only influence advice-taking behaviors in the current situation, but also generalizes to social behaviors in subsequent interactions. Thus, based on the other's honest behavior, individuals formed trustworthiness impressions that informed their future behavior toward the partner.

Crucially, signs of dishonesty were immediately integrated into one's reputational priors. Thus, when individuals realized that their trust was misplaced, they revised their behavior, discounting the advice of dishonest social partners and entrusting less money to them in a subsequent interaction. Surprisingly, showing at the same time some signs of honesty was not enough to regain full trust. In fact, participants' trust in advisers without reputation was higher as compared with their trust in inconsistently honest advisers. These findings suggest that although honesty has a higher value in building and maintaining trust than simple behavioral predictability (e.g., being predictably dishonest; Jones & George, 1998), signs of dishonesty have deleterious consequences on an individual's positive, initial expectations of others' behaviors, as a person may not be able to regain the other's entire trust once it is lost (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). A possible explanation for this difficulty in regaining trust after signs of dishonesty might be that individuals attribute strong moral labels to the other person, which have been shown to deeply affect approach-avoidance tendencies (Fiske et al., 2007; Peeters, 2002).

## Good Reputation Impairs Learning

However, this does not seem to happen if someone first has the opportunity to build a good reputation. In particular, a reputation of being honest impaired participants' ability to learn and update trustworthiness impressions of the other person on the basis of new incoming information. That is, once an individual has formed a belief that someone is trustworthy, trust does not easily break down because of signs of dishonesty. This suggests that individuals might keep trusting someone who has ceased to be trustworthy because reputation signals behavioral consistency. Thus, participants may have thought that a trustworthy behavior is more consistent with the character of a person who has a reputation of being honest and tried to explain away inconsistencies in the other's current actions on the basis of her reputation. This is similar to the psychological mechanism that has been observed when individuals rationalize inconsistent policy contents on the basis of party membership (Cohen, 2003).

However, a similar behavioral rigidity was not observed for dishonest others. In this case, individuals were seen to readily update their behavior, adapting it to the current trustworthiness of the initially dishonest partner. A reinforcement-learning model suggests that these different behavioral patterns hinged on an asymmetry in information weighting that impaired reputational learning for honest but not dishonest others. In particular, our participants valued truthful and untruthful advice from the initially honest partner in a significantly different fashion. On the contrary, no differences were observed in information valuation for the initially dishonest partner. Specifically, truthful advice from the initially honest adviser was, on the one hand, valued more than untruthful advice. On the other hand, it was also valued more than truthful advice from the initially dishonest other. This asymmetry indicates a positivity bias in integration of information from honest others, which has likely led participants to disregard information inconsistent with their priors about the partner with a good reputation. In turn, such biased information valuation might have contributed to the observed resistance to behavior change, which chimes with previous work showing that supportive communicated beliefs reduce the exploration of options other than the supported one (Pilditch & Custers, 2018).

These findings provide an explanation for perceptual and judgmental biases in different domains. For instance, recent work has indicated a perceptual bias that makes individuals perceive young Black men as more threatening (Wilson, Hugenberg, & Rule, 2017). As perceptual evidence is integrated on the basis of a

reinforcement learning mechanism (Badcock et al., 2019), such bias can be likely due to an asymmetry in perceptual information weighting that distorts judgmental responses. Similarly, social phenomena such as the "do-gooder derogation" or "self-licensing" might involve a similar asymmetry in information valuation (Merritt, Effron, & Monin, 2010; Minson & Monin, 2012; Monin, Sawyer, & Marquez, 2008). For instance, past good deeds might license to present, morally dubious actions because those past good deeds have contributed to the formation of a good reputation that leads others to more strongly discount evidence inconsistent with that reputation (Merritt et al., 2010).

These results further indicate that individuals, when uninformed, might be more susceptible to the other's good reputation. This particular bias may have gone overlooked in previous studies, especially in advice-giving paradigms, for participants in these tasks are often provided with some information about the structure of the task that allows them to weight the quality of the other's advice during the decisional process (Diaconescu et al., 2014; Mahmoodi et al., 2018; Yaniv & Kleinberger, 2000). With this information at hand, participants may have been able to make more informative decisions, reducing any positivity bias. On the contrary, in our paradigm, participants could judge the quality of the other's advice only during the feedback phase, after a decision was made. This implies that the honest character of the adviser, inferred during repeated past interactions, was the only information participants had to judge the advice's quality when making their choices during the decision phase. This dynamic closely resembles real-life scenarios, in which individuals often need to blindly rely on another person's advice because uninformed and thus unable to weight its quality before deciding whether to take or discount it. For instance, when facing a legal or medical issue without knowledge of the law system or medicine, the reputation of the lawyer or the doctor may be the only available information to decide whether to follow their advice. In contexts like this, an uninformed decision-maker may be more susceptible to positive reputational influences because she lacks knowledge that could serve as an anchor for her decisions. In line with this, previous evidence has shown that naïve judges who have not committed to a position before receiving advice are more prone to accept the advice (Sniezek & Buckley, 1995).

Finally, our computational model provides a general and simple, mechanistic explanation of the biased learning processes that contribute to belief rigidity and resistance to behavior change. The processing dynamics unearthed by our computational analyses further concur with previous neuroimaging evidence, providing a possible neurobiological model of biased information sampling and processing. For instance, previous work has suggested that decreased activity in the orbitofrontal cortex is associated with stronger resistance to political belief change, suggesting that reduced revision of one's political beliefs is associated with decreased valuation of counterevidence (Kaplan, Gimbel, & Harris, 2016). On the contrary, stronger weight on expert or honest advice is related to increased activity in the orbitofrontal cortex, suggesting that increased advice utilization depends on the extent to which individuals value the advice (Bellucci, Molter, & Park, 2019; Meshi et al., 2012). As the orbitofrontal cortex undertakes an important role in reinforcement learning (Gottfried & Dolan, 2004; Tsuchida et al., 2010), this region may be crucial to how individuals value information inconsistent with one's priors for belief updating and learning.

## Conclusions

Taken together, our results demonstrate that honesty is a central determinant of trustworthiness impressions and trusting behavior across contexts. Furthermore, we show that different learning mechanisms underlie integration of social information from honest and dishonest others. In particular, participants did not revise their trustworthiness impressions of those with an established good reputation, whereas they successfully tracked changes in trustworthiness of those with a bad reputation.

These patterns may be an unfortunate consequence or a byproduct of a learning-strategy optimization in social interactions that has led to the rise and need of reputational priors. For instance, developing a successful strategy to quickly identify and continuously keep track of the behavior of free-riders is of pivotal importance to an individual's survival chances, as free-riders may exploit an individual at every time and point, jeopardizing her existence. However, keeping track of the intentions and motives of every single action of our social fellows implies enormous energy costs that are unsustainable on the long run. Thus, we may speculate that reputation has evolved as a tool for efficient resource distribution in social behavior control. In this regard, using a reputational tag to quickly identify who deserves our trust and who does not is a simple but efficient solution to focus energy and resources on a limited number of interactions with a limited number of individuals in which the risk of exploitation is more likely. It follows that an individual might relax her alertness during the safer interactions with trustworthy group members, while maintaining high alertness during interactions with untrustworthy others to detect signs of exploitation. Social norms might further aid individuals during this efficient resource distribution to track others' behavior, as social norms establish general criteria for acceptable behaviors in a group that support quick recognition of norm-deviant behaviors for belief formation about the other's character.

Nonetheless, even though this simple solution works well in general for a successful maximization of an animal's resource utilization during social interactions, it bears the danger of biased estimations when judging the behavior of those with an established good reputation. In particular, when the current behavior of these social partners does not reflect their good reputation any longer, people may dangerously be at the mercy of the partner's exploitation, since the ability to optimally revise one's behavioral attitudes is impaired.

Finally, our study provides evidence that when deciding whether to take advice, individuals integrate different types of information. First, we were able to disentangle the confound in the previous literature between personal gains and social information about others, showing that social characteristics of the adviser (such as her reputation and trustworthiness) guide advice-taking behaviors beyond the personal benefits derived from the decision to take a piece of advice. Second, we demonstrated that these social characteristics of the adviser have even a stronger influence on participants' decisions than the adviser's actual accuracy and competence (McGinnies & Ward, 1980), impairing participants' ability to optimally revise their advice-taking behavior. Future studies are

needed to investigate whether and how this learning impairment for partners with a good reputation can be overcome. For instance, another trustworthy source might provide competing information about the intentions and behavior of the partner with a good reputation that might help individuals revise their beliefs and so form more accurate impressions of the other (Hertz et al., 2017). Hence, it is of pivotal importance for future investigations on advice-taking behaviors to take these social aspects of advisers into consideration.

By providing novel insights into the learning dynamics that underlie the formation of reputational priors relevant for social behaviors, we believe that this work contributes to a better understanding of how interpersonal interactions unfold, pointing to new testable hypotheses for future investigations in a wide range of scientific fields from psychology and experimental economics to political sciences and neuroscience.

## References


Aimone, J. A., & Houser, D. (2011). Beneficial betrayal aversion. *PLoS ONE, 6,* e17725. http://dx.doi.org/10.1371/journal.pone.0017725

Aimone, J. A., & Houser, D. (2012). What you don't know won't hurt you: A laboratory analysis of betrayal aversion. *Experimental Economics, 15,* 571–588. http://dx.doi.org/10.1007/s10683-012-9314-z

Aimone, J. A., & Houser, D. (2013). Harnessing the benefits of betrayal aversion. *Journal of Economic Behavior & Organization, 89,* 1–8. http://dx.doi.org/10.1016/j.jebo.2013.02.001

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics, 9,* 193–208. http://dx.doi.org/10.1007/s10683-006-9122-4

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11,* 150–166. http://dx.doi.org/10.1177/1088868306294907

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86,* 356–366. http://dx.doi.org/10.1037/0022-3514.86.2.356

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*. Advance online publication. http://dx.doi.org/10.1016/j.plrev.2018.10.002

Baker, R., Honeyford, K., Levene, L. S., Mainous, A. G., III, Jones, D. R., Bankart, M. J., & Stokes, T. (2016). Population characteristics, mechanisms of primary care and premature mortality in England: A cross-sectional study. *British Medical Journal Open, 6,* e009981. http://dx.doi.org/10.1136/bmjopen-2015-009981

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science, 4,* 170193. http://dx.doi.org/10.1098/rsos.170193

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games. *European Journal of Psychological Assessment, 30,* 178–192. http://dx.doi.org/10.1027/1015-5759/a000183

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron, 64,* 756–770. http://dx.doi.org/10.1016/j.neuron.2009.11.017

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America, 114,* E5070–E5076. http://dx.doi.org/10.1073/pnas.1615978114

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

Bellucci, G., Molter, F., & Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. *Nature Communications, 10,* 5184. http://dx.doi.org/10.1038/s41467-019-13261-8

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 122–142. http://dx.doi.org/10.1006/game.1995.1027

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511616037

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science, 33,* 206–242. http://dx.doi.org/10.1111/j.1551-6709.2009.01010.x

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

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *The American Economic Review, 98,* 294–310. http://dx.doi.org/10.1257/aer.98.1.294

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization, 55,* 467–484. http://dx.doi.org/10.1016/j.jebo.2003.11.004

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,* 127–151. http://dx.doi.org/10.1016/j.obhdp.2006.07.001

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford, UK: Oxford University Press.

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica, 104,* 371–398. http://dx.doi.org/10.1016/S0001-6918(00)00037-8

Burnham, K. P., & Anderson, D. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin, Germany: Springer.

Burt, R. S., & Knez, M. (1995). Kinds of Third-Party Effects on Trust. *Rationality and Society, 7,* 255–292. http://dx.doi.org/10.1177/1043463195007003003

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76,* 1–29. http://dx.doi.org/10.18637/jss.v076.i01

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61,* 87–105. http://dx.doi.org/10.1016/j.cogpsych.2010.03.001

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology, 85,* 808–822. http://dx.doi.org/10.1037/0022-3514.85.5.808

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior, 31,* 73–98. http://dx.doi.org/10.1016/j.riob.2011.10.004

Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes, 112,* 11–23. http://dx.doi.org/10.1016/j.obhdp.2009.11.007

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8,* 1611–1618. http://dx.doi.org/10.1038/nn1575

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., . . . Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology, 10,* e1003810. http://dx.doi.org/10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience, 12,* 618–634. http://dx.doi.org/10.1093/scan/nsw171

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology, 87,* 611–628. http://dx.doi.org/10.1037/0021-9010.87.4.611

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior, 62,* 287–303. http://dx.doi.org/10.1016/j.geb.2007.06.001

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425,* 785–791. http://dx.doi.org/10.1038/nature02043

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140. http://dx.doi.org/10.1177/001872675400700202

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11,* 77–83. http://dx.doi.org/10.1016/j.tics.2006.11.005

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *The Journal of Neuroscience, 33,* 3602–3611. http://dx.doi.org/10.1523/JNEUROSCI.3086-12.2013

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37,* 1673–1682. http://dx.doi.org/10.1016/S0042-6989(96)00286-6

Gordon, R., & Spears, K. (2012). You don't act like you trust me: Dissociations between behavioural and explicit measures of source credibility judgement. *The Quarterly Journal of Experimental Psychology, 65,* 121–134. http://dx.doi.org/10.1080/17470218.2011.591534

Gottfried, J. A., & Dolan, R. J. (2004). Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. *Nature Neuroscience, 7,* 1144–1152. http://dx.doi.org/10.1038/nn1314

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic, 29,* 337. http://dx.doi.org/10.22329/il.v29i4.2903

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science, 40,* 1496–1533. http://dx.doi.org/10.1111/cogs.12276

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,* 2191. http://dx.doi.org/10.1038/s41467-017-02314-5

Hilbig, B. E., Thielmann, I., Hepp, J., Klein, S. A., & Zettler, I. (2015). From personality to altruistic behavior (and back): Evidence from a double-blind dictator game. *Journal of Research in Personality, 55,* 46–50. http://dx.doi.org/10.1016/j.jrp.2014.12.004

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLoS Computational Biology, 14,* e1005935. http://dx.doi.org/10.1371/journal.pcbi.1005935

JASP Team. (2019). JASP (Version 0.11.1) [Computer software]. Retrieved from https://jasp-stats.org

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review, 23,* 531–546. http://dx.doi.org/10.5465/amr.1998.926625

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports, 6,* 39589. http://dx.doi.org/10.1038/srep39589

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences, 22,* 651–665. http://dx.doi.org/10.1016/j.tics.2018.04.003

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology, 89,* 104–118. http://dx.doi.org/10.1037/0021-9010.89.1.104

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308,* 78–83. http://dx.doi.org/10.1126/science.1108062

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50,* 569–598. http://dx.doi.org/10.1146/annurev.psych.50.1.569

Lee, K., & Ashton, M. C. (2004). Psychometric Properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39,* 329–358. http://dx.doi.org/10.1207/s15327906mbr3902_8

Levine, E. E., Bitterly, T. B., Cohen, T. R., & Schweitzer, M. E. (2018). Who is trustworthy? Predicting trustworthy intentions and behavior. *Journal of Personality and Social Psychology, 115,* 468–494. http://dx.doi.org/10.1037/pspi0000136

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes, 126,* 88–106. http://dx.doi.org/10.1016/j.obhdp.2014.10.007

Li, J., Xiao, E., Houser, D., & Montague, P. R. (2009). Neural responses to sanction threats in two-party economic exchange. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 16835–16840. http://dx.doi.org/10.1073/pnas.0908855106

Madsen, J. K. (2016). Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 165–170). Austin, TX: Cognitive Science Society.

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature Communications, 9,* 2474. http://dx.doi.org/10.1038/s41467-018-04925-y

Masuda, N., & Nakamura, M. (2012). Coevolution of trustful buyers and cooperative sellers in the trust game. *PLoS ONE, 7,* e44169. http://dx.doi.org/10.1371/journal.pone.0044169

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20,* 709–734. http://dx.doi.org/10.5465/amr.1995.9508080335

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization, 52,* 267–275. http://dx.doi.org/10.1016/S0167-2681(03)00003-9

McElreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., . . . Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior, 26,* 483–508. http://dx.doi.org/10.1016/j.evolhumbehav.2005.04.003

McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin, 6,* 467–472. http://dx.doi.org/10.1177/014616728063023

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25,* 1106–1115. http://dx.doi.org/10.1177/0956797614524255

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology*

*Compass, 4,* 344–357. http://dx.doi.org/10.1111/j.1751-9004.2010.00263.x

Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PLoS ONE, 7,* e49748. http://dx.doi.org/10.1371/journal.pone.0049748

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons. *Nature, 415,* 424–426. http://dx.doi.org/10.1038/415424a

Minson, J. A., & Monin, B. (2012). Do-gooder derogation. *Social Psychological and Personality Science, 3,* 200–207. http://dx.doi.org/10.1177/1948550611415695

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology, 95,* 76–93. http://dx.doi.org/10.1037/0022-3514.95.1.76

Nelson, W. R., Jr. (2002). Equity or intention: It is the thought that counts. *Journal of Economic Behavior & Organization, 48,* 423–430. http://dx.doi.org/10.1016/S0167-2681(01)00245-1

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Peeters, G. (2002). From good and bad to can and must: Subjective necessity of acts associated with positively and negatively valued stimuli. *European Journal of Social Psychology, 32,* 125–136. http://dx.doi.org/10.1002/ejsp.70

Pereira Gray, D. J., Sidaway-Lee, K., White, E., Thorne, A., & Evans, P. H. (2018). Continuity of care with doctors-a matter of life and death? A systematic review of continuity of care and mortality. *British Medical Journal Open, 8,* e021161. http://dx.doi.org/10.1136/bmjopen-2017-021161

Pilditch, T. D., & Custers, R. (2018). Communicated beliefs about action-outcomes: The role of initial confirmation in the adoption and maintenance of unsupported beliefs. *Acta Psychologica, 184,* 46–63. http://dx.doi.org/10.1016/j.actpsy.2017.04.006

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rode, J. (2010). Truth and trust in communication: Experiments on the effect of a competitive context. *Games and Economic Behavior, 68,* 325–338. http://dx.doi.org/10.1016/j.geb.2009.05.008

Rodriguez Buritica, J. M., Heekeren, H. R., & van den Bos, W. (2019). The computational basis of following advice in adolescents. *Journal of Experimental Child Psychology, 180,* 39–54. http://dx.doi.org/10.1016/j.jecp.2018.11.019

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23,* 393–404. http://dx.doi.org/10.5465/amr.1998.926617

Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S., & Murray, E. A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nature Neuroscience, 16,* 1140–1145. http://dx.doi.org/10.1038/nn.3440

Sah, S., Loewenstein, G., & Cain, D. M. (2013). The burden of disclosure: Increased compliance with distrusted advice. *Journal of Personality and Social Psychology, 104,* 289–304. http://dx.doi.org/10.1037/a0030527

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes, 101,* 1–19. http://dx.doi.org/10.1016/j.obhdp.2006.05.005

Semmann, D., Krambeck, H.-J., & Milinski, M. (2004). Strategic investment in reputation. *Behavioral Ecology and Sociobiology*. Advance online publication. http://dx.doi.org/10.1007/s00265-004-0782-9

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes, 62,* 159–174. http://dx.doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., May, D. R., & Sawyer, J. E. (1990). Social uncertainty and interdependence: A study of resource allocation decisions in groups. *Organizational Behavior and Human Decision Processes, 46,* 155–180. http://dx.doi.org/10.1016/0749-5978(90)90027-7

Staudinger, M. R., & Buechel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *NeuroImage, 76,* 125–133. http://dx.doi.org/10.1016/j.neuroimage.2013.02.074

Thielmann, I., & Hilbig, B. E. (2015). The traits one can trust: Dissecting reciprocity and kindness as determinants of trustworthy behavior. *Personality and Social Psychology Bulletin, 41,* 1523–1536. http://dx.doi.org/10.1177/0146167215600530

Tsuchida, A., Doll, B. B., & Fellows, L. K. (2010). Beyond reversal: A critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. *The Journal of Neuroscience, 30,* 16868–16875. http://dx.doi.org/10.1523/JNEUROSCI.1958-10.2010

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

van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108,* 796–803. http://dx.doi.org/10.1016/j.cognition.2008.07.002

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology, 12,* 1012–1015. http://dx.doi.org/10.1016/S0960-9822(02)00890-4

Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology, 113,* 59–80. http://dx.doi.org/10.1037/pspi0000092

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

Yaniv, I. (2016). The benefit of additional opinions. *Current Directions in Psychological Science, 13,* 75–78. http://dx.doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83,* 260–281. http://dx.doi.org/10.1006/obhd.2000.2909

Zhao, K., & Smillie, L. D. (2015). The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review, 19,* 277–302. http://dx.doi.org/10.1177/1088868314553709

Received March 31, 2019
Revision received November 20, 2019
Accepted November 21, 2019 ■