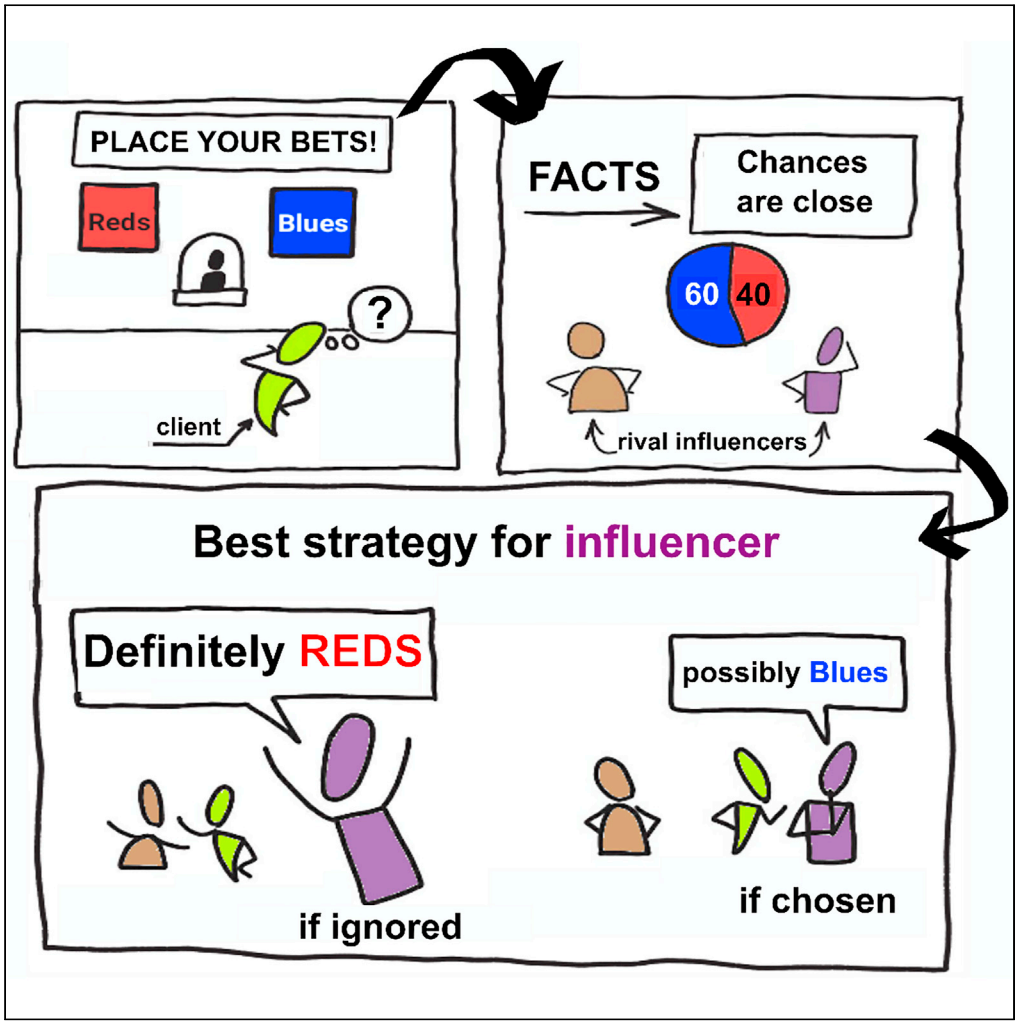
iScience

CellPress
OPEN ACCESS

Article

Strategic disinformation outperforms honesty in competition for social influence



Ralf H.J.M. Kurvers, Uri Hertz, Jurgis Karpus, Marta P. Balode, Bertrand Jayles, Ken Binmore, Bahador Bahrami

kurvers@mpib-berlin.mpg.de

Highlights

We investigate the conditions under which disinformation gains social influence

Game theory shows that disinformation pays for advisers when they are ignored

Such a strategic adviser outcompeted an honest one in swaying human participants

Individuals, communicating dyads, and majority groups followed a strategic adviser

Kurvers et al., iScience *24*, 103505
December 17, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.103505

iScience

CellPress
OPEN ACCESS

Article

Strategic disinformation outperforms honesty in competition for social influence

Ralf H.J.M. Kurvers,[1,7,8,*] Uri Hertz,[2,7] Jurgis Karpus,[3,4,7] Marta P. Balode,[1] Bertrand Jayles,[1] Ken Binmore,[5] and Bahador Bahrami[1,3,6]

SUMMARY

Competition for social influence is a major force shaping societies, from baboons guiding their troop in different directions, to politicians competing for voters, to influencers competing for attention on social media. Social influence is invariably a competitive exercise with multiple influencers competing for it. We study which strategy maximizes social influence under competition. Applying game theory to a scenario where two advisers compete for the attention of a client, we find that the rational solution for advisers is to communicate truthfully when favored by the client, but to lie when ignored. Across seven pre-registered studies, testing 802 participants, such a strategic adviser consistently outcompeted an honest adviser. Strategic dishonesty outperformed truth-telling in swaying individual voters, the majority vote in anonymously voting groups, and the consensus vote in communicating groups. Our findings help explain the success of political movements that thrive on disinformation, and vocal underdog politicians with no credible program.

INTRODUCTION

Social influence is a fundamental organizing principle across human and non-human societies (Conradt, 2012; Conradt and List, 2009; Turner, 1991). Social influence is invariably a competitive exercise because the influencer is rarely in a one-to-one relationship with their potential followers whose choices they wish to influence. Instead, they have to compete with others to gain and maintain influence. From baboons competing to guide their troop to different preferred directions (Strandburg-Peshkin et al., 2015, 2017), and fish competing for directing their shoal to different preferred food sources (Couzin et al., 2011; Miller et al., 2013), to politicians competing for voters (Grossman and Helpman, 1996; Kitschelt, 2000), consultancy firms competing for harking in clients (McAfee and Brynjolfsson, 2012; Pine and Gilmore, 1998), and social influencers competing for "likes" and followers on social media platforms (Lorenz-Spreen et al., 2019; Weng et al., 2012), multiple influencers compete for gaining social influence. Approaching the process of persuasion and influence from a competitive viewpoint is important because the influencers' ultimate goal is often not to provide the best information or service for their clients, but to outcompete their rivals (e.g., to become the leader of the group or to gain political power).

We identify three hallmarks of competition for social influence: information asymmetry, delegation of future decisions, and intractable uncertainty. Information asymmetry occurs when influence seekers (e.g., politicians or advisers) know more about an issue than do the people they seek to influence (e.g., voters or clients) (Healy and Palepu, 2001). For example, in the political arena, the issues at stake are often multidimensional and too complex for people to be fully informed about. In the Brexit vote, for example, the regions most strongly favoring Leave were also—to the surprise of many voters—the most dependent on European Union markets (Los et al., 2017). Competition for social influence also often involves future decisions and delegations (Baron and Holmström, 1980); for example, voters or clients granting politicians or fund managers the power to make future decisions on their behalf. Finally, predicting the future is hard (Silver, 2012). Pundits who are regularly tasked to predict uncertain future events in finance, politics, or sports often turn out to be wrong (Tetlock, 2017). Competition for social influence thus tends to take place under high outcome uncertainty (Hertwig et al., 2019). That makes it difficult to evaluate advice accuracy and creates opportunities for competing advisers to seek influence strategically (e.g., by masking a strategic lie as a prediction error).

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[2]Department of Cognitive Sciences, University of Haifa, 199 Aba Khoushy Avenue Mount Carmel, Haifa, Israel

[3]Faculty of Psychology and Educational Sciences, General and Experimental Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich, Germany

[4]Faculty of Philosophy, Philosophy of Science and the Study of Religion, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

[5]Department of Economics, University College London, Drayton House, 30 Gordon St, London WC1H 0AX, UK

[6]Department of Psychology, Royal Holloway University of London, Egham Surrey, London TW20 0EX, UK

[7]These authors contributed equally

[8]Lead contact

*Correspondence: kurvers@mpib-berlin.mpg.de

https://doi.org/10.1016/j.isci.2021.103505

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Information asymmetry, delegation of future decisions, and intractable uncertainty all shape the way competitors for social influence communicate their opinions and recommendations to their clients strategically, thereby effecting how these competitions unfold as well as their eventual outcomes. Here we describe the influence-seeking strategy that best succeeds under these conditions, studying this competition as a zero-sum game between advisers: if one adviser wins, the other loses. In the following, we will first study a scenario in which two influencers—henceforth, advisers—compete for the attention of a single client. Using game theory, we propose an intuition about what strategies are rational for the two advisers to adopt when all they care about is to be favored by the client. We show that a strategic, rational adviser communicates information to the client honestly when the client favors them, but lies about it when the client favors the competitor. Next, taking an empirical approach, we show, across seven experiments, that such a strategic adviser is indeed able to outperform an honest adviser in swaying individuals, the majority vote in anonymously voting groups, and the consensus vote in communicating groups of clients. Finally, we show the psychological mechanisms driving the strategic adviser's success.

RESULTS

Strategic dishonesty as a rational strategy

We start by investigating whether there is a dominant, rational strategy for two advisers to adopt when they compete for the attention of a single client. The advisers propose bets to the client on one of two possible outcomes of a lottery that the client cares about. The client can, in each round, only select one adviser for placing her bet, and the advisers thus compete for the client's attention. The advisers' goal is to be chosen by the client as often as possible. The client's goal is to win as many bets as she can. In a finite number of rounds, the client starts each round by selecting one of the two advisers to place a bet on her behalf. The bet is placed on either the black or white color. Both advisers (but not the client) then receive the same probabilistic evidence (p) about the probability of the winning color being black. They then simultaneously offer their respective recommendations s_1 and s_2 to the client indicating their estimates of the probability of black winning. The client follows the selected adviser's recommendation s_i and bets on black (white) if s_i>0.5 (s_i<0.5). Next, the winning color is publicly announced and the client evaluates the recommendations that she received from both advisers in light of the outcome of the lottery to select the adviser they wish to follow in the next round.

Advisers know that, provided $s_1 \neq s_2$, the client updates the competence weights of the two advisers (w_1 and $w_{2,\ \text{with}\ w1+w2=1}$) according to

$$w_i^* = \frac{w_i c_i^2}{w_1 c_1^2 + w_2 c_2^2} \qquad \text{(Equation 1)}$$

where c_i=s_i when the winning color is black and c_i=1−s_i otherwise. If s_1=s_2, the weights remain unchanged: $w_i^* = w_i$. The client's updating rule rewards highly confident correct advice (e.g., high confidence that the winning color is black when the outcome is black, with "high confidence" meaning a report that the probability of winning from betting on black is high) and penalizes highly confident wrong advice (e.g., high confidence that the winning color is black when the outcome is white). Similar to reinforcement learning (RL) updating rules, like Rescorla-Wagner (Lockwood and Klein-Flügge, 2021), more recent observations have greater influence on the influence weight w_i, with the highest weight given to the current confidence and accuracy c_i. Unlike standard RL updating rules, the updating is in the relative accuracy of two advisers and, therefore, the update is normalized with the update of both advisers. In the first round, the client selects an adviser at random. In the following rounds, she selects the adviser with the higher updated weight. If updated weights are equal, the client retains the adviser selected in the previous round. Note that, previous theoretical studies of competitive advice-giving investigated a scenario in which the client also follows a game-theoretically rational strategy (e.g., Krishna and Morgan, 2001); here, however, we fix the client's updating rule as one inspired by previous empirical observation of adviser selection by human clients (Bayarri and De Groot, 1989; Hertz et al., 2017, 2020a, 2020b) and ask how a strategic adviser can best respond to this empirical (rather than ideal) updating rule.

For the two strategic advisers, this is a zero-sum game: whenever one wins, the other loses. This permits us to employ game-theoretic methods developed to study this class of games to ask if, given the client's updating rule and the uncertainty concerning the outcome of a lottery in each round, a rational advising strategy can be found. The advisers' decision problem in any round of the game is solved by backward induction, whereby we first work out their optimal choices in the last round and then work our way back through preceding rounds. In Box 1, we demonstrate this analysis for the last two rounds of the game to gain an intuition about what these rational strategies may be. We find a consistent pattern of a rational advising strategy emerging.

CellPress
OPEN ACCESS

Box 1. The game-theoretic analysis of the advisers' game

We demonstrate the backward induction procedure for the last two rounds of the advisers' game. In our analysis we restrict the advisers' choices s_i to the set [0, 1/9, 2/9, 3/9, 4/9, 5/9, 6/9, 7/9, 8/9, 1]. Relaxing this restriction does not change our conclusions. We also assume that the probability that the winning color is black (p) is drawn from a uniform distribution on [0,1] in each round.

The last round of interest is the one in which advisers' choices still matter to them, i.e., they can influence whom the client will select for her final bet. To illustrate the backward induction procedure, we consider the particular case when, at the start of this round, Adviser 1 is selected and has high influence over the client with w_1=0.8 and, by extension, Adviser 2 has low influence with w_2=0.2. The advisers can use Equation 1 to compute their updated weights for all possible combinations s_1 and s_2, conditional on whether the winning color in the current round will be black or white (Figure 1A). Using these weights, they generate advisers' payoffs, i.e., probabilities of being selected for the client's final bet (Figure 1B) in terms of p and q=1–p, the probability that the winning color is white. As this is a zero-sum game, Adviser 2's payoffs are Adviser 1's payoffs subtracted from 1 and maximizing Adviser 2's payoff is equivalent to minimizing that of Adviser 1.

In a rational solution of the game—a Nash equilibrium—each adviser maximizes her expected payoff given her opponent's choice. We find one equilibrium by iteratively deleting weakly dominated strategies. Adviser 1's strategy s_1=4/9 dominates all s_1<4/9, since, irrespective of Adviser 2's choice, it always yields the same or higher payoff to Adviser 1 as any s_1<4/9 (Figure 1C). Hence, we delete all s_1<4/9. Similarly, we delete all s_1>5/9. For Adviser 2, after these deletions, all 0<s_2<1 are dominated by s_2=0 and s_2=1. Deleting all 0<s_2<1 leaves advisers with two strategies each (Figure 1C). In the Nash equilibrium of this reduced game, the selected Adviser 1 randomizes between the two most cautious advice strategies s_1=4/9 and s_1=5/9 with probabilities q and p, respectively (see STAR methods for full derivation). Provided 0<p<1, the ignored Adviser 2 randomizes between the two most extreme advice strategies s_2=0 and s_2=1 with probabilities p and q. Adviser 1's and 2's equilibrium payoffs (i.e., their expected payoffs when both randomize as above) are p^2-p+1 (which is at least 0.75) and $p-p^2$ (at most 0.25), respectively, which illustrates the selected adviser's advantage.

Although deletion of weakly dominated strategies eliminates other equilibria in the non-reduced game of Figure 1B, in zero-sum games like this one, a player's expected payoff from playing any equilibrium strategy against any equilibrium strategy of her opponent is always the same (Binmore, 2007). This means that advisers do not care which equilibrium strategy they play, and one equilibrium is sufficient to determine advisers' equilibrium payoffs from any strategic (i.e., rational) play.

In the penultimate round, each adviser aims to maximize the probability of being selected at the end of the penultimate *and* the last round. Again, we consider the particular case when, at the start of this round, w_1=0.8. Advisers' weights and, hence, probabilities of being selected at the end of the penultimate round for all possible combinations s_1 and s_2 are the same as before (Figures 1A and 1B). To obtain advisers' payoff matrices in the penultimate round, we need to add their expected probabilities of being selected at the end of the last round to those of being selected after the penultimate round. We focus on Adviser 1 and illustrate this here for the diagonal s_1=s_2 (see Figure 1D and STAR Methods for all combinations s_1 and s_2). In this case, irrespective of the lottery outcome in the penultimate round, her weight at the start of the last round will be 0.8. As already shown, her payoff in the last round will be p^2-p+1. At this stage, advisers do not know the value p in the last round, but they know that it will be drawn from a uniform distribution on [0,1]. Therefore, the expected value of her payoff in the last round is obtained by integrating $\int_0^1 (p^2-p+1)dp \approx 0.83$.

Hence, Adviser 1's expected payoff in the penultimate round is the sum of her expected probabilities of being selected at the end of the penultimate and the last round: 1+0.83=1.83. Adviser 2's payoff in the penultimate round is Adviser 1's payoff subtracted from 2.

Figure 1D shows advisers' payoff matrices for the particular case of p=0.4 in the penultimate round. As can be seen, these are similar to the advisers' payoff matrices in the last round (Figure 1B). The selected Adviser 1 maximizes her expected payoff by using "moderate" strategies close to the truth, i.e., p=0.4, whereas the ignored Adviser 2's best response is to select "extreme" strategies. Indeed, in the only Nash equilibrium in this scenario, Adviser 1 randomizes between s_1=4/9 and s_1=5/9 with probabilities 0.65 and 0.35 respectively, while Adviser 2 randomizes between s_2=0 and s_2=1 with probabilities 0.4 and 0.6. In the STAR Methods, we solve the game when w_1=0.8 and 0.6 for p=0.4, 0.25, and 0.1 to corroborate the emerging pattern of advisers' strategy choices.

When selected, a strategic adviser maximizes their likelihood of maintaining an advantage (i.e., higher weight) by providing moderate recommendations that stay relatively true to the observed evidence ($s_i \sim p$). When ignored, the strategic adviser seeks to strike lucky by offering confident recommendations that contradict the selected adviser's recommendation. The key intuition is that if the lottery outcome turns out to be in line with the ignored adviser's confident counterfactual recommendation, the client's updating

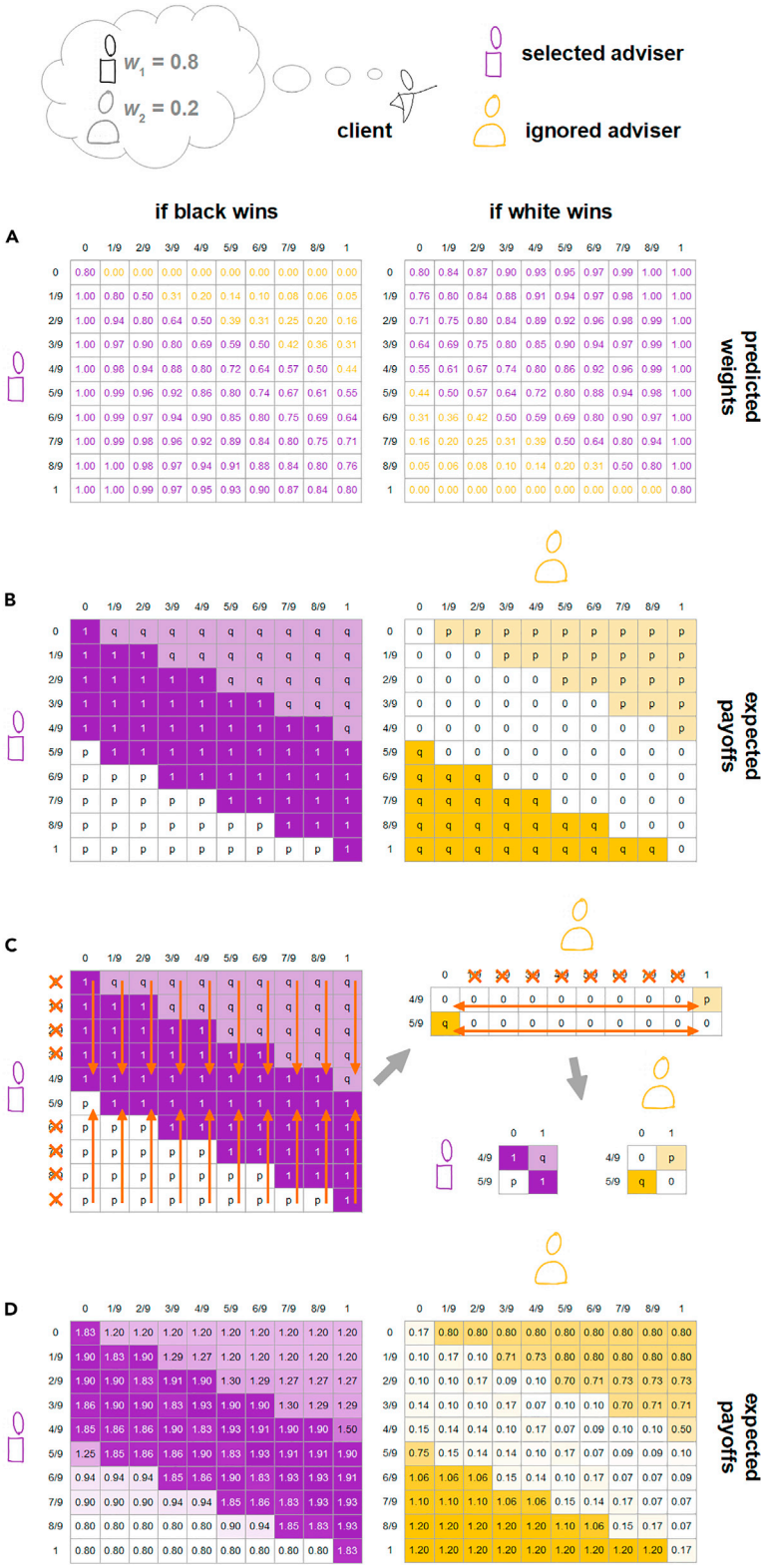

w_1 = 0.8
w_2 = 0.2
client
selected adviser
ignored adviser
A
if black wins
if white wins
predicted weights
B
expected payoffs
C
D
expected payoffs

Figure 1. The game-theoretic analysis of the advisers' game

The game-theoretic analysis for the case where the selected Adviser 1 has high influence over the client with w_1=0.8 and the ignored Adviser 2 has low influence with w_2=0.2 in the last (A–C) and the penultimate (D) round. In all matrices, Adviser 1 chooses between s_1 identified by rows; Adviser 2 chooses between s_2 identified by columns.

(A) Adviser 1's predicted (updated) weights, starting from w_1=0.8, conditional on whether the winning color is black (matrix on the left) or white (right) for all possible combinations s_1 and s_2. Weights greater than or equal to 0.5 are shown in purple, resulting in Adviser 1 being selected for the following round. Weights below 0.5 are shown in yellow, resulting in Adviser 2 being selected for the following round. Note that w_1+w_2=1.

(B) Expected payoffs, i.e., probability of being selected for the following round for Adviser 1 (left) and Adviser 2 (right). For Adviser 1 (2) these are obtained by taking p in each cell where Adviser 1's predicted weight, conditional on the winning color being black, is shown in purple (yellow), and adding q where Adviser 1's predicted weight, conditional on the winning color being white, is shown in purple (yellow). Color scaling indicates lowest (white) to highest (dark) payoff, assuming p=0.4.

(C) Iterative deletion of weakly dominated strategies leaves advisers with two strategies each: Adviser 1 randomizes between two cautious (4/9, 5/9) and Adviser 2 between two extreme (0, 1) advice strategies.

(D) Expected payoffs in the penultimate round for Adviser 1 (left) and Adviser 2 (right). Payoffs correspond to the sum of probabilities of being selected at the end of the penultimate round and at the end of the last round. Here, at the start of the penultimate round, w_1=0.8 and p=0.4. The color scaling is similar to (B).

rule (Equation 1) will take a sizable notice of this missed opportunity. Regrettably, our analysis indicates that in this game, principled honesty does not pay: An adviser that always (i.e., when selected as well as when ignored) communicates the evidence truthfully ($s_i = p$) does worse in swaying the client's vote than a strategic adviser who mixes the "sensible moderate" and "radical contrarian" strategies depending on being selected or ignored. In the STAR Methods, we further corroborate these results, showing that this rational strategy emerges under a wide range of conditions (including when a strategic adviser competes with an honest adviser, and when the client uses a probabilistic decision rule to select advisers).

Strategic dishonesty sways individual voters

Having derived the rational strategy, we next conducted seven preregistered experiments (https://osf.io/9gjyc/) to empirically test whether a strategic adviser employing this mixed strategy would indeed win a client's attention more often compared with an honest adviser who reports truthfully. In all experiments, human participants acted as clients. Over 20 rounds, they attempted to maximize their winnings by deciding, at the beginning of every round, which one of two advisers, symbolized by cartoon figures (Figure 2), to hire for that round. When participants decided for an adviser, they received the lottery ticket recommended by the selected adviser, and also observed the ignored adviser's recommendation. At the end of each round, the lottery outcome (win or loss) was randomly drawn with a probability p to be black (see STAR Methods for details and https://osf.io/9gjyc/for screenshots of experimental instructions of all treatments).

In each round, both advisers received the same information (p), the likelihood that the lottery outcome is black. Programmed by the experimenter, one adviser was honest and the other strategic. The honest adviser always provided truthful predictions (e.g., recommending black with low [high] confidence when there was weak [strong] evidence in favor of black). The strategic adviser also recommended honestly when selected. Crucially, when it was ignored and received weak evidence (i.e., when the evidence was only weakly informative of the correct outcome), the strategic adviser lied by recommending, with medium confidence, the opposite of what the evidence had indicated. By contradicting the weak evidence (and by extension the honest adviser) when ignored, the strategic adviser thus distinguished themselves from the selected honest adviser. The strategic adviser did not contradict strong evidence even if they were ignored, thereby avoiding too many blatant errors. In the STAR Methods, we derive the rational strategy for a strategic adviser who believes that their opponent is honest. This derived strategy closely matches the strategy of the strategic adviser that we programmed into the experiments. Note that in none of our experiments was the strategic adviser's recommendations more likely than the honest adviser's recommendation to be correct. Therefore, if clients were only persuaded by advisers' accuracy, they would remain indifferent between the two advisers. To test whether the strategic adviser was more popular than the honest adviser, we determined whether there was a significant positive effect of round on the likelihood to select the strategic adviser, using hierarchical Bayesian regression models (brms; see Table S3 for model results, and https://osf.io/9gjyc/for data and analysis code). We preregistered exclusion criteria for all experiments. In Experiments 1 and 2, we, however, decided to deviate in one aspect from the preregistered exclusion criteria. In the preregistration of both experiments, we announced that we would exclude participants who did not

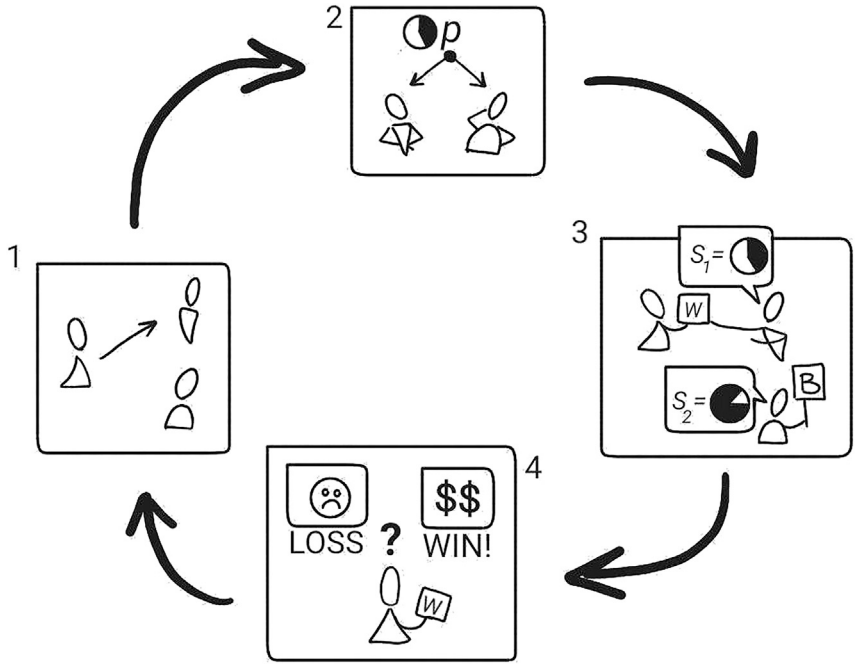


Figure 2. The experimental paradigm for testing the success of the strategic adviser

(1) At the beginning of each round, participants select an adviser to choose a lottery ticket on their behalf. (2) Both advisers then observe the evidence. The pie chart indicates that the evidence (*p*) weakly favors white. (3) The selected adviser provides participants with a ticket (here, White lottery with low confidence corresponding to the weak evidence for white). The ignored adviser also states its recommendation (here Black with high confidence). (4) The lottery is played out and participants may win or lose. Note that the ignored adviser depicted here follows the game-theoretic rational strategy by contradicting the available evidence, effectively lying with high confidence.

sample both advisers. In all seven studies we observed participants who did not sample both advisers, but in all studies it was more likely that these participants always selected the strategic adviser and not the honest one (Figure 3). Therefore, we consider this behavior a feature of participants' strategy and not a lack of engagement and included these participants in the analysis of Experiments 1 and 2 (and removed this criterion in the preregistrations of the subsequent experiments).

We started by investigating whether the strategic adviser can draw the attention of single clients across different levels of evidence strength (i.e., the observed likelihood of winning a bet/the dominance of a given color) and incentive regimes. In a pilot study, we observed that participants (N = 28) were more likely to select the strategic, not honest, adviser (brm: β [confidence interval (CI)] = 0.04 [0.01–0.08]; Figures 3A and 4A). Using these results, we performed numerical simulations to examine the impact of evidence uncertainty (i.e., distance between *p* and chance) on the strategic adviser's success (see preregistration https://osf.io/rsn8h/). These simulations predicted the strategic adviser's influence to increase with increasing uncertainty. Experiment 1 (N = 160) tested this prediction across four levels of evidence strength. As predicted, the strategic adviser's influence was strongest at the weakest level of evidence (i.e., the highest level of uncertainty; evidence 1: β [CI] = 0.10 [0.07–0.14]; evidence 2: β [CI] = 0.03 [0.00–0.06]; evidence 3: β [CI] = 0.06 [0.03–0.09]; evidence 4: β [CI] = 0.03 [-0.00 to 0.06]; Figures 3B and 4B). In Experiment 2 (N = 140) we tested whether the strategic adviser's success depended on the client's incentive to win more lotteries. In contrast to Experiment 1, which incentivized participants for correct lottery outcomes, Experiment 2 did not incentivize participants for correct lottery outcomes; participants received a flat payment, independent of the number of winning rounds. Testing the same four levels of evidence strength, participants still preferred the strategic over the honest adviser when evidence was weakest, and progressively less with increasing evidence (evidence 1: β [CI] = 0.04 [0.01–0.07]; evidence 2: β [CI] = 0.03 [0.00–0.06]; evidence 3: β [CI] = 0.00 [-0.02 to 0.03]; evidence 4: β [CI] = −0.01 [-0.04 to 0.01]; Figures 3C and 4C). In Experiment 3 (N = 45), with uncertainty at maximum and incentives for correct outcomes reinstated, we replicated our key finding that participants preferred strategic over honest advisers, albeit

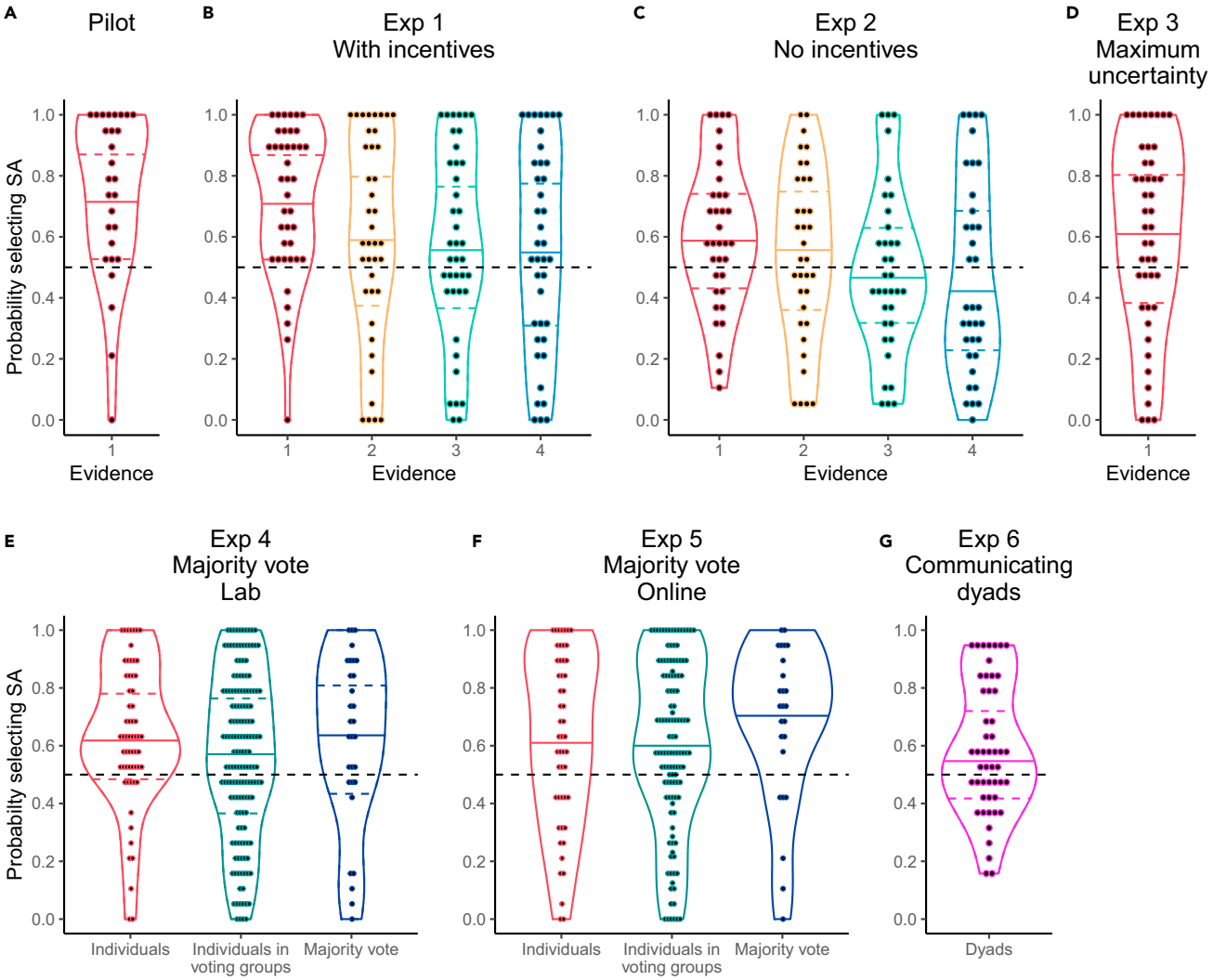


Figure 3. The probability to select the strategic adviser (SA) per experiment

(A–D) Each dot shows a participant's mean likelihood to select the strategic adviser across the 20 rounds (i.e., each dot represents one unique individual). (E–G) (E and F) Red dots show a participant's mean likelihood to select the strategic adviser across the 20 rounds for individuals playing alone. Dark green dots show a participant's mean likelihood to select the strategic adviser across the 20 rounds for individuals playing in majority voting groups. Dark blue dots show the mean likelihood to select the strategic adviser across the 20 rounds for groups using the majority vote (i.e., each dot represents a unique group). (G) Each dot shows a dyad's mean likelihood to select the strategic adviser across the 20 rounds (i.e., each dot represents one unique dyad). In all panels, the first choice (i.e., round 1) was excluded as this constituted a random choice. Violin plots show median and interquartile ranges. Black dashed horizontal lines indicate chance level (0.5).

not significantly (β [CI] = 0.01 [-0.01 to 0.04]; Figures 3D and 4D). Note that across the seven experiments, this was the only case in which this treatment was not significant.

Strategic dishonesty sways voting and communicating groups

Having established the effectiveness of the game-theoretic rational strategy in winning individual clients' attention, we next examined whether this strategy could sway a crowd of voters. If the individuals in the crowd voted entirely independently and, as observed so far, favored the strategic adviser, Condorcet's jury theorem (Boland, 1989; marquis de Condorcet, 1785)—which states that combining independent binary decisions amplifies individual preferences—would predict that the majority vote would favor the strategic adviser even more strongly (see preregistration https://osf.io/z8k3c/ for detailed predictions). In Experiment 4, participants were recruited in groups of five clients (N = 30 groups) whose anonymous votes

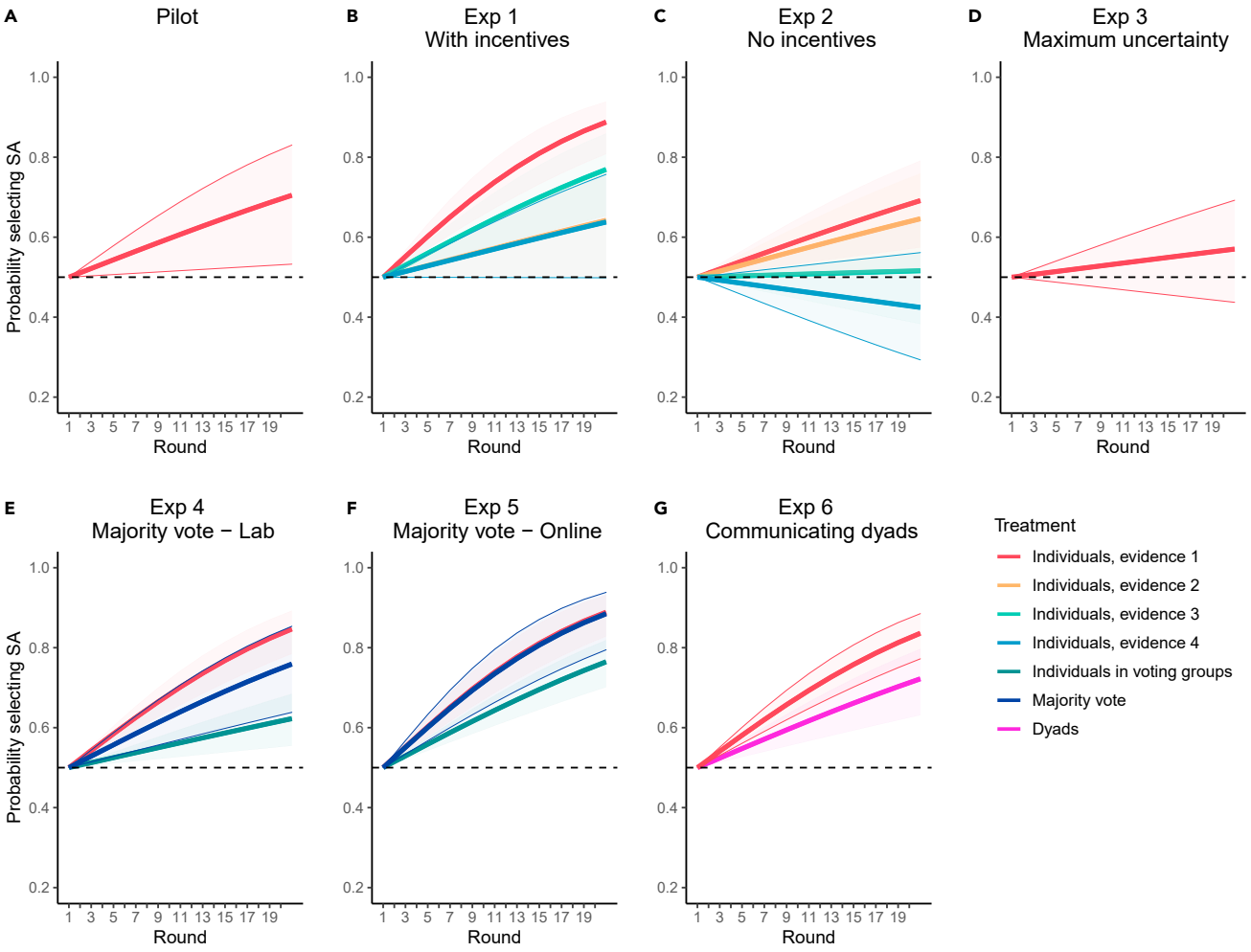


Figure 4. The strategic adviser (SA) exerts a larger influence over single clients, majority-voting groups, and communicating dyads, and at low-evidence strength (i.e., high uncertainty)

The likelihood to select the strategic adviser over 20 rounds across the seven studies with 0.5 being the chance expectation (horizontal dashed line).

(A) In the pilot, single participants were more likely than chance to select the strategic adviser.

(B and C) Single participants were most likely to select the strategic adviser at the lowest evidence level (i.e., highest level of uncertainty), independent of whether lottery outcomes were incentivized (B) or not (C). In (B) the blue and yellow lines overlap.

(D) Under maximum uncertainty, single clients showed a trend for favoring the strategic adviser.

(E and F) Individuals in majority voting groups (dark green line) were more likely than chance, but less likely than single participants (red line), to select the strategic adviser. The majority vote (i.e., aggregating the independent decisions of a group; dark blue line) was as likely as single participants to select the strategic adviser, both in the laboratory (E) and online (F). In (F) the dark blue and red lines overlap.

(G) Communicating dyads were more likely than chance to select the strategic adviser. Thick lines show the mean of the posterior distributions, and bands show 95% credible intervals of Bayesian regression models.

were aggregated by majority rule, a common procedure in elections. The selected adviser's recommendation was the same for all group members. Experiments 1–3 were conducted online, whereas Experiment 4 was conducted in the laboratory. For direct comparison, a separate control experiment—also in the laboratory—was conducted with individual clients (N = 60). Figures 3E and 4E show that single individuals (β [CI] = 0.09 [0.06–0.11]), individual votes within groups (β [CI] = 0.03 [0.01–0.04]), and majority vote decisions (β [CI] = 0.06 [0.03–0.09]) all favored the strategic adviser. The design of Experiment 5 was identical to that of Experiment 4, with one exception: it was conducted online, not in the laboratory. All main findings were replicated (single individuals: β [CI] = 0.10 [0.08–0.13], N = 50; individual votes within groups: β [CI] = 0.06 [0.04–0.08], N = 25 groups; majority vote decisions: β [CI] = 0.10 [0.07–0.14]); Figures 3F and 4F). In Experiments 4 and 5 the magnitude of the strategic adviser's success was similar across individuals and majority

vote, indicating that groups were similarly vulnerable to being swayed by the rational strategy and that individuals within groups did not vote entirely independently (see also next section).

Finally, Experiment 6 investigated the strategic adviser's influence in persuading communicating individuals making joint decisions. Participants were recruited in dyads (N = 50 dyads) and instructed to discuss and agree on which adviser to follow in each round. Previous works have shown that both perceptual decisions under uncertainty (Bahrami et al., 2010) and logical problem solving requiring reasoning by argumentation (Mercier and Sperber, 2011) benefit from face-to-face communication. These findings raise the possibility that face-to-face interacting clients may be able to see through the strategic adviser's tactic. However, dyadic decisions also favored the strategic adviser over the honest adviser (β [CI] = 0.05 [0.03–0.07]; Figures 3G and 4G), thereby lending further support to the generality of the rational strategy for persuasion.

The psychological basis of the success of strategic dishonesty

We had two non-mutually exclusive hypotheses about the underlying psychological basis of the success of the game-theoretic rational strategy. First, following the instrumental-learning literature, we hypothesized that clients' choice of adviser would follow a "win-stay, lose-shift" strategy (Imhof et al., 2007; Nowak and Sigmund, 1993). This strategy predicts that the client's likelihood to switch after a loss does not depend on the ignored adviser's advice. Our second hypothesis was more specific and followed directly from the game-theoretic analysis of the client's updating rule (Equation 1). This hypothesis too predicted that clients would be more likely to shift if the selected adviser gave the wrong advice (as the "win-stay, lose-shift" strategy), but that the shifting likelihood would, additionally, be higher when the ignored adviser had offered a contradicting recommendation. Intuitively, a client who sees that they would have fared better with the ignored adviser is more likely to switch in the next round. Critically, one key insight emerging from our work is that such common sense would be misguided under high uncertainty, when the available information is only weakly predictive of outcomes and can be exploited by a strategic contrarian such as our strategic adviser. To test these two hypotheses, we determined whether there was a significant positive effect of "negative lottery outcome," "contradicting advice," and their interaction on the likelihood to change adviser in the next round using hierarchical brms (see Table S4 for model results, and https://osf.io/9gjyc/for data and analysis code).

Figure 5 shows the results of this analysis. For single individuals at evidence strength levels 1, 2, and 3 (Figures 5A–5C), individuals were most likely to change adviser if they lost and the ignored adviser's recommendation opposed the selected adviser's recommendation (evidence level 1: interaction: β [CI] = 1.03 [0.70–1.36]; level 2: lost: β [CI] = 1.15 [0.75–1.53], contradicting: β [CI] = 0.38 [-0.02 to 0.78]; level 3: interaction: β [CI] = 0.88 [0.20–1.56]). This illustrates the success of the strategic adviser's strategy of distinguishing itself from its competitor when ignored and implies that the client did not entirely disregard the ignored adviser's advice. At evidence level 4, the available evidence was always high, preventing the strategic adviser from using its contrarian strategy, effectively turning into an honest adviser. Hence, we could not test the effect of contradicting, but we did find an effect of "negative lottery outcome" (β [CI] = 0.94 [0.57–1.30]; Figure 5D). The behavior of individual voters within the majority-voting groups showed a more complex pattern. Individuals supporting the majority vote in a given round showed a similar switching pattern to single clients (interaction: β [CI] = 1.18 [0.78–1.59]; Figure 5E), whereas individuals in the minority were most likely to switch when the group won and the ignored adviser presented opposing advice (β [CI] = interaction: −1.57 [-2.22 to −0.93]; Figure 5F) adding support to the currently selected adviser. Finally, dyads were also more likely to change adviser when they lost (β [CI] = 0.75 [0.45–1.05]) and received opposing advice from the ignored adviser (β [CI] = 0.46 [0.14–0.78]; Figure 5G). Taken together, we find strong evidence across treatments not only for a "win-stay lose-shift" strategy but also for shifting when this situation is combined with the ignored adviser having provided opposing advice (i.e., hypothesis 2).

In all treatments, individuals' likelihood to change advisers decreased over the course of the experiment (Figure 6; Table S4). We suggest that this is the result of two processes. First, as shown in Figure 4, the strategic adviser was more likely to be selected over the course of the experiment. When selected, the strategic adviser gave the same advice as the honest adviser. Participants were thus increasingly confronted with identical advice over the course of the experiment, making them less likely to switch (see also Figure 5). Second, individuals may, over time, have moved from an exploration to an exploitation phase as commonly observed in finite games.

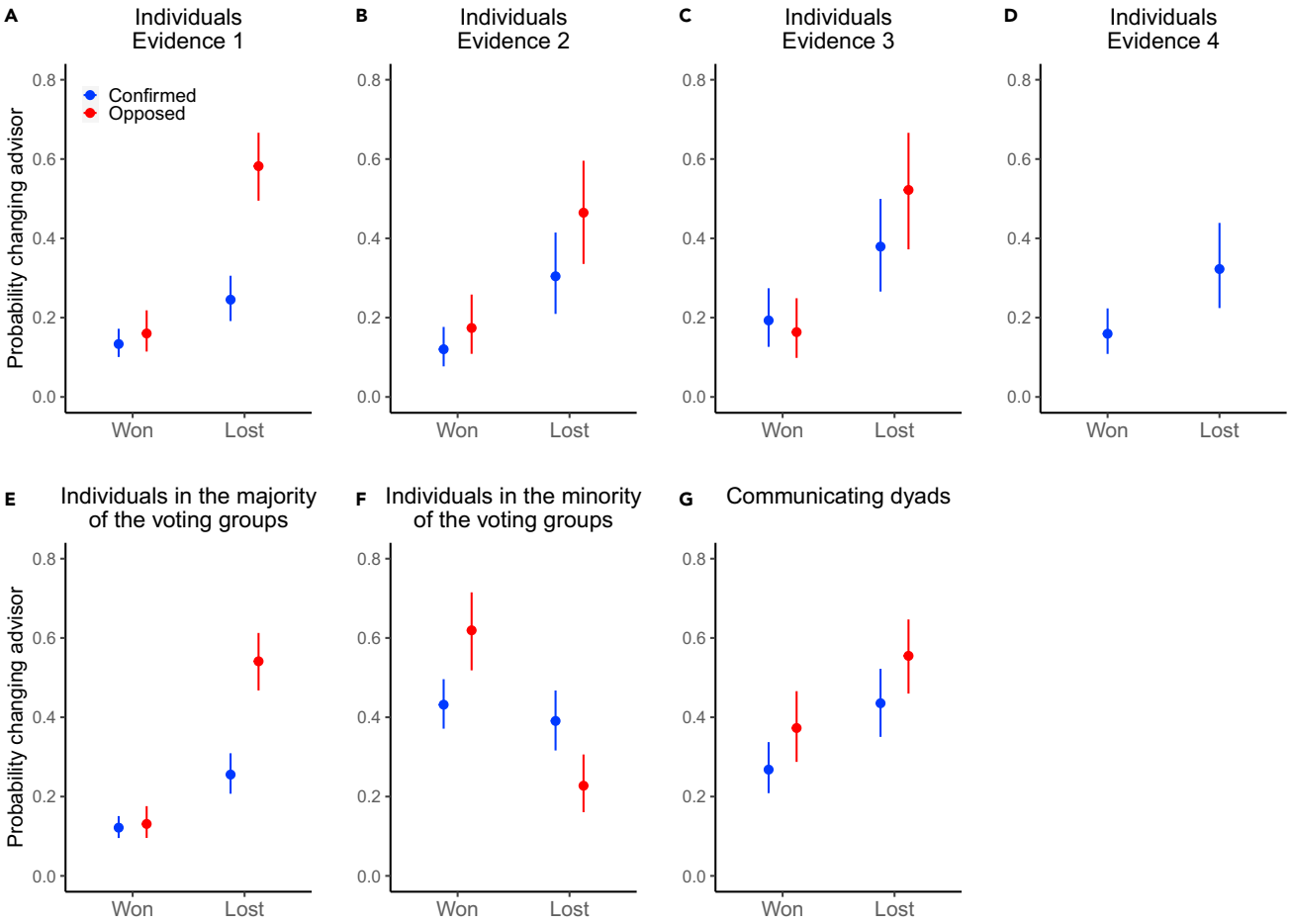


Figure 5. Participants were most likely to change adviser when losing a bet and when the ignored adviser provided opposing advice to the selected adviser

The likelihood to change adviser in the next round as a function of whether the participant(s) won/lost and the ignored adviser confirmed/opposed the advice of the selected adviser per treatment.

(A) At the lowest evidence level (i.e., highest uncertainty level) participants were most likely to change adviser when they lost and the ignored adviser gave the opposing color advice (data are collapsed across Pilot + Experiments 1–6).

(B and C) Similarly at evidence level 2 (B) and 3 (C) participants were most likely to change adviser when they lost and the ignored adviser gave the opposing color advice.

(D) At evidence level 4, the available evidence was always high, preventing the strategic adviser from using its contrarian strategy. (B–D) Data are collapsed across Experiments 1 + 2.

(E) Individuals in the majority of the voting groups were most likely to change adviser when they lost and the ignored adviser gave the opposing color advice.

(F) Individuals in the minority of the voting groups were, however, most likely to change adviser when their group won (against the minority's opinion) and the ignored adviser gave the opposing color advice. (E, F) Data are collapsed across Experiments 4 + 5.

(G) Communicating dyads were most likely to change adviser when they lost and the ignored adviser gave the opposing color advice. Shown are the mean of the posterior distributions and 95% credible intervals of Bayesian regression models.

DISCUSSION

Our theoretical and empirical findings provide converging evidence that by strategically sending out disinformation advisers can gain social influence when competing with other advisers. Our results hark back to Aristotle, who defined politics as a socially interactive game of persuasion between "orators" and "members of assembly" about an uncertain future (Aristotle, 2004). Echoing Aristotle's insight, the sobering observation from our results is that individuals, majority-voting groups, and consensual groups can indeed be swayed by a disingenuous strategy that is not committed to truth, but to beating the competition. Casting the strive for influence as a competition, our results may help explain the presence of truth distortion across many domains of social influence, be it politics, economics, or social media (Lewandowsky et al., 2017). It, for example, helps to explain why advantaged (e.g., incumbent) political candidates are expected

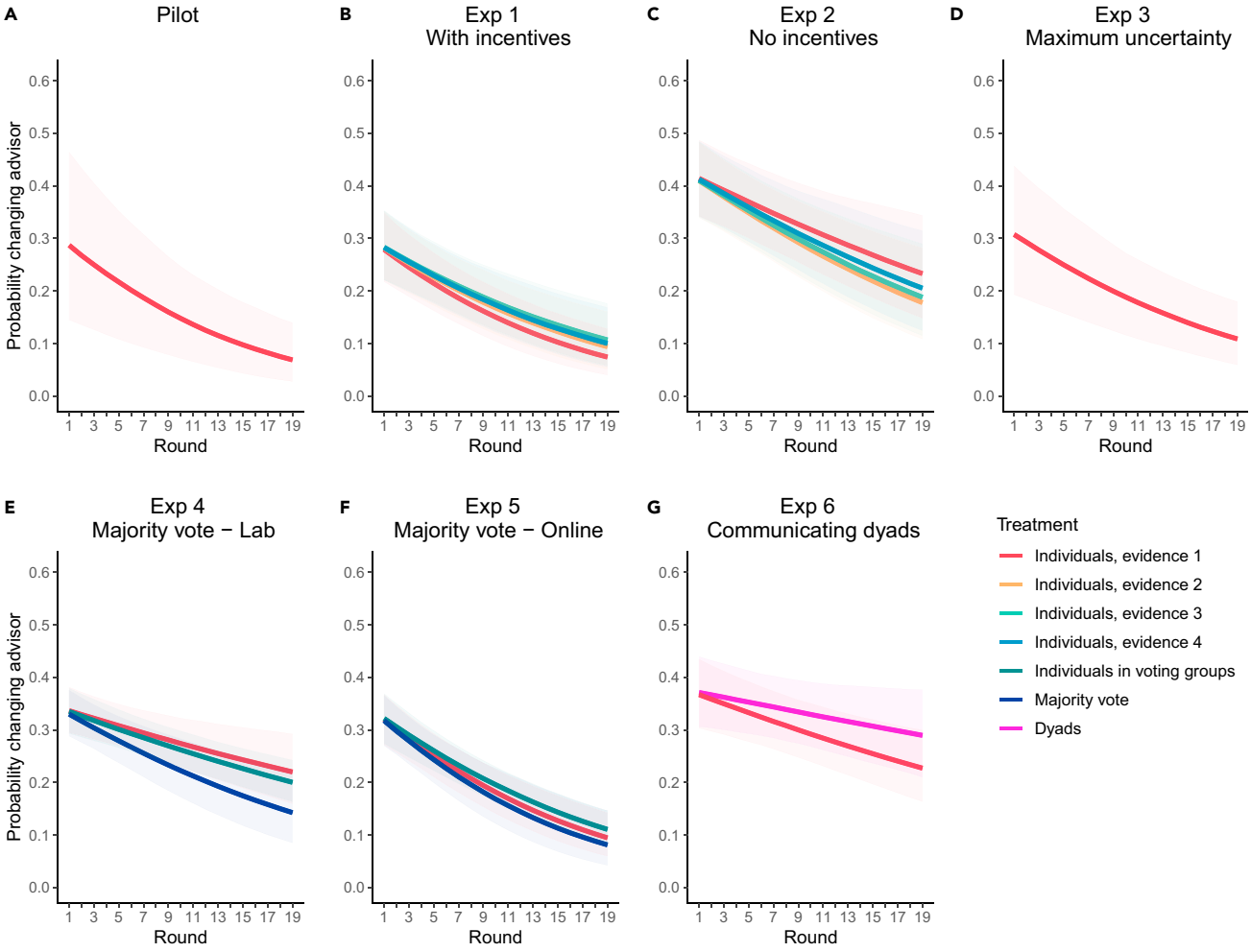


Figure 6. The likelihood to change adviser decreased over time across all studies

(A–D) Single participants became increasingly less likely to change adviser over the course of the experiment, independent of the level of evidence.

(E and F) Also individuals in majority voting groups (dark green line) and the majority vote itself (i.e., the aggregation of the independent decisions in a group; dark blue line) became less likely to change adviser over time.

(G) Communicating dyads were also less likely to change adviser over time. Thick lines show the mean of the posterior distributions, and bands show 95% credible intervals of Bayesian regression models.

to adopt more moderate positions than disadvantaged candidates (e.g., opposition) who (are expected to) take up more extreme (or deviant) positions (Groseclose, 2001; Stone and Simas, 2010). According to former Prime Minister David Cameron, this is exactly what Boris Johnson did in the run up to the Brexit referendum. Cameron claimed that Johnson "risked an outcome he didn't believe in because it would help his political career" (Cameron, 2019). Also, new companies entering competitive markets (Pollock and Gulati, 2007) or job seekers (Levinson and Perry, 2011) are advised to stand out of the crowd.

A key assumption in our game-theoretic analysis is that the client updates the weights it assigns to the advisers based on the perceived accuracy of their advice. In other words, the client herself is not strategic (e.g., Krishna and Morgan, 2001). We used this updating rule because it is widely observed, both in adviser-selection paradigms (Bayarri and De Groot, 1989; Hertz et al., 2017, 2020a, 2020b) and more broadly in social influence studies (Tenney et al., 2019). Highly confident individuals are generally trusted more (Anderson et al., 2012; Tenney et al., 2019; Von Hippel and Trivers, 2011), but when overconfidence is exposed, individuals generally lose their influence. For example, eyewitnesses who were confident but wrong about a memory were consequently judged as less believable when testifying with confidence about other memories (Tenney et al., 2007). A similar situation applies to overconfident job applicants (Tenney

and Spellman, 2011). The reason why this updating rule is widely observed may be because it captures a well-known cognitive bias—i.e., the outcome bias—in people's assessments of the quality of received advice, whereby we evaluate our decisions concerning uncertain events in terms of their consequences (Baron and Hershey, 1988). Another reason may be that clients perceive a highly confident adviser as being more informed about the present lottery and extrapolate from that that this adviser may also be better informed about the lotteries to come. Nevertheless, we note that what was crucial for the success of the strategic adviser was to deviate from the selected adviser when being ignored. This behavior drew the attention to the ignored adviser whenever the selected adviser's advice turned out to be wrong. This was robustly observed across a wide range of settings, including differences in (1) singletons, anonymously voting groups, and discussion groups; (2) incentive structures; and (3) evidence levels. This suggests that this is a robustly observed phenomenon that can be utilized by advisers.

A key psychological insight emerging from our work is that the slogan of "voting for change" can be exploited by a manipulative adviser that follows the game-theoretic optimal strategy. Our results provide a compelling argument why opinions at odds with mainstream views appeal to a broad audience of voters. They further suggest that voting for change is especially appealing when voters experience economic losses (e.g., a reduction in income, or job loss) (Guiso et al., 2017), even, and this is crucial, when this promise of change is neither based on any credible evidence nor of any benefit to the voter. This can, for example, explain the mismatch between local voting and local economic consequences in the Brexit vote (Los et al., 2017). Future research is needed to test the boundary conditions of such strategies (e.g., by relaxing the three hallmarks of competition for social influence describing our experimental paradigm: information asymmetry, delegation of future decisions, and intractable uncertainty), investigate which character traits are especially vulnerable to such strategies, and develop ways to inoculate people from such strategies.

Limitations of the study

In our game-theoretic analysis, we used a simple client's updating rule (Equation 1) to model human clients' adviser-selection process. From this rule, we derived an optimal strategy for advisers seeking influence. Although our empirical results showed that this strategy was useful in swaying human clients (Figures 3 and 4), it is not necessarily the case that this rule was the exact one used to select advisers in the experiments. In our experiments, we tested a number of different scenarios, in terms of group composition (singletons, communicating dyads and majority voting groups), incentive structure, and the uncertainty in the evidence available to the advisers. It is likely that different conditions will alter how clients update the weights they assign to both advisers. Therefore, future work is needed to uncover the variations in such updating rules and how they may change with context. For example, one may use RL models to understand in more detail how human clients update the weights assigned to advisers as a function of group condition, incentive structure, and evidence level (Sutton and Barto, 2018). Another open question is to test whether clients learn about the selected and ignored adviser symmetrically or not (e.g., with different learning rates for each). Such learning models can, in turn, inspire new game-theoretic work for more sophisticated advising strategies.

Another limitation of our work was our focus on scenarios with only two advisers. In many real-world social influence systems (e.g., Twitter, elections, and animal groups) more than two individuals compete for attention. We suspect that, with increased competition between advisers, the need to differentiate oneself from rivals will make the optimal advising strategy diverge from the truth to an even greater extent. Future work could extend both the game-theoretic and empirical work to situations with more than two advisers to study whether our results extend to such scenarios and/or whether other rational strategies emerge.

It is important to note that the human clients in our experiments did not make any irrational decisions. As the strategic adviser aligned its advice with that of the honest adviser once selected, the expected payoff to the client was actually independent of the client's choice. Our experiments were deliberately designed such that random choice or simply following one adviser all the time would have resulted in the same expected payoff to the client. This design feature enables us to offer a key insight, i.e., that the human client's behavior deviates from such simple strategies in a systematic way as humans switch more often when they lose and observe opposing evidence. An adviser who understands this can make use of the client's behavioral tendencies to get selected.

Finally, although our game-theoretic strategy was successful in gaining influence over participants, this does not imply that the strategy is widely used by advisers seeking influence. In previous works using the advice-giving task, many participants playing the role of advisers—either when competing with other human advisers for influence over a human client or when interacting with bot-players—used an attenuated version of the strategy described here. Instead of contradicting the other adviser when ignored by the client, human participants exaggerated their hand and gave overconfident advice (Hertz et al., 2017, 2020a, 2020b). When selected by the client, they stopped exaggerating and gave better-calibrated advice. Using this attenuated version of the strategy allowed participants acting as advisers to distinguish themselves from their rival adviser, while avoiding lying blatantly (Fischbacher and Föllmi-Heusi, 2013; Gneezy, 2005). It is interesting to consider why human advisers deviate from the optimal strategy described here. Possibly, people are motivated by other goals beyond gaining influence, such as prosocial behavior, moral signaling, self-image, or longer-term considerations of reputation beyond the context of the experiment (Arkin et al., 1980; Cheng et al., 2013; Sperber et al., 2010; Zaki, 2014). Advisers with high levels of social anxiety, for example, were less likely to engage in the game-theoretic strategy, suggesting that motivations such as anxiety also play a role in information sharing (Hertz et al., 2017). However, in cases in which gaining influence is the dominant motivation, for example, in political campaigns, we may expect to observe the use of the deceitful strategy as described here more frequently.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Game-theoretical analysis
  - Experiments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Probability of selecting strategic adviser
  - Probability of changing adviser

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103505.

ACKNOWLEDGMENTS

We thank Lucas Molleman for help with programming the experiments and Deborah Ain for feedback on the manuscript. We acknowledge financial support by the Max Planck Institute for Human Development. B.B. and J.K. were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (819040; acronym: rid-O). B.B. was supported by the NOMIS foundation and the Humboldt Foundation. J.K. was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder. U.H. was supported by the National Institute of Psychobiology in Israel (211-19-20) and the Israel Science Foundation (1532/20).

AUTHOR CONTRIBUTIONS

R.H.J.M.K., U.H., and B.B. conceived the original idea. J.K., K.B., and B.B. conducted the game-theoretic analysis. R.H.J.M.K., U.H., M.P.B., B.J., and B.B. designed the experiments. R.H.J.M.K., U.H., and M.P.B. conducted the experiments. R.H.J.M.K. and M.P.B. analyzed the experimental data. R.H.J.M.K., J.K., and B.B. wrote the manuscript with critical input from all other authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 13, 2021
Revised: June 18, 2021
Accepted: November 22, 2021
Published: December 17, 2021

REFERENCES

Anderson, C., Brion, S., Moore, D.A., and Kennedy, J.A. (2012). A status-enhancement account of overconfidence. J. Pers Soc. Psychol. *103*, 718.

Aristotle, R. (2004). Chapter 3. Rhetoric (Mineola, N.Y: Dover Publications).

Arkin, R.M., Appelman, A.J., and Burger, J.M. (1980). Social anxiety, self-presentation, and the self-serving bias in causal attribution. J. Pers Soc. Psychol. *38*, 23.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science *329*, 1081–1085. https://doi.org/10.1126/science.1185718.

Baron, D.P., and Holmström, B. (1980). The investment banking contract for new issues under asymmetric information: delegation and the incentive problem. J. Finance *35*, 1115–1138.

Baron, J., and Hershey, J.C. (1988). Outcome bias in decision evaluation. J. Pers Soc. Psychol. *54*, 569.

Bayarri, M., and De Groot, M. (1989). Comparison of experiments with weighted distributions. In Statistical Data Analysis and Inference, Y. Dodge, ed. (Elsevier), pp. 185–197.

Binmore, K. (2007). Playing for Real: A Text on Game Theory (Oxford University Press).

Boland, P.J. (1989). Majority systems and the Condorcet jury theorem. Statistician *38*, 181. https://doi.org/10.2307/2348873.

Bürkner, P.-C. (2017). brms: an R package for Bayesian multilevel models using Stan. J. Stat. Softw. *80*, 1–28.

Cameron, D. (2019). For the Record (Harper Collins Publ).

Cheng, J.T., Tracy, J.L., Foulsham, T., Kingstone, A., and Henrich, J. (2013). Two ways to the top: evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. J. Pers Soc. Psychol. *104*, 103.

Conradt, L. (2012). Models in animal collective decision-making: information uncertainty and conflicting preferences. Interf. Focus *2*, 226–240. https://doi.org/10.1098/rsfs.2011.0090.

Conradt, L., and List, C. (2009). Group decisions in humans and animals: a survey. Philos. T Roy Soc. B *364*, 719–742. https://doi.org/10.1098/rstb.2008.0276.

Couzin, I.D., Ioannou, C.C., Demirel, G., Gross, T., Torney, C.J., Hartnett, A., Conradt, L., Levin, S.A., and Leonard, N.E. (2011). Uninformed individuals promote democratic consensus in animal groups. Science *334*, 1578–1580. https://doi.org/10.1126/science.1210280.

Fischbacher, U., and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. J. Eur. Econ. Assoc. *11*, 525–547.

Giamattei, M., Yahosseini, K.S., Gächter, S., and Molleman, L. (2020). LIONESS Lab: a free web-based platform for conducting interactive experiments online. J. Econ. Sci. Assoc. 1–17.

Gneezy, U. (2005). Deception: the role of consequences. Am. Econ. Rev. *95*, 384–394.

Groseclose, T. (2001). A model of candidate location when one candidate has a valence advantage. Am. J. Polit. Sci. 862–886.

Grossman, G.M., and Helpman, E. (1996). Electoral competition and special interest politics. Rev. Econ. Stud. *63*, 265. https://doi.org/10.2307/2297852.

Guiso L., Herrera H., Morelli M., and Sonno T. (2017). Populism: Demand and Supply. CEPR Discussion Paper No. DP11871.

Healy, P.M., and Palepu, K.G. (2001). Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. J. Account. Econ. *31*, 405–440. https://doi.org/10.1016/s0165-4101(01)00018-0.

Hertwig, R., Pleskac, T.J., and Pachur, T. (2019). Taming Uncertainty (Mit Press).

Hertz, U., Bell, V., Barnby, J.M., McQuillin, A., and Bahrami, B. (2020a). The communication of metacognition for social strategy in psychosis: an exploratory study. Schizophr. Bull. Open *1*, sgaa058.

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C.D., and Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. Nat. Commun. *8*, 2191. https://doi.org/10.1038/s41467-017-02314-5.

Hertz, U., Tyropoulou, E., Traberg, C., and Bahrami, B. (2020b). Self-competence increases the willingness to pay for social influence. Sci. Rep. *10*, 1–11.

Imhof, L.A., Fudenberg, D., and Nowak, M.A. (2007). Tit-for-tat or win-stay, lose-shift? J. Theor. Biol. *247*, 574–580.

Kitschelt, H. (2000). Linkages between citizens and politicians in democratic polities. Comp. Pol. Stud. *33*, 845–879. https://doi.org/10.1177/001041400003300607.

Krishna, V., and Morgan, J. (2001). A model of expertise. Q. J. Econ. *116*, 747–775.

Levinson, J.C., and Perry, D.E. (2011). Guerrilla Marketing for Job Hunters 3.0: How to Stand Out from the Crowd and Tap into the Hidden Job Market Using Social Media and 999 Other Tactics Today (John Wiley & Sons).

Lewandowsky, S., Ecker, U.K.H., and Cook, J. (2017). Beyond misinformation: understanding and coping with the "Post-Truth" era. J. Appl. Res. Mem. Cogn. *6*, 353–369.

Lockwood, P.L., and Klein-Flügge, M.C. (2021). Computational modelling of social cognition and behaviour—a reinforcement learning primer. Soc. Cogn. Affect. Neurosci. *16*, 761–771.

Lorenz-Spreen, P., Mønsted, B.M., Hövel, P., and Lehmann, S. (2019). Accelerating dynamics of collective attention. Nat. Commun. *10*, 1759. https://doi.org/10.1038/s41467-019-09311-w.

Los, B., McCann, P., Springford, J., and Thissen, M. (2017). The mismatch between local voting and the local economic consequences of Brexit. Reg. Stud. *51*, 786–799. https://doi.org/10.1080/00343404.2017.1287350.

marquis de Condorcet, M.J.A. (1785). Essai sur l'application de l'analyse a la probabilite des decisions: rendues a la pluralite de voix (De l'Imprimerie Royale).

McAfee, A., and Brynjolfsson, E. (2012). Big data: the management revolution. Harv. Bus. Rev. *90*, 60.

Mercier, H., and Sperber, D. (2011). Argumentation: its adaptiveness and efficacy. Behav. Brain Sci. *34*, 94.

Miller, N., Garnier, S., Hartnett, A.T., and Couzin, I.D. (2013). Both information and social cohesion determine collective decisions in animal groups. Proc. Natl. Acad. Sci. USA *110*, 5263–5268. https://doi.org/10.1073/pnas.1217513110.

Nowak, M., and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature *364*, 56–58. https://doi.org/10.1038/364056a0.

Pine, B.J., and Gilmore, J.H. (1998). Welcome to the experience economy. Harv. Bus. Rev. *76*, 97.

Pollock, T.G., and Gulati, R. (2007). Standing out from the crowd: the visibility-enhancing effects of IPO-related signals on alliance formation by entrepreneurial firms. Strateg. Organ. *5*, 339–372.

Savani, R., and von Stengel, B. (2015). Game Theory Explorer: software for the applied game theorist. Comput. Manag. Sci. *12*, 5–33.

Silver, N. (2012). The Signal and the Noise: Why So Many Predictions Fail-Bbut Some Don't (Penguin).

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., and Wilson, D. (2010). Epistemic vigilance. Mind Lang. *25*, 359–393.

Stone, W.J., and Simas, E.N. (2010). Candidate valence and ideological positions in US house elections. Am. J. Polit. Sci. *54*, 371–388.

Strandburg-Peshkin, A., Farine, D.R., Couzin, I.D., and Crofoot, M.C. (2015). Shared decision-making drives collective movement in wild baboons. Science *348*, 1358–1361.

Strandburg-Peshkin, A., Farine, D.R., Crofoot, M.C., and Couzin, I.D. (2017). Habitat and social factors shape individual decisions and emergent group structure during baboon collective movement. Elife *6*, e19505.

Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction, Second Edition (MIT Press).

Tenney, E.R., MacCoun, R.J., Spellman, B.A., and Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. Psychol. Sci. *18*, 46–50.

Tenney, E.R., Meikle, N.L., Hunsaker, D., Moore, D.A., and Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. J. Pers Soc. Psychol. *116*, 396.

Tenney, E.R., and Spellman, B.A. (2011). Complex social consequences of self-knowledge. Social Psychol. Personal. Sci. *2*, 343–350.

Tetlock, P.E. (2017). Expert Political Judgment: How Good Is it? How Can We Know?, New Edition (Princeton University Press).

Turner, J.C. (1991). Social Influence (Thomson Brooks/Cole Publishing Co)).

Von Hippel, W., and Trivers, R. (2011). The evolution and psychology of self-deception. Behav. Brain Sci. *34*, 1.

Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. Sci. Rep. *2*, 1–8. https://doi.org/10.1038/srep00335.

Zaki, J. (2014). Empathy: a motivated account. Psychol. Bull. *140*, 1608.

STAR★METHODS

KEY RESOURCES TABLE

RESOURCE AVAILABILITY

REAGENT or RESOURCE	IDENTIFIER	SOURCE
Deposited data		
Data and statistical analysis	NA	https://osf.io/9gjyc/
Software and algorithms		
Lioness Lab Version 1.1	LIONESS Lab	https://lioness-lab.org/
R version 4.0.4	R Project	https://www.r-project.org/
RStudio version 1.4.1106	RStudio	https://www.rstudio.com/

Lead contact

Further information and requests for resources should be directed to the lead contact, Ralf Kurvers (kurvers@mpib-berlin.mpg.de).

Materials availability

No materials were newly generated for this paper.

Data and code availability

- Data: The datasets generated during this study are available at https://osf.io/z8k3c/.
- Code: The statistical code and the code used for the numerical simulations are available at https://osf.io/z8k3c/.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All experimental studies were either approved by the Institutional Review Board of the University College London (UCL ICN; Pilot, Exp. 2, 3; Ethics approval number: 5375/001) or of the Max Planck Institute for Human Development (MPIB; Exp. 1, 4, 5, 6; Ethics approval numbers: A2020-7, A2019/39, A2019/18, A2019/38). For lab studies, participants signed a consent form prior to starting the experiment, and for online studies, participants checked a box, indicating their consent. Participants' age and gender, as well as sample size, are detailed below for each of the seven studies.

METHOD DETAILS

- Game-theoretic analysis
- Experiments
- Quantification and statistical analysis

Game-theoretical analysis

Nash equilibria in the last round when w_1=0.8. As described in Box 1 in the main text, iterative deletion of weakly dominated strategies leaves advisers with two strategies each. We distinguish between pure and mixed strategies. In the reduced game, i.e., the remaining game after deletions, Adviser 1's pure strategies are s_1=4/9 and s_1=5/9; Adviser 2's pure strategies are s_2=0 and s_2=1. In a Nash equilibrium, each adviser maximizes her expected payoff given her opponent's strategy. The cases when p=0 and p=1 are trivial, since, irrespective of Adviser 2's strategy, Adviser 1 guarantees a sure win by playing her pure strategy s_1=4/9 and s_1=5/9 respectively. When $0<p<1$, there is no equilibrium in pure strategies: if Adviser 1 plays s_1=4/9, Adviser 2 maximizes her payoff with s_2=1, but if Adviser 2 plays s_2=1, Adviser 1 maximizes her payoff with s_1=5/9, and so on. In other words, Adviser 1 tries to align her advice with Adviser 2 by choosing the s_1

closest to s_2, while Adviser 2 tries to differentiate from Adviser 1 by choosing the s_2 furthest from s_1. As a result, their best response choices of pure strategy are in a continuous cycle. In equilibrium, both advisers thus play mixed strategies, randomizing between their pure strategies with some probabilities.

We find these probabilities by using the fact that, in a mixed-strategy equilibrium, the expected payoffs from all pure strategies that a player plays with positive probability must be equal. Let ε and $1-\varepsilon$ be the probabilities with which Adviser 2 plays s_2=0 and s_2=1 respectively. Adviser 1's expected payoffs from playing her pure strategies s_1=4/9 and s_1=5/9 are $\varepsilon+q(1-\varepsilon)$ and $p\varepsilon+1-\varepsilon$ respectively. Equating the two and making use of $q=1-p$ yields $\varepsilon=p$. Similarly, it can be derived that Adviser 1 plays s_1=4/9 with probability q. Thus, when $0<p<1$, there is one Nash equilibrium in the reduced game of Figure 1b. Adviser 1 randomizes between her pure strategies s_1=4/9 and s_1=5/9 with probabilities q and p respectively, while Adviser 2 randomizes between s_2=0 and s_2=1 with probabilities p and q. Note that the lower the p, the higher the likelihood that Adviser 1 announces s_1=4/9 and Adviser 2 announces s_2=1 (her extreme strategy that is furthest from p). Adviser 1's expected equilibrium payoff is obtained by plugging $\varepsilon=p$ into the payoff from playing any of her pure strategies to which she assigns positive probability in mixed strategy equilibrium play (i.e., $\varepsilon+q(1-\varepsilon)$ or $p\varepsilon+1-\varepsilon$). This yields p^2-p+1, the lowest value of which is 0.75 when p=0.5. Adviser 2's payoff is Adviser 1's payoff subtracted from 1: $p-p^2$.

Deletion of weakly dominated strategies eliminates other equilibria in the non-reduced game of Figure 1b. However, Adviser 2's equilibrium strategy is the same in all Nash equilibria of the non-reduced game. This is because, in zero-sum games like this one, equilibria are equivalent, meaning that a player's expected payoff in all equilibria is the same, and interchangeable, meaning that if strategy pairs (s_1,s_2) and (s_1^*,s_2^*) constitute equilibria, then so do (s_1,s_2^*) and (s_1^*,s_2) (Binmore, 2007). Therefore, an adviser's equilibrium strategy must yield the same expected payoff and be payoff-maximizing against any possible equilibrium strategy of her opponent. From Figure 1B it can be seen that no deviation from Adviser 2's equilibrium strategy found earlier satisfies these criteria for Adviser 1. There are, however, deviations from Adviser 1's equilibrium strategy above that satisfy these criteria for Adviser 2. For example, Adviser 1 may use probabilities q and p to randomize between s_1=0 and s_1=5/9. Altogether, Adviser 1 has 15 equilibrium strategies to choose from. In each, she randomizes between some $s_1 \leq 4/9$ and some $s_1 \geq 5/9$ with probabilities q and p respectively (Table S1).

Nash equilibria in the last round when $\mathbf{w_1 \geq w_2}$. When $w_1 \geq w_2$, iterative deletion of weakly dominated strategies reduces Adviser 1's payoff matrix to a $n \times n$ matrix in which Adviser 2 always retains her extreme strategies $s_2=y_1=0$ and $s_2=y_n=1$ whenever $n \geq 2$ (Figure S1). Figure S2 shows this for w_1=0.9, 0.8, 0.7, 0.6, and 0.5. When p=0 or p=1, no matter what Adviser 2 does, Adviser 1 guarantees a sure win by playing her pure strategy $s_1=x_1$ or $s_1=x_n$ respectively. When $0<p<1$, there is no Nash equilibrium in pure strategies. As previously, the probabilities with which advisers randomize between pure strategies in a mixed-strategy Nash equilibrium can be found by using the fact that, in a mixed-strategy equilibrium, the expected payoffs from all pure strategies that a player plays with positive probability must be equal.

Let ε_i be the probability with which Adviser 2 plays $s_2=y_i$. Adviser 1's expected payoff from playing her pure strategy $s_1=x_1$ is $\varepsilon_1+q(1-\varepsilon_1)$, that from playing $s_1=x_2$ is $p\eta_1+\varepsilon_2+q(1-\varepsilon_1-\varepsilon_2)$, and so on. In general, the expected payoff from playing $s_1=x_i$ is $p(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_{i-1})+\varepsilon_i+q(1-\varepsilon_1-\varepsilon_2-\ldots-\varepsilon_i)$. Equating expected payoffs from any two adjacent x_{i-1} and x_i yields $\varepsilon_i=(q/p)\varepsilon_{i-1}$, from which it follows that $\varepsilon_i=(q/p)^{i-1}\varepsilon_1$. Making use of $\sum\varepsilon_i=1$ gives $\varepsilon_1[1+q/p+(q/p)^2+\ldots+(q/p)^{n-1}]=1$, where the elements in square brackets are the sum of terms in a geometric series. Thus, whenever $p \neq 0.5$, this yields:

$$\varepsilon_1 = \frac{1-q/p}{1-(q/p)^n}$$

Inserting this into the formula for ε_i gives:

$$\varepsilon_i = (q/p)^{i-1}\frac{1-q/p}{1-(q/p)^n}$$

It can be similarly derived that Adviser 1 plays $s_1=x_i$ with probability:

$$\eta_i = (p/q)^{i-1}\frac{1-p/q}{1-(p/q)^n}$$

When p=0.5, advisers play all x_i and y_i with equal probabilities $\frac{1}{n}$.

When p<0.5 (in which case p/q<1) η_i is decreasing in i, i.e., Adviser 1 uses higher probabilities for $s_1 \leq 4/9$ than for $s_1 \geq 5/9$ (vice versa when p>0.5). When $n \geq 2$, this is reversed for Adviser 2. (When n=1, it does not matter what Adviser 2 does because Adviser 1 always wins.) The fact that in zero-sum games equilibria are equivalent and interchangeable implies that this is the only Nash equilibrium in the reduced game of Figure S1, and the above conclusions hold in every equilibrium of a non-reduced game too.

When $p \neq 0.5$, Adviser 1's expected equilibrium payoff can be computed by plugging ε_1 derived above into the formula for the expected payoff from playing her pure strategy $s_1=x_1$ (i.e., $\varepsilon_1+q(1-\varepsilon_1)$). Simplified and rearranged, this yields:

$$P_n = \frac{p^{n+1}-q^{n+1}}{p^n-q^n}$$

When p=0.5:

$$P_n = \frac{1}{2}+\frac{1}{2n}$$

Adviser 2's expected payoff is $1-P_n$. Since P_n>0.5 for all p and n, the selected adviser always has an advantage.

Generating payoff matrices in the penultimate round. As noted earlier, iterative deletion of weakly dominated strategies in the last round yields the $n \times n$ payoff matrix (Figure S1) where n is determined by advisers' weights at the end of the penultimate round. When $w_1 \geq w_2$, n increases from 1 to 10 as w_1 decreases from 1 to 0.5 (Figure S2). Figure S3 shows Adviser 1's payoff matrices in the last round when, at the end of the penultimate round, w_1=0.836 and w_1=0.835. In these cases, iterative deletion of weakly dominated strategies yields payoff matrices of sizes n=1 and n=2 respectively. As can be seen, n changes from 1 to 2 when Adviser 1's payoff associated with the pair of strategies s_1=5/9 and s_2=0 changes from 1 to p, i.e., when Adviser 1's updated end-of-round weight when the winning color is white falls below 0.5. Advisers know that the client updates their competence weights using (Equation 1). Thus, solving

$$\frac{w_1(1-5/9)^2}{w_1(1-5/9)^2+w_2(1-0)^2} = 0.5$$

shows that this happens when w_1 falls below $81/97 \approx 0.8351$. Cut-off weights for other values n can be obtained similarly (Table S2).

We use this information to generate Adviser 1's payoff matrix in the penultimate round. In the main text we consider the particular case when, at the start of this round w_1=0.8, and illustrate the procedure for the diagonal $s_1=s_2$. Here we show the procedure for other combinations s_1 and s_2.

Suppose the advisers were to announce s_1=1/9 and s_2=2/9 in the penultimate round. Adviser 1 would be certain to be selected at the end of the round and her updated weight would be either w_1=0.5 or w_1=0.84, depending on whether the winning color in the penultimate round is black (the probability of which is p) or white (the probability of which is q) respectively (Figures 1A and 1B). The general formula for Adviser 1's equilibrium payoff P_n in the last round was derived in the previous section. The corresponding values n when, at the start of the last round, w_1=0.5 and w_1=0.84, are 10 and 1 respectively (Table S2). At this stage advisers do not know the value p in the last round, but they know that it will be drawn from a uniform distribution on [0,1]. What matters, thus, are the expected values of Adviser 1's payoffs in the last round that are obtained by integrating P_1 and P_{10} with respect to p. These values are given in Table S2. Thus, Adviser 1's expected payoff from the pair of strategies s_1=1/9 and s_2=2/9 in the penultimate round is $1+p\int_0^1 P_{10}dp+q\int_0^1 P_1 dp$ = 1.75+0.25q. Adviser 1's complete payoff matrix is shown in Figure S4.

Nash equilibria in the penultimate round when w_1=0.8 and w_1=0.6. We next solve the game in the penultimate round when w_1=0.8 and w_1=0.6 for p=0.4, 0.25, and 0.1. Adviser 1's payoff matrices are shown in Figure S4 (w_1=0.8) and Figure S5 (w_1=0.6). To compute the advisers' equilibrium strategies, we use the freely available software Game Theory Explorer (Savani and von Stengel, 2015) (http://www.

gametheoryexplorer.org/). There is only one Nash equilibrium in each considered scenario. The probabilities with which advisers randomize between their pure strategies in each equilibrium are shown in Figure S6. The selected adviser (i.e., Adviser 1) randomizes between moderate strategies closer to the truth as compared to the ignored adviser (i.e., Adviser 2). The ignored adviser assigns high probabilities to strategies that are contrary to what the selected adviser reports. Also, when the ignored adviser's weight is low, she is most likely to announce s_2=0 or s_2=1 that is furthest from truth.

Strategic versus honest adviser. While honest reporting of truth—announcing s_i=p with probability 1—is often a payoff-maximizing strategy in the last round (because of multiplicity of Nash equilibria), this is rarely the case in earlier stages of the game. This plays to a strategic (i.e., payoff-maximizing) adviser's advantage: whenever honest reporting is not part of equilibrium play, the strategic adviser's payoff is higher than in an equilibrium and, hence, her chances of being selected go up. If a strategic adviser believes her opponent to be also strategic, her choice of strategy is determined the same way as before. However, if she knows or believes her opponent to be honest, she can increase her chances of being selected even further by choosing the best, i.e., a payoff-maximizing, strategy in response to honest play. Here we consider this scenario.

Once selected, a strategic adviser can switch to honest reporting of truth, since this guarantees her a sure win in all subsequent rounds of the game. To determine her payoff-maximizing strategy when she is not selected, we start by analysing the last round of the game. Consider the case when, at the start of this round, her weight is w_1=0.2. Figure S7A shows her updated weights at the end of this round for all possible combinations s_1 and s_2, conditional on whether the winning colour in this round will be black or white. Using these weights we can generate Adviser 1's payoffs, i.e., probabilities of being selected for the client's final bet (Figure S7B). The honest adviser always chooses s_2 that is closest to p. Thus, having observed p, the strategic adviser knows what her opponent will do. Differently from payoff matrices analyzed earlier, each column here, therefore, represents the strategic adviser's decision problem for a known value p. For example, when p is close to 1/9 (and the honest adviser, therefore, chooses s_2=1/9) the strategic adviser maximizes her chances of winning by announcing $s_1 \geq 3/9$. She will win if the winning colour is black, the probability of which is p, i.e., close to 1/9. While announcing s_1=0 or s_1=1 that is furthest from p is always the strategic adviser's payoff-maximizing strategy in this scenario, she can afford to choose moderate s_1 that are closer to truth when p is sufficiently small or large.

Figure S8 shows the case when w_1=0.4. In this scenario, for intermediate values p, the strategic adviser does best by overstating the evidence. For example, when p is close to 3/9, she maximizes her payoff by announcing $s_1 \leq 1/9$. She can also afford to play moderate strategies that are not far from truth by just slightly overstating the strength of observed evidence for the winning colour being either black or white.

In the penultimate round, the strategic adviser aims to maximize the probability of being selected at the end of the penultimate *and* the last round. Consider the case when, at the start of the penultimate round, her weight is w_1=0.2. To generate her payoff matrix, we need to first compute her expected payoffs in the last round for the possible updated weights w_1 at the end of the penultimate round (Figure S7A). Suppose that the advisers in the penultimate round were to announce s_1=1/9 and s_2=1/9. Irrespective of the colour drawn in the penultimate round, her weight at the start of the last round would be w_1=0.2. The relevant payoff matrix in the last round is, therefore, that of Figure S7B. Her payoffs for all possible p in the last round are the maximum values of each column in this matrix. Since the value p in the last round is at this stage unknown, her expected payoff is a weighted sum of these maximum values: $2\times1/18\times1/36+2\times1/9(1/9+2/9+3/9+4/9)=0.25$ (because the ranges of values p that are associated with s_2=0 and s_2=1 are half the size of the ranges associated with any other s_2, we use weights 1/18 and 1/9 accordingly). The strategic adviser knows that she will not be selected at the end of the penultimate round when s_1=1/9 and s_2=1/9. As such, her expected payoff from this pair of strategies in the penultimate round is 0.25 derived above.

Performing the remaining calculations yields the penultimate round payoff matrix of Figure S7C. For intermediate values p, a strategic adviser does best by playing strategies that are far from the truth and contrary to the received evidence. But as p becomes sufficiently small or large, her payoff-maximizing strategies are closer to the truth. Figure S8C shows the case when w_1=0.4. In this scenario, for intermediate values p, the strategic adviser does best by playing moderate strategies that are not far from truth and just slightly overstate the strength of the observed evidence.

To summarize, when ignored, i.e., not selected by the client, what the strategic adviser does depends on what she believes her current weight to be. If she thinks that her weight is relatively high, she plays moderate strategies that are not far from truth by slightly overstating the observed evidence. However, as she begins to suspect that her weight may be low, she becomes increasingly likely to opt for extreme strategies that are contrary to the observed evidence and the honest adviser's reports. Once selected, however, the strategic adviser reverts to reporting truth.

Softmax decision rule. Thus far, we made two important assumptions about the client. We assumed i) a particular way in which the client updates the weights of their two advisers (Equation 1), and ii) a particular decision rule with which the client chooses their adviser in every round of the game based on the assigned weights. For the latter, the decision rule was simply: "choose the adviser with the highest weight". In other words, the higher-weighted adviser was chosen with probability 1 (conversely, the lower-weighted adviser was chosen with probability 0). Next to this deterministic rule, we here consider a probabilistic rule: "use the two advisers' weights as probabilities with which you choose them". This implies that i) both advisers are chosen with positive probabilities and these probabilities add up to 1, and ii) the higher-weighted adviser is chosen with a higher probability than the lower-weighted adviser. This is sometimes referred to as the softmax decision rule. (The actual softmax function has the same logic as the one we use here, except that it normalizes the input weights using an exponential function. In our case, the advisers' weights can be used as decision probabilities without the need to transform them first.)

The analysis is the same as before: we consider the last round and then work our way back to the penultimate round of the game. Figures S9 and S10 show the advisers' rational strategies in the last and the penultimate rounds of the game respectively. Comparing the advisers' choices in the penultimate round of the game between Figures S6 (the original decision rule) and S10 (the softmax decision rule), we can conclude the following about the softmax decision rule implementation:

First, when the difference between the two advisers' weights is high, e.g., w_1=0.8 and w_2=0.2 (the left hand side in the Figures), the higher-weighted adviser reports the truth or something very close to it. The lower-weighted adviser strategically misreports the observed evidence by either overstating it or understating it. When the observed evidence is "strong", e.g., when p is close to 0 in our examples, the lower-weighted adviser remains close to the truth but still differentiates themselves from the higher-weighted adviser in an attempt to "strike lucky". When the observed evidence is "weak", e.g., when p=0.4, the lower-weighted adviser is likely to report extremes (s_2=0 or s_2=1) using higher probability for the extreme that is furthest from the truth.

Second, when the difference between the two advisers' weights is small, e.g., w_1=0.6 and w_2=0.4 (the right hand side in the Figures), both advisers tend to stick close to the truth and are almost indistinguishable in terms of their advice strategies. This makes sense, because in this scenario they both know that they are already nearly equally likely to be selected by the client. When the observed evidence is "strong", they pretty much report the same thing to the client. When the observed evidence is "weak" the lower-weighted adviser will just slightly deviate from the truth in an attempt to "strike lucky" and thus to tilt the client's weight a little more in their advantage.

Experiments

Experimental protocols. All lab studies were conducted at the behavioral lab of the Center for Adaptive Rationality (ARC) at the MPIB. All online studies were conducted at Amazon Mechanical Turk (MTurk), only including individuals from the United States with a minimum HIT approval rating of 90%, and a history of at least 100 approved HITs. All six experiments (but not the pilot) were preregistered (https://osf.io/9gjyc/). The order of presentation of the studies in the Main Text slightly deviates from the presentation order of the preregistrations (Table S5). All studies were programmed in LIONESS (Giamattei et al., 2020).

Pilot experiment. In the pilot experiment, single participants observed two advisers, symbolized by cartoon figures, and had to decide for 20 rounds which of the two advisers they wanted to hire (Figure 2, see https://osf.io/vybak/ for screenshots). The sequence of a round was as follows: 1) The participant selected which adviser to follow. 2) Both advisers observed the available evidence. The evidence was generated as follows: a rack of 100 balls was filled with a mix of white and black balls and both advisers observed the same (randomly sampled) 75 balls from the rack. 3) Both advisers communicated their advice

to the client. Their advice consisted of (i) color (i.e., black or white), and (ii) confidence level (1–5 scale). The participant betted on the colour advice of the selected adviser, but also observed the recommendation of the ignored adviser. 4) One ball from the rack was randomly drawn. If the colour advice of the selected adviser matched (did not match) the colour of the drawn ball, the participant won (lost). The two advisers played different strategies: honest or strategic. The honest adviser reported the colour honestly: if the majority of balls it observed was white (black), it reported white (black) to the participant. It also reported confidence honestly, using a linear mapping of evidence level and confidence: majority of one colour: 51-60%: Confidence (CF) = 1. 61-70%: CF = 2. 71-80%: CF = 3. 81-90%: CF = 4. 91-100%: CF = 5. A small amount of noise was added: in 25% of cases the CF level was increased (or decreased) with one unit, provided this was possible.

The strategy of the strategic adviser depended on whether or not it was selected. When selected, the strategic adviser reported the colour and confidence honestly. When ignored, its strategy depended on the strength of the evidence: if the evidence it observed was strong (> 75% of balls it observed were of the same colour), the strategic adviser reported honestly. However, if the evidence it observed was weak (<= 75% of observed balls were of the same colour), it reported the colour of the minority of the balls with a (randomly sampled) confidence level of 2, 3 or 4. Its strategy is thus to deviate from the observed evidence whenever it was ignored and there was only weak evidence. The 20 rounds encompassed five "easy" rounds with 90 balls of one colour (either black or white), and 15 "difficult" rounds with 50 balls of one colour. The twenty rounds were shown in random order. We used a mix of easy and difficult rounds to increase variation in task difficulty within a participant for a more realistic sampling experience. Participants who started the experiment but did not finish were removed from all analyses. Prior to starting the experiment, participants were required to read the instructions. Participants needed to pass a comprehension check before being allowed into the experiment. The experiment took approximately 10 minutes and participants who successfully completed the experiment received a $3 participation fee. In total 28 individuals (mean ± standard deviation (S.D.) age = 36.1 ± 8.7 years; 25% female, and 75% male) completed the experiment.

Experiment 1. Experiment 1 (preregistration: https://osf.io/qkncz/) followed the same setup as the pilot experiment, with the exception that we used four levels of evidence, varying the ratio of black vs. white balls. All four treatments included five "easy" rounds with 90 balls of one colour (either black or white). The four treatments differed in the remaining 15 rounds. These rounds consisted either of 50, 60, 70 or 80 balls of one colour. From 50 to 80 balls of one colour, the outcome of the rounds becomes increasingly easier to predict. The five easy rounds and the remaining 15 rounds were shown in random order. We generated predictions for the different evidence levels using simulations (see preregistration https://osf.io/z8k3c/). Participants received a $3 participation fee, and an additional $0,10 for each correct outcome (i.e., a win). We collected data for 40 participants per treatment, resulting in 160 participants in total (mean ± S.D. age = 37.2 ± 10.3 years; 40% female, 59% male, and 1% other).

Experiment 2. Experiment 2 (preregistration: https://osf.io/rsn8h/) was similar to Experiment 1, with the exception that correct choices were not incentivized. That is, participants only received a $3 participation fee and did not receive a bonus payment for correct outcomes. This was done to test if we could replicate the results of Experiment 1 without incentivizing correct choices. In each treatment, we collected data for 35 participants, resulting in a total of 140 participants (mean ± S.D. age = 35.2 ± 9.5; 38% female, 62% male). The results of Experiments 1 and 2 showed that, as predicted, the strategic adviser gains the highest influence in the most uncertain condition (i.e., 15 rounds with 50/50 ratio of balls, and 5 rounds with 90/10 ratio; Figure 3B and 3C). We, therefore, continued with this condition in all subsequent experiments.

Experiment 3. Experiment 3 (preregistration: https://osf.io/bpngu/) was similar to Experiment 1, with the exception that we only collected data in the environment with the weakest evidence. This was done to test if we could replicate the effect in this environment once more. Participants received a $3 participation fee, and $0,10 for each correct decision. In total 45 individuals (mean ± S.D. age = 35.7 ± 10.8 years; 49% female, 51% male) completed the experiment.

Experiment 4. Experiment 4 (preregistration: https://osf.io/2ydh3/; screenshots: https://osf.io/hfkuy) investigated whether the strategic adviser can also gain influence in a group of participants whose decisions are combined under an anonymous majority vote. This experiment was done at the lowest evidence

level. We generated predictions for the influence of the strategic adviser in groups versus individuals using simulations (see https://osf.io/z8k3c/). We conducted two treatments: an individual and a group treatment. The individual treatment served as a control and was the same as in Experiments 1–3. In the group treatment, five participants performed the experiment together as a group. In each round, each of the five individuals made an individual decision which adviser to follow. The adviser chosen by most group members was selected, and all group members followed the selected adviser's recommendation. Group members only saw the outcome of the majority vote (i.e., which adviser was selected) but not the size of the majority nor the decisions of individual group members. In both treatments, participants could only enter the experiment if they completed a list of comprehension questions. This study was conducted in the lab. In the individual treatment, participants performed the experiment alone sitting behind a desktop. In the group treatment, participants started as soon as all five group members completed the comprehension test. In the group treatment, individuals worked independently using their own tablet. The tablets were controlled by a central server. Group members were sitting in the same room, and made aware that they were doing this experiment with the people in the same room. In both treatments, participants received a participation fee of €6 plus a bonus payment of €0.10 for each correct outcome. The study took approximately 15 minutes for the individual treatment and 20 minutes for the group treatment. For the individual treatment, we collected data for 60 participants (mean age = 27.6 ± 5.5 years; 59% female, 41% male). For the group treatment, we collected data on 30 groups of five individuals, resulting in 150 participants (mean ± S.D. age = 27.1 ± 5.1 years; 63% female, 35% male, and 2% other).

Experiment 5. Experiment 5 (preregistration: https://osf.io/z8k3c/) was a replication of Experiment 4 but conducted online to test if we could replicate our findings from Experiment 4 online. We again conducted an individual and a group treatment (group size five). In the group treatment, participants entered a virtual waiting room after completing the comprehension questions, and waited until they were paired with four other group members upon which the experiment started. As this was an online group study, we needed to implement a policy for non-responders. Participants that did not respond (i.e., did not decide which adviser to follow) within 10 seconds (except round 1: 30 seconds, and round 2: 20 seconds), were removed from the group ("drop-outs"), to assure that the experiment would move forward in the case of non-responders. Participants that dropped out of the experiment were not replaced, hence these groups continued with a smaller group size. Participants were not informed about the number of dropouts and in case of a tie (i.e., equal amount of support for both advisers), one of the advisers was selected randomly. In both treatments, participants received a $3 flat fee for participation, plus a bonus payment of $0,10 for each correct outcome. For the individual treatment, we planned to collect data for 50 individuals, and for the group treatment, we planned to collect data for 25 groups successfully completing the experiment. A successful completion was defined as having at least three participants remaining in the last round. In total 147 individuals completed the experiment: 50 singletons (mean ± S.D. age = 35.4 ± 9.5 years; 37% female, 61% male, and 2% other) and 97 individuals distributed over 25 groups (mean ± S.D. age = 36.2 ± 11.4 years; 42% female, 58% male). Five groups finished with five participants; twelve with four participants, and eight with three participants.

Experiment 6. Experiment 6 (preregistration: https://osf.io/8h47m/; screenshots: https://osf.io/b5gxq/) took place in the lab and investigated the strategic adviser's ability to gain influence in communicating groups. Participants performed the experiment in dyads, sitting together at one computer screen. Dyads were instructed to discuss their opinions with each other and reach a consensual agreement on which adviser to follow. Participants received a participation fee of €6 plus a bonus payment of €0.10 for each correct outcome. The study took approximately 20 minutes to complete. We did not perform an additional individual treatment, but used the individual treatment of experiment 4 as control because all participants were from the same participant pool (i.e., the participant pool of the lab of ARC of the MPIB). We collected data of 50 dyads, resulting in a total of 100 participants (mean ± S.D. age = 27.3 ± 5.7 years; 57% female, 41% male, and 2% other).

QUANTIFICATION AND STATISTICAL ANALYSIS

In the earliest preregistrations we announced that we would exclude participants that did not sample both advisers. In all seven studies we observed participants that did not sample both advisers but in all seven studies it was more likely that these participants always chose the strategic adviser and not the honest one (Figure 3). Therefore, we consider this behaviour a feature of participants' strategy and not a lack of engagement, and included these participants in the statistical analysis. For all statistical analyses we

used R (version 4.0.4). We used Bayesian hierarchical generalized linear models using the brm function from the brms package (Bürkner, 2017) and its default priors. For each model, we ran three chains in parallel with 6,000 iterations, of which the first 3,000 were discarded as burn-in to reduce autocorrelations. Visual inspection of the Markov chains and the Gelman-Rubin statistic (Rhat) indicated that all Markov chains converged. Unless stated otherwise, the points and error bars reported reflect the mean estimates and the 95% credible intervals (CI) of the posterior distribution.

Probability of selecting strategic adviser

To test if individuals were more likely to select the strategic adviser than the honest adviser, we fitted "Selected Adviser" (strategic or honest) as binomial response variable, and "Round Number" (either alone (Pilot + Exp. 3) or in interaction with Treatment (Exp. 1, 2, 4, 5 and 6)) as population-level ("fixed") effect(s). Since in the first round, individuals did not have any information about the advisers, and their choices were thus random, we fixed the intercept at Round 1 at 0.5. "Individual (or Dyad) Identity" was included as group-level ("random") effect. In the majority voting experiments (Experiments 4+5) "Individual Identity" was nested in "Group Identity". A preference of the strategic adviser over the honest adviser was inferred by evaluating whether there was a positive and credible (i.e., non-overlapping with 0) effect of "Round Number" on "Selected Adviser". See Table S3 for models results, and https://osf.io/9gjyc/ for data and code.

Probability of changing adviser

To test how lottery outcome and the ignored adviser's advice direction affected the likelihood to change adviser across the different treatments, we fitted "Changing Adviser" (yes/no) as binomial response variable, and "Lottery Outcome" (lost/won), "Ignored Adviser's Advice" (opposing/confirming), and their interaction, as population-level effects. To test how time affected the likelihood to change adviser, we also fitted "Round Number" as a population-level effect. "Individual (or Dyad) Identity" was, again, included as group-level effect. In the majority voting experiments (Experiments 4+5) "Individual Identity" was, again, nested in "Group Identity". As inference criterion, we evaluated whether the effects were credibly different from 0 (either the main effects or their interaction). See Table S4 for model results, and https://osf.io/9gjyc/ for data and code.