WINPEC Working Paper Series No. E2532
March 2026

Damned if You Do, Damned if You Don't: The Dilemma of Consistency and Revision in Expert Communication

Yoshio Kamijo Daiki Kishishita Satoru Shimokawa

Waseda INstitute of Political EConomy
Waseda University
Tokyo, Japan

Damned if You Do, Damned if You Don't: The Dilemma of Consistency and Revision in Expert Communication*

Yoshio Kamijo† Daiki Kishishita‡ Satoru Shimokawa§

March 11, 2026

Abstract

This paper identifies a fundamental expert communication dilemma: citizens distrust consistent advice as a signal of bias, yet distrust revision as indicating limited knowledge. We formalize this in a repeated cheap-talk model with bias uncertainty and gradually accumulating evidence. Theoretically, perfect compliance obtains without private information but collapses otherwise. Experiments reveal the dilemma is more severe than predicted, emerging even without private information. Importantly, private signals do not reduce welfare but instead filter incorrect advice. Finally, compliance substantially recovers with algorithmic advisors, suggesting automated advice mitigates communication failures in controversial policy environments.

Keywords: Reputational cheap talk; Expert advice; Informed receiver; Algorithmic advice; Experiment

JEL classification: D83; C92

*We are grateful to the participants at the 2025 Summer Workshop on Economic Theory (SWET) in Hokkaido and the 19th Annual Conference of the Association of Behavioral Economics and Finance at Waseda University for their valuable comments and suggestions. Generative AI tools, specifically Gemini and ChatGPT, were employed during the preparation of this manuscript: Gemini was used for language polishing and grammatical corrections, while ChatGPT assisted in translating the experimental instructions from Japanese into English. This work was supported by JSPS KAKENHI Grant Number 22K01397. All remaining errors are our own.

†Faculty of Political Science and Economics, Waseda University. 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo, Japan. 169-8050. E-mail: yoshio.kamijo@gmail.com

‡Graduate School of Economics, Hitotsubashi University. 2-1 Naka, Kunitachi, Tokyo, Japan. 186-8601. E-mail: daiki.kishishita@gmail.com

§Faculty of Political Science and Economics, Waseda University. 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo, Japan. 169-8050. E-mail: satoru.shimokawa@waseda.jp

1 Introduction

Public compliance with expert advice is a cornerstone of effective policymaking. During the COVID-19 pandemic, for example, governments and scientists issued recommendations on mask usage, physical distancing, and vaccination that were crucial for limiting the spread of infection (Adjodah et al., 2021). Yet compliance was far from universal, and a key driver of non-compliance was distrust in experts (Bargain and Aminjonov, 2020; Algan et al., 2021). Understanding the sources of this distrust is therefore of first-order importance. This paper develops a theoretical framework that identifies a novel dilemma in expert communication and evaluates its implications for non-compliance through laboratory experiments.

A striking and underappreciated feature of expert communication is that experts face distrust no matter what they do; that is, they are punished both for sticking to their advice and for revising it. First, prior research documents that some experts strategically distort information to advance their own interests (Oreskes and Conway, 2010; Wang et al., 2010; Bes-Rastrollo et al., 2013).[1] Thus, although citizens may wish to follow experts' advice, they also fear being misled and having their behavior manipulated. Specifically, if an expert's recommendations remain unchanged over time, citizens may suspect ulterior motives, interpreting consistency as evidence of bias.

Second, for many complex and rapidly evolving issues such as the COVID-19 pandemic, even experts possess only limited and provisional knowledge. Recommendations must be made based on the best available evidence, with the understanding that they may later be revised as new findings emerge (Intemann, 2023). For example, early in the COVID-19 pandemic, health authorities recommended against mask-wearing by the general public; within months, the same authorities reversed course and advocated universal masking (Intemann, 2023).[2] Although such revisions reflect responsible scientific practice, they may be perceived by citizens as inconsistency or incompetence, thereby eroding trust.

The result is a fundamental dilemma: citizens distrust consistency as a signal of bias, yet distrust revision as a signal of limited knowledge. Experts are, in a word, "damned if they do and damned if they don't."

Importantly, this dilemma is not unique to public health. A financial advisor who persistently recommends the same portfolio allocation may be suspected of earning commissions regardless of market conditions, while one who frequently revises recommendations may be seen as lacking genuine expertise. A climate scientist who maintains the same projections over decades may be accused of ideological commitment, while one who updates them in light of new data may be dismissed as unreliable. Wherever experts' motives are not fully transparent and their knowledge is gradually accumulated, the same tension between consistency and

[1] Bes-Rastrollo et al. (2013) show that studies funded by beverage companies are five times more likely to reject a link between sugary drinks and obesity than independent studies.

[2] Another example is that during the COVID-19 pandemic, early optimism about chloroquine (CQ) and hydroxychloroquine (HCQ) (Chen et al., 2020; Gautret et al., 2020) was tempered by subsequent analyses (Tang et al., 2020).

revision arises. Understanding this dilemma and its effect on public compliance with expert advice is therefore of broad importance across many domains.

Nevertheless, this dilemma has received little formal analysis. Existing models of strategic information transmission capture only one side of this problem (e.g., Sobel, 1985; Morris, 2001; Ottaviani and Sørensen, 2006*a*,*b*). That is, cheap-talk models with reputational concerns typically focus either on distrust in intentions or on uncertainty about expertise, but never both simultaneously. The reason is straightforward: in standard models, the state of the world is drawn independently each period, so there is no tension between consistency and revision—an advisor who changes recommendations is simply responding to a new state. Formalizing the consistency-revision dilemma requires a model in which the state is time-invariant and evidence accumulates gradually. Unlike canonical models with i.i.d. states, our assumption of a time-invariant state allows for the gradual accumulation of evidence, which is the prerequisite for the trade-off between perceived competence and perceived impartiality. The contribution of the present study is to formalize this dilemma in such a setting and design a novel framework for laboratory experiments.

For this purpose, we develop a repeated cheap-talk model, which we refer to as *the advisor–guesser game*. In each period, an advisor sends a binary recommendation to a guesser, who then chooses a binary action. The guesser's optimal action coincides with the unknown, time-invariant state of the world.

In every period, the advisor receives an imperfect signal about the state, capturing the idea that evidence is gradually accumulated over time. The advisor can be one of three types: a biased type who always prefers action 0 regardless of the state, a biased type who always prefers action 1 regardless of the state, or an unbiased type. The probabilities of being biased toward action 0 and toward action 1 are equal. The advisor's type is private information and unknown to the guesser. This feature captures the possibility that experts may have ulterior motives or ideological commitments that distort their recommendations.

Within this environment, we obtain the following results. First, when the guesser has no private information about the state—so that the advisor's recommendation is the sole source of information—there exists a unique equilibrium satisfying natural regularity conditions, namely *the truth-telling and obedience equilibrium*. In this equilibrium, the unbiased advisor reports truthfully, each biased advisor persistently recommends her preferred action, and the guesser always follows the recommendation. In this benchmark case, the dilemma in information transmission does not arise despite the two sources of distrust. Intuitively, in the absence of any alternative source of information, the guesser has no basis on which to deviate from the advisor's recommendation. Anticipating that her advice will be followed, the unbiased advisor truthfully communicates her signal, while biased advisors consistently advocate their preferred action.

However, the assumption that the advisor's recommendation is the only available source of information may be unrealistic depending on contexts. Even ordinary citizens often have access their own information, which may reflect *naive advice* from peers or their own experi-

ence rather than *professional advice* (Schotter, 2003). To capture this effect of naive advice, we extend the model to allow the guesser to receive private signals about the state. This seemingly modest change fundamentally alters equilibrium behavior. The guesser may now rationally reject advice in two distinct situations: when recommendations change abruptly in ways that contradict private signals (reflecting distrust in knowledge), and when recommendations remain excessively persistent despite opposing private evidence (reflecting distrust in intentions). Once the guesser's private information becomes sufficiently informative, the benchmark equilibrium collapses. The dilemma in information transmission emerges endogenously: both consistency and revision can trigger noncompliance. This suggests that citizens' acquisition of private information, such as through social media platforms, may undermine the effectiveness of public information transmitted by experts.

Whether these theoretical implications hold in practice remains an open question. Individuals may deviate from them due to behavioral biases. Moreover, the environment with alternative sources of information is analytically complex and does not admit a tractable closed-form equilibrium characterization. We therefore complement our theoretical analysis with laboratory experiments designed to assess how citizens respond to expert advice under varying degrees of private information and potential bias.

The experiment uses a ball-and-pot design that cleanly implements the model. A hidden pot is either blue or red; the advisor draws balls from the pot and sends a color recommendation; the guesser observes the recommendation, possibly draws her own ball, and guesses the pot's color. The experiment employs a two-by-two design that manipulates two factors: whether the guesser has access to private information and the probability that the advisor is unbiased.

In addition, we include treatments in which the advisor is an automated agent rather than a human. This design serves two purposes. First, it disciplines advisor behavior to conform to theoretical benchmarks, providing a cleaner test of guesser-side compliance. Second, and independently, it allows us to test whether the source of advice—human versus machine—affects trust and compliance, a question of growing practical relevance as AI-generated recommendations become increasingly prevalent.

The main findings can be summarized as follows. First, the consistency-revision dilemma is more severe than theory predicts. Even when guessers have no private information, where the theory predicts full compliance, participants follow advice only about 75% of the time. Compliance falls significantly when advice is revised across periods and also when the same advice is repeated for many consecutive periods, confirming that both revision and consistency independently trigger skepticism.

Second, theory predicts that the presence of additional information sources may undermine information transmission and potentially reduce the guesser's welfare. However, we do not observe such an effect in the data: there is no evidence that the rate of correct guesses declines when additional information sources are available.

Finally, when the advisor is an automated agent, guessers are more likely to follow the advice, and the dilemma is less pronounced. Our results suggest that participants exhibit

relatively high compliance with machine-generated advice, providing little support for the algorithm aversion documented in prior studies (Ishowo-Oloko et al., 2019; Commerford et al., 2022). Therefore, algorithmic advice may help mitigate communication failures in controversial policy environments.

Related literature: This study contributes to and bridges three strands of literature on strategic communication with partially informed receivers, dynamic information transmission with reputation, and advice-taking behavior.

First, following Crawford and Sobel (1982), strategic information transmission in static cheap-talk environments has been extensively studied both theoretically and experimentally (see Blume, Lai and Lim (2020) for a comprehensive review). Experimental contributions include, among others, Dickhaut, McCabe and Mukherji (1995), Blume et al. (1998), Cai and Wang (2006), and Jin, Luca and Martin (2021).

Within this literature, several studies theoretically analyze cheap-talk games with partially informed receivers (Ishida and Shimizu, 2016; Lai, 2014; de Barreda, 2024), showing that greater precision in receivers' private signals can impede information transmission. This result resonates with our theoretical prediction. However, these models are static: they cannot generate the consistency-revision dilemma, which is inherently dynamic. We provide the first framework in which both consistency and revision are simultaneously penalized. Moreover, while prior work is largely theoretical,[3] our study directly tests this insight in laboratory experiments and shows that this dilemma indeed arises, and that its presence is independent of whether receivers have access to alternative sources of information.

Second, our study is related to the literature on repeated cheap-talk games with reputational concerns. Similar to our framework, several theoretical contributions examine environments in which the sender may be biased or unbiased (Sobel, 1985; Benabou and Laroque, 1992; Morris, 2001). Furthermore, Ettinger and Jehiel (2021) and Choi, Lee and Lim (2025) provide experimental evidence based on this class of models.[4] They share our concern with advisor bias, but assume states are *i.i.d.* across periods. This assumption rules out the gradual accumulation of evidence that drives our results. Our key theoretical novelty is a time-invariant state, which endogenously generates a penalty for updating advice over time. Furthermore, our laboratory experiments succeed in documenting that guessers avoid overly

[3] Charness and Grosskopf (2004) conduct a laboratory experiment in which participants play a coordination game with pre-play cheap-talk communication about which action to take. They examine how outcomes depend on whether receivers observe the sender's past actions. However, their setting differs substantially from the communication games a la Crawford and Sobel (1982). Relatedly, a strand of literature experimentally examine the effect of transparency about conflicts of interests on strategic information transmission (e.g., Loewenstein, Cain and Sah, 2011; Kartal and Tremewan, 2018). However, their analysis focuses on information about the bias rather than information about the underlying state.

[4] Chung and Harbaugh (2019) also conduct a laboratory experiment in a static framework in which the sender's bias is uncertain. Another strand of the reputation literature assumes uncertainty about the sender's ability rather than her bias (Ottaviani and Sørensen, 2006*a*,*b*). Meloso, Nunnari and Ottaviani (2023) test this type of model in a laboratory experiment. Experiments on repeated cheap-talk games without reputation include Wilson and Vespa (2020) and Huang and Lim (2025).

consistent advice.

Finally, our model simultaneously incorporates professional advice and naive advice and provides a framework for studying how their interaction shapes individuals' compliance behavior. While previous studies examine each type in isolation using different models (e.g., Schotter (2003) and Çelen, Kariv and Schotter (2010) for naive advice, and Li, Özer and Subramanian (2022) and Meloso, Nunnari and Ottaviani (2023) for expert advice), we consider both types within a unified framework by defining expert advice as highly accurate but potentially biased, and naive advice as less accurate but unbiased. Our experimental results do not suggest that exposure to naive advice improves individuals' welfare or increases their likelihood of following professional advice. However, naive advice proves useful for identifying incorrect professional advice.

Taken together, our contributions are threefold. First, we identify a novel dilemma in dynamic information transmission whereby experts are penalized both for maintaining consistent advice and for updating it. Second, we develop a theoretical framework that isolates this mechanism in a repeated cheap-talk environment with time-invariant states and gradual evidence accumulation. Third, we provide laboratory evidence and document how the interaction between professional and naive advice shapes compliance behavior. More broadly, our findings underscore the importance of dynamic trust formation in expert communication.

The remainder of this paper is organized as follows. Section 2 introduces the advisor–guesser game and conducts the theoretical analysis. The experimental design is detailed in Section 3, followed by a discussion of the experimental results in Section 4. Section 5 concludes the paper.

2 Theory

2.1 Model

The advisor–guesser game is a two-player dynamic game over multiple periods ($t = 1, ..., T$), where one player is assigned the role of an advisor and the other acts as a guesser. The role is fixed across the periods.

Information structure: There is a binary state of the world: $\omega \in \{B, R\}$, which will correspond to the color of a pot (blue vs. red) in the subsequent laboratory experiment. This state is fixed across periods, and the prior probability of ω being R is $1/2$.

At the beginning of each period, the advisor receives $m > 0$ private signals regarding the state ω. Let $s_{it}^A \in \{B, R\}$ denote the i-th private signal received by the advisor in period t. The signals are conditionally independent and satisfy

$$\Pr(s_{it}^A = B \mid \omega = B) = \Pr(s_{it}^A = R \mid \omega = R) = \frac{3}{5}.$$

The precise value of the signal accuracy is not essential for the theoretical analysis; we adopt this parameterization because it will be implemented in the laboratory experiment described below. The signals received by the advisor are not observable to the guesser.

Before taking an action, the guesser receives $n \in \{0, \ldots, m-1\}$ private signals regarding the state ω. Let $s_{it}^G \in \{B, R\}$ denote the i-th private signal received by the guesser in period t. These signals have the same accuracy, $3/5$, and are not observable to the advisor. Since $n < m$, the advisor has an informational advantage over the guesser.

Actions: In each period, after receiving her private signals, the advisor sends a binary recommendation (i.e., a hint) to the guesser, $h_t \in \{B, R\}$.[5] After observing this recommendation and her own private signals, the guesser chooses an action $a_t \in \{B, R\}$, which represents her guess about the value of ω. This sequence constitutes a single round of the game. Although this stage game is repeated $T > 0$ times, ω remains unchanged.

Payoff structure and advisor's type: The payoffs for the advisor and guesser depend solely on the guesser's decision. The guesser's payoff is straightforward. If her guess regarding ω is correct, she receives $b_G > 0$. However, if her guess is incorrect, she receives no payoff. That is, the guesser's payoff is given by

$$\sum_{t=1}^{T} \mathbf{1}\{a_t = \omega\} b_G.$$

In contrast, multiple types of advisors exist, and each type has a different payoff function. There are three types of advisors: blue-biased, red-biased, and unbiased. The *blue-biased advisor* receives $b_B > 0$ if the guesser chooses $a_t = B$ as her guess and receives nothing otherwise. That is, the blue-biased advisor's payoff is given by

$$\sum_{t=1}^{T} \mathbf{1}\{a_t = B\} b_B.$$

Conversely, the *red-biased advisor* receives $b_R > 0$ if the guesser chooses $a_t = R$ and receives nothing otherwise. That is, the red-biased advisor's payoff is given by

$$\sum_{t=1}^{T} \mathbf{1}\{a_t = R\} b_R.$$

We refer to these two types of advisors as biased advisors, and they have an incentive to

[5] We restrict the advisor's message space to binary recommendations. This assumption is motivated by a common feature of real-world expert communication: even when experts possess rich and nuanced information, communicating its full detail to non-specialist audiences is often impractical. In many policy contexts, experts distill their assessments into a single actionable recommendation rather than conveying the underlying evidence. Our binary message space captures this essential feature of expert communication.

steer the guesser toward their preferred directions regardless of the value of ω. These types capture situations in which the advisor has ideological or financial interests that distort their recommendations.

In contrast, the *unbiased advisor* receives $b_W > 0$ if the guesser's guess is correct (i.e., $a_t = \omega$) and receives nothing otherwise. That is, the advisor's payoff is given by

$$\sum_{t=1}^{T} \mathbf{1}\{a_t = \omega\} b_W.$$

Thus, the unbiased advisor aims for the guesser to make the correct guess.

The type of advisor is determined at the beginning of the game, but this information is kept private from the guesser. The probability of an advisor being unbiased is $p \in (0, 1)$, while the probabilities of being a blue-biased or red-biased type are each $(1 - p)/2$.

A number of theoretical models on reputation building have examined settings where the alignment of preferences between the sender (i.e., advisor) and the receiver (i.e., guesser) is uncertain for the receiver (e.g., Sobel, 1985; Benabou and Laroque, 1992; Morris, 2001). However, these studies typically assume that the state of the world ω is independently drawn in each period. The novelty of our study is to consider the case in which ω is time-invariant, so that information gradually accumulates for the sender over time.

Timing of the game: The sequence of events in each period t is summarized as follows:

1. The advisor observes m number of private signals about ω: $(s^A_{1t}, ..., s^A_{mt})$. Then, the advisor provides a recommendation h_t to the guesser.

2. The guesser receives the recommendation from the advisor, and observes n number of private signals about ω: $(s^G_{1t}, ..., s^G_{nt})$. Then, the guesser takes a binary action, a_t.

The payoffs are realized not at the end of each period, but at the end of the entire game.

2.2 Equilibrium

A distinctive feature of our framework is the time-invariant nature of the state ω, which allows for the gradual accumulation of evidence over time. Unlike canonical reputational models where states are *i.i.d.* across periods, this setting endogenously generates a dynamics of trust in reliability of the advisor's knowledge. In our model, the advisor's history serves as a reference point: a revision of advice suggests a prior lack of information, while rigid consistency triggers suspicions of strategic bias.

We characterize a pure-strategy perfect Bayesian equilibrium of this game.

Equilibrium refinement: Cheap-talk games generically admit multiple equilibria, including uninformative babbling equilibria in which messages carry no meaning, and perverse

equilibria in which the language is inverted so that the hint "Blue" is understood to mean "choose Red" (Crawford and Sobel, 1982). To focus on economically meaningful equilibria, we impose two restrictions that formalize two natural requirements on how players use language and information.

Let $a^t := (a_1, ..., a_t)$ be the sequence of actions chosen by the guesser and $h^t =: (h_1, ..., h_t)$ be the sequence of hints. Furthermore, let

$$\Delta_t := |\{(i,\tau) \in \{1,...,m\} \times \{1,...,t\} | s^A_{i\tau} = B\}| - |\{(i,\tau) \in \{1,...,m\} \times \{1,...,t\} | s^A_{i\tau} = R\}|.$$

That is, Δ_t is the number of blue signals minus red signals observed by the advisor until period t. Based on these notations, we let $h^*_{Ut}(\Delta_t, h^{t-1}, a^{t-1}) \in \{B, R\}$ be the equilibrium strategy of the unbiased advisor and $h^*_{kt}(\Delta_t, h^{t-1}, a^{t-1}) \in \{B, R\}$ be the equilibrium strategy of k-biased advisor ($k \in \{B, R\}$).

Assumption 1. *We focus on the equilibrium where*

(i). *for any* (h^{t-1}, a^{t-1})*, there exists* $\bar{\Delta}_t(h^{t-1}, a^{t-1}) \in \{-m(t-1)+1, ..., m(t-1)\}$ *such that* $h^*_{Ut} = B$ *if* $\Delta_t > \bar{\Delta}_t(h^{t-1}, a^{t-1})$ *and* $h^*_{Ut} = R$ *if* $\Delta_t < \bar{\Delta}_t(h^{t-1}, a^{t-1})$*, and*

(ii). *the biased advisors' equilibrium strategies are independent of* Δ_t.

Condition (i) imposes two requirements. First, it implies the threshold property such that the unbiased strategy is always *monotonic* with respect to his or her private information. In general, cheap talk games face message indeterminacy (i.e., the meaning of messages is arbitrary); thus, there is an unreasonable equilibrium where hint R means "action B should be chosen" and hint B means "action R should be chosen." Condition (i) eliminates the possibility of such an equilibrium by fixing the interpretation of the language used by players.

Second, condition (i) also requires that $\bar{\Delta}$ takes an interior value, implying that the unbiased sender never takes the perfectly uninformed strategy of sending a specific hint regardless of their private information. This rules out a possibility of the babbling equilibrium.

Furthermore, condition (ii) requires that biased advisors' strategies be independent of their private signals. This reflects the fact that the state is payoff-irrelevant for biased advisors—a blue-biased advisor earns a reward whenever the guesser chooses Blue, regardless of the true state—so there is no reason for their behavior to vary with Δ_t.

In the following, we characterize the equilibrium satisfying these regularity conditions.

2.2.1 Equilibrium with $n = 0$ (case with uninformed guesser)

We start with the case where the guesser cannot obtain any private information about ω (i.e., $n = 0$). In such a case, the equilibrium is characterized as follows:

Proposition 1. (Full compliance of uninformed guessers).

In the perfect Bayesian equilibrium satisfying Assumption 1,

(i). *The guesser always follows the advisor's hint;*

(ii). *The blue (resp. red)-biased advisor always sends $h_t = B$ (resp. R); and*

(iii). *The unbiased advisor sends $h_t = B$ if $\Delta_t > 0$ and $h_t = R$ if $\Delta_t < 0$.*

Proof. See the Appendix. □

Therefore, in the unique equilibrium, the guesser follows the advice, and the unbiased advisor sends the hint sincerely based on his or her private information. Hereafter, we will call this equilibrium *the truth-telling and obedience equilibrium* (hereafter, *TO-equilibrium*).

Given that the guesser follows the advice, it is straightforward that the red (resp. blue)-biased advisor sends a red (resp. blue) hint, and the unbiased advisor sends a hint that is likely to be true. Therefore, the key is the guesser's incentive to follow the advice.

To understand why the guesser follows the advice, note that the advisor is unbiased with positive probability. As a result, the advisor's hint is always partially informative about ω. Given that the prior probability of $\omega = B$ is 0.5 and the guesser has no private information about ω, the advisor's hint is the only useful input for decision-making. Therefore, the guesser always follows the advice.

Notably, this conclusion always holds as long as $n = 0$, irrespective of the value of $p \in (0, 1)$. Even if the advisor is highly likely to be biased, the guesser follows the advice.

In summary, the guesser does not face the dilemma in information transmission discussed in the introduction: in this environment, the guesser has no meaningful alternative to following the advisor's recommendation, regardless of whether the advice is revised or remains consistent. Consequently, the advisor also does not face the dilemma.

2.2.2 Equilibrium with $n > 0$ (case with partially informed guesser)

Next, we examine the case where the guesser obtains their own private information about ω (i.e., $n > 0$). In such a case, the above discussion does not straightforwardly apply; that is, the TO-equilibrium does not necessarily exist.

From the proof of Proposition 1, it is straightforward that the advisors follow the specified strategy as long as the guesser always follows the advisor's hint. Therefore, it suffices to examine the guesser's incentive. Given this consideration, to characterize when the TO-equilibrium breaks down, we ask whether the guesser has an incentive to deviate from the equilibrium strategy—that is, to ignore the advisor's recommendation—taking the advisors' strategies as given. Lemmas 1 and 2 identify two distinct situations in which such a deviation is individually rational. Together, they establish the formal basis for Proposition 2.

The first situation arises when the advisor revises her recommendation. A revision carries two pieces of information simultaneously, and they pull in opposite directions. On the one hand, because biased advisors never revise—they persistently advocate their preferred action—a revision is a signal that the advisor is unbiased, which speaks in favor of compliance.

On the other hand, a revision implies that the cumulative signal Δ_t crossed zero between periods, meaning the advisor's private evidence only weakly favors the current recommendation. When the guesser's own private signals point in the opposite direction, this weak reliability may be insufficient to override her private information. The net effect can therefore make deviation individually rational.

Lemma 1 formalizes this tension. It shows that, taking the equilibrium strategies of Proposition 1 as given, a guesser who observes a revision may strictly prefer to follow her own signal rather than the advisor's hint, provided the guesser's information is sufficiently precise in aggregate ($nT > m$).

Lemma 1. (Revision of advice may be penalized). *Assume that the unbiased and biased advisors follow the strategy profile specified in Proposition 1. Furthermore, suppose that $h_t \neq h_{t-1}$ for some t. Then, the guesser does not always follow h_t if $nT > m$.*

Proof. See the Appendix. □

Intuitively, a revision of advice (a "flip") signals that the advisor is likely unbiased but also reveals that their informational confidence—represented by a low absolute value of Δ_t—is minimal. Consequently, even a small amount of private information held by the guesser can be sufficient to override the advisor's recommendation.

The second situation arises when the advisor maintains consistent recommendations over multiple periods. Consistency is also a double-edged signal. An unbiased advisor who genuinely observes evidence favoring the same action repeatedly will naturally be consistent—so consistency can reflect reliability. But biased advisors are also maximally consistent, always recommending their preferred action regardless of evidence. Thus, observing the same recommendation for many consecutive periods increases the posterior probability that the advisor is biased, eroding the credibility of the hint. Again, when this erosion is severe enough relative to the guesser's private signals, deviation becomes individually rational.

Lemma 2 formalizes this second source of breakdown, showing that excessive consistency can also trigger non-compliance when the prior probability of bias is sufficiently high (i.e., p is sufficiently small). Let $L_t(\omega) := \Pr(\Delta_\tau \geq 0 \; \forall \tau \leq t \mid \omega)$, which represents the probability that Δ_τ continues to be positive up to period t given state ω.

Lemma 2. (Consistency of advice may be penalized). *Assume that the unbiased and biased advisors follow the strategy profile specified in Proposition 1. Furthermore, suppose that h_τ is the same for any $\tau \in \{1, ..., t\}$. Then, the guesser does not always follow h_t in period t if*

$$\frac{p}{1-p} < \frac{\left(\frac{3}{2}\right)^{nt} - 1}{2L_t(B) - 2L_t(R)\left(\frac{3}{2}\right)^{nt}} \text{ or } L_t(B) \leq L_t(R)\left(\frac{3}{2}\right)^{nt}$$

holds.

Proof. See the Appendix. □

Although the condition is much more complicated than the condition in Lemma 1, a sufficient condition is that p is sufficiently small. While n and T must be sufficiently large for the deviation in Lemma 1, it is not necessary for the deviation in Lemma 2. Only small p (i.e., the high initial probability of the advisor being biased) is sufficient for the deviation in this lemma as long as $n > 0$. Specifically, the right-hand side in the condition is greater than

$$\frac{1}{2}\left[\left(\frac{3}{2}\right)^{nT} - 1\right].$$

Thus, as long as

$$\frac{p}{1-p} \leq \frac{1}{2}\left[\left(\frac{3}{2}\right)^{nT} - 1\right]$$

holds,[6] the condition in Lemma 2 is always satisfied.

In summary, the guesser does not necessarily follow the advisor's recommendation, either when the advisor revises the recommendation or when the guesser repeatedly observes the same recommendation. In this sense, the guesser faces a dilemma in trusting the advisor. Anticipating this response, the advisor in turn faces a dilemma of whether to revise advice or maintain consistency. As a result, the strategies characterized in Proposition 1 no longer constitute an equilibrium. We summarize this observation as follows:

Proposition 2. (Non-compliance of partially informed guessers).
Suppose that $n > 0$. TO-equilibrium does not exist either when n and T are sufficiently large or when p is sufficiently small.

As a direct consequence of the proposition, we also obtain the following corollary:

Corollary 1. *When $n = 0$, the guesser follows the hint with probability one. On the contrary, when $n > 0$, the guesser may not follow the hint.*

Taken together, the above analysis reveals that the dilemma in information transmission arises only when the guesser has access to an information source other than the advisor.

3 Experimental design

The environment with $n > 0$ does not admit a tractable closed-form equilibrium characterization, a difficulty that itself reflects the strategic complexity inherent in real-world expert

[6] Later, in the experiment, we use $T = 10$ and $n = 1$. In such a case,

$$\frac{p}{1-p} \leq \frac{1}{2}\left[\left(\frac{3}{2}\right)^{nT} - 1\right] \Leftrightarrow p < 0.966.$$

This is a fairly weak condition.

communication. Rather than treating this as a limitation, we adopt a complementary approach: theory identifies the conditions under which the truth-telling and obedience equilibrium collapses, while experiments reveal the behavioral patterns that emerge in its absence.

3.1 Basic setting

We implemented the advisor-guesser game using a ball-and-pot design. The state $\omega \in \{B, R\}$ corresponds to the color of a virtual pot (see Figure 1 (a)). A blue pot ($\omega = B$) contains three blue balls and two red balls, while a red pot ($\omega = R$) contains three red and two blue balls. Consequently, any single draw yields a signal with an accuracy of 0.6. The state ω is determined at the beginning of each session and remains constant across all T periods, allowing for the gradual accumulation of information.

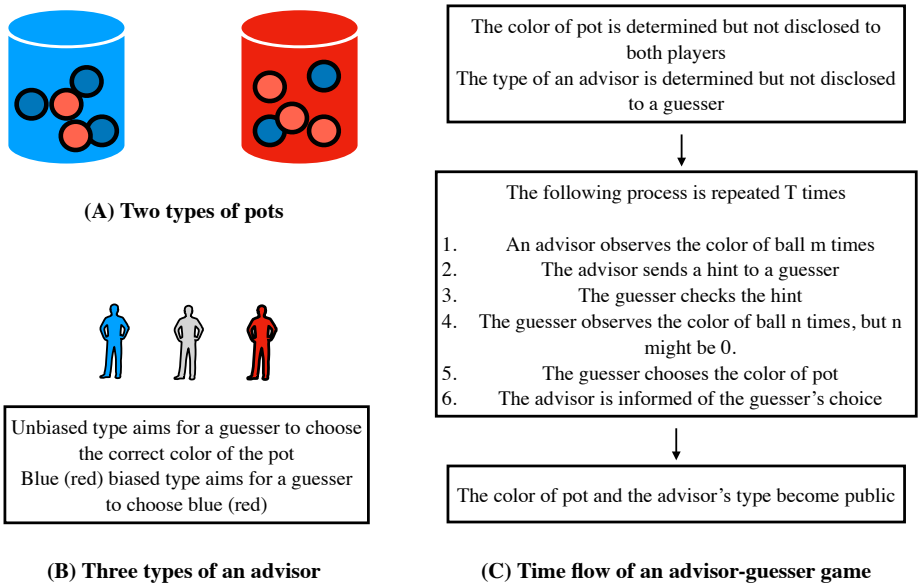

Figure 1: Game structure

The color of the pot is determined at the beginning of the game and remains unchanged across the periods. The probability of selecting a blue or red pot is equal (50%). However, neither the advisor nor the guesser knows the color of the pot.

Instead of directly observing the color of the pot, the advisor draws a ball, checks its color, and returns it to the pot. This process is repeated $m > 0$ times in each period. Since the ball is put back in the pot each time, the probability of drawing a red ball or a blue ball remains the same throughout each trial. In other words, in each period, the advisor observes m number of signals about ω with accuracy $3/5$. After observing the colors of the balls, the advisor sends a message to the guesser regarding the color of the pot. The message can be either blue or red.

Next, the guesser takes a turn. In addition to checking the message from the advisor, the guesser can observe the color of the ball drawn from the same pot n times. In other words, in each period, the guesser observes n number of signals about ω with accuracy $3/5$. Afterward, the guesser makes a final decision regarding the color of the pot, choosing from either blue or red.

This sequence is repeated T times, although the color of the pot remains unchanged. Note that the color of the pot remains undisclosed to both the advisor and the guesser, meaning that it is unknown whether the recommendation and guess are correct. This sequence is summarized in Figure 1 (c).

3.2 Experimental conditions

We manipulated three dimensions in our experiments:

Presence of alternative information source: First, to explore the influence of the alternative information source available to the guesser, we introduced two distinct values for the parameter n: 0 and 1. $n = 0$ means that the guesser cannot access an alternative source other than the advisor's type, whereas $n = 1$ means that the guesser can observe a single private signal in each period. Although appearing as a subtle differentiation, it is of considerable importance, as highlighted by Propositions 1 and 2.

Uncertainty about advisor intent: The second treatment variable is the probability that an advisor is unbiased, i.e., p. We specifically chose three values: $p = 0.5$, $p = 0.8$ and $p = 1$. These values represent scenarios where the level of trust in advisors' intentions ranges from low to very high.

Human vs. computer bot: Third, we also included a treatment in which the advisor is the algorithmic advisor (hereafter "bot") rather than a human participant. This design allows us to discipline the advisor's behavior to follow equilibrium strategies, thereby providing a cleaner test of whether guessers themselves behave in accordance with equilibrium predictions.

The unbiased bot advisor generates recommendations according to Bayes' rule, using the colors of previously drawn balls to infer the most likely pot color. When the posterior probabilities of the two colors are equal, the bot simply repeats its previous hint. Real human advisors do not implement optimal strategies perfectly; even unbiased advisors occasionally send a hint that does not reflect their best assessment of the evidence. To ensure that the bot's behavior is comparable to that of human advisors in this respect, we program the unbiased bot to make an error with 10% probability, recommending the less likely color rather than the more likely one. Similarly, the biased bot advisor always recommends its preferred color, independent of the observed balls, again with a 10% probability of error. Participants assigned

to the guesser role are fully informed about the bot advisor's decision rules, including the possibility of mistakes.

Beyond disciplining advisor play to conform to theoretical benchmarks, the contrast between human and bot advisors allows us to directly investigate the degree of trust in algorithmic advice. While a growing body of literature documents "algorithm aversion"—a tendency for individuals to discount machine-generated advice more heavily than human advice after observing errors (e.g., Ishowo-Oloko et al. 2019)—our design tests whether the transparency of a bot's non-strategic decision rule can actually enhance trust by mitigating the consistency-revision dilemma.

Other parameters: Regarding the number of balls observed by the advisor in each period and the number of periods, we chose $m = 3$ and $T = 10$ or 20. This selection is intentional, as it ensures that even in cases involving an unbiased advisor, discerning the true color of the pot at the beginning of the game remains challenging. However, as the game progresses, this design allows for an improvement in predictive accuracy. Note that when $n = 1$ and $T = 10$ or $T = 20$, the conditions identified in Lemmas 1 and 2 hold as long as $p \neq 1$. Therefore, the theoretical prediction is that TO-equilibrium is played only when (i) $n = 0$ or (ii) $n = 1$ and $p = 1$.

Assignment of each condition: Our experiment consists of eight conditions in total, summarized in Table 1. The core is a 3 × 2 factorial design holding $n = 0$, which varies the advisor type (human vs. bot) and the probability of an unbiased advisor ($p \in \{0.5, 0.8, 1\}$). In addition, we included two extension treatments featuring $n = 1$ for human advisors with $p = 0.5$ and $p = 0.8$ to examine how access to private information alters compliance.

Table 1: Conditions of experiment

Condition	Advisor	n	p	m	T	N of Sbj.
Human-50	Human	$n = 0$	$p = 0.5$	$m = 3$	$T = 10$	42
Human-80		$n = 0$	$p = 0.8$			42
Human-OB-50		Observable ($n = 1$)	$p = 0.5$			40
Human-OB-80		Observable ($n = 1$)	$p = 0.8$			40
Human-100		$n = 0$	$p = 1$		$T = 20$	38
Bot-50	Bot	$n = 0$	$p = 0.5$			20
Bot-80		$n = 0$	$p = 0.8$			20
Bot-100		$n = 0$	$p = 1$			18

Note: p = the prior probability of an advisor being an unbiased type. n = the number of balls a guesser can observe in a period. m = the number of balls an advisor can observe in a period. T = the length of a game.

3.3 Incentive structure and risk neutrality

To control for subjects' risk attitudes, we implemented a binarized lottery incentive scheme (Harrison, Martínez-Correa and Swarthout, 2013; Hossain and Okui, 2013). Each correct guess earned points that linearly increased the probability of winning a fixed monetary prize of 500 JPY. As shown below, this linear payoff structure ensures that an expected-utility-maximizing participant should choose the color with the higher subjective posterior probability, regardless of the curvature of their utility function.

Consider first the case $T = 10$. In this case, a bonus of $100/T = 10$ points was awarded for each successful response in every period; that is, we set $b_G = b_R = b_B = b_W = 10$. The accumulated points determine the probability of receiving a fixed reward of 500 Japanese yen (JPY). For example, if a guesser answers correctly in 7 out of 10 periods, she earns 70 points, corresponding to a 70% chance of receiving the additional reward.

Under this reward scheme, participants' choices are independent of the curvature of their utility function over monetary payoffs. To see this, let n_r denote the number of times the guesser chooses "red" at time t, and let $n_b = t - n_r$ denote the number of times she chooses "blue." Let α be the subjective probability that the pot is red at the time of choice, with $1 - \alpha$ denoting the probability that it is blue. The expected payoff is then

$$(n_r/10)\alpha U(500 \text{ JPY}) + (n_b/10)(1 - \alpha)U(500 \text{ JPY}).$$

Because this expression is linear in n_r and n_b, the guesser maximizes expected payoff by choosing the color associated with the higher subjective probability.

When $T = 20$, the bonus is adjusted to $100/T = 5$ points.

3.4 Post-experiment belief elicitation

After completing the game, but before receiving any feedback, guessers in treatments with $p < 1$ were asked to report their posterior beliefs about whether their advisor was unbiased. To incentivize truthful belief elicitation, we employed a multiple pricing list method. Specifically, guessers were repeatedly presented with two options: (A) a payment of 200 JPY if their advisor was unbiased, and (B) the same payment delivered with probability q, independent of the advisor's type. Given a prior probability p—set to either 0.5 or 0.8—the probability q in Option (B) varied from $p - 0.2$ to $p + 0.2$ in increments of 0.02.[7] Accordingly, the first question involved a choice between options (A) and (B) with $q = p - 0.2$, while the final (21st) decision corresponded to $q = p + 0.2$. Among the 21 questions, one was randomly selected to determine the guesser's final payment.

[7] If the subjective posterior beliefs are in the range $(p - 0.2, p + 0.2)$, guessers should switch from option (A) to option (B) at some point, since the probability of winning in option (B) is higher in the later questions. We constrained them to switch from (A) to (B) at most once due to the interface design of the experiment. This was done to prevent people from switching repeatedly, or conversely from (B) to (A), and to guide them toward correct understanding of this questionnaire.

By contrast, participants assigned to the advisor role were incentivized to predict whether their paired guesser perceived them as more unbiased than the prior probability specified at the beginning of the game.[8]

3.5 Protocol

Participants were recruited from Sona Systems, a standard research participant pool that includes both undergraduate and graduate students at Waseda University in Japan.[9] While IRB approval was not required at this university, we strictly adhered to the standard protocols of experimental economics.

A total of 262 individuals participated in the experiment, with 103 females, 158 males, and 1 participant whose sex was not specified. The average age of the participants was 21.13 years. The experiment was conducted in January and July 2022, January 2023, and November 2025. Importantly, owing to the between-subjects design, no participant took part in more than one session.

The advisor–guesser game constitutes a novel experimental paradigm and therefore involves an exploratory component (for example, equilibrium strategies are not theoretically characterized when $n > 0$). Consequently, this study was not preregistered. Participant assignment was designed to yield approximately 20 guessers per experimental condition; the exact sample sizes are reported in Table 1.

Because the initial sessions were conducted during the COVID-19 pandemic, all sessions were held online. We conducted the experiment via Zoom in combination with oTree (Chen, Schonger and Wickens, 2016). Upon entering the Zoom meeting, participants were instructed to change their display names to their assigned experimental IDs. To maintain communication control and ensure anonymity, participants were required to keep their webcams off and microphones muted throughout the session.

Once all participants confirmed their readiness, the experimenter distributed a link to the oTree application. Within the platform, participants proceeded through the instructions at their own pace, followed by a comprehension quiz before starting the task. To mitigate potential declines in engagement in the online setting, the comprehension quiz was incentivized (with a maximum bonus of 200 JPY). After completing the experiment, participants filled out a post-experimental questionnaire that collected demographic information and additional survey responses.

Participants received a fixed participation fee of 1,500 JPY, in addition to performance-based earnings from the main task, the comprehension quiz, and the belief elicitation task. On average, total compensation amounted to 2,120 JPY (approximately 18.4 USD, 16.0 USD, and 13.7 USD based on the exchange rates in January 2022, January 2023, and November 2025, respectively). Each session lasted approximately 70 minutes.

[8] In the treatments with bots, advisors were not human participants; thus, this procedure was omitted.

[9] This is a system for participant management. Refer to https://www.sona-systems.com/ for more details.

The English translation of the instructions is provided in Appendix C.

4 Results

4.1 Data summary: Compliance rates, accuracy, and deviations from theoretical predictions

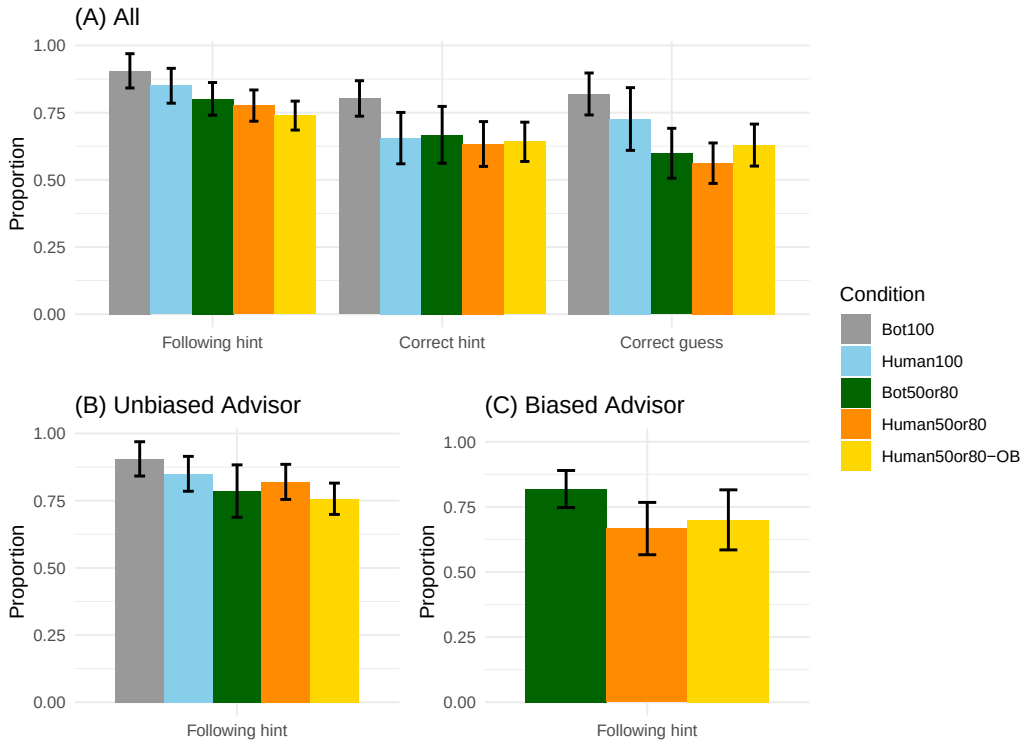

Figure 2: Proportion of compliance, correct hint and correct guess

Note: Error bars indicate 95% confidence intervals. Since each individual provides multiple observations, intervals are calculated based on individual-level clustering.

Panel (A) of Figure 2 reports three measures: (i) the proportion of guesses that followed the advisor's hint ("Following hint"), (ii) the proportion of hints that correctly matched the true color of the pot ("Correct hint"), and (iii) the proportion of guesses that correctly identified the true color ("Correct guess"). Although the experiment consists of eight conditions, we pool the $p = 0.5$ and $p = 0.8$ treatments for ease of visual presentation.

Across all five cases, the rate of following the advisor's hint ranged between approximately 70% and 90%. The theory predicts full compliance (100%) in all cases except when guessers have access to alternative information sources (the Human 50-OB and Human 80-OB conditions). In contrast to this prediction, compliance is far from perfect. In particular, in the Human 50 and Human 80 conditions, the compliance rate is only about 75%.

Moreover, the theory predicts no difference between human and bot advisors and no difference between $p = 1$ and $p < 1$. In the data, however, both the use of a bot advisor and the elimination of suspicion regarding advisor bias ($p = 1$) increase compliance.

These patterns are confirmed by a regression analysis in which the dependent variable is an indicator for following the advisor's hint, and the independent variables include three treatment dummies: Bot, $p = 1$, and OB (observability). All regression results reported in this subsection are presented in Table A1 at Appendix. The estimated coefficients are $\beta_{Bot} = 0.061$ (SE = 0.035) and $\beta_{p=1} = 0.076$ (SE = 0.035), indicating positive effects of both treatments on compliance. In the same specification, observability has no statistically significant effect on compliance ($\beta_{OB} = -0.011$, SE = 0.043).

Similar positive effects of the bot advisor and the elimination of bias concerns are observed for both hint accuracy and guess accuracy. Importantly, observability has no negative effect on the rate of correct guesses. From a theoretical perspective, the presence of additional information sources (observability) may undermine information transmission and potentially reduce the guesser's welfare. However, we find no evidence of such a negative effect in the data.

To further refine the analysis, we split the sample by advisor type and examine compliance with hints separately (Panels (B) and (C) of Figure 2). For unbiased advisors, differences in compliance across conditions become smaller. In the corresponding regression, only the $p = 1$ dummy remains weakly significant ($\beta_{p=1} = 0.072$, SE = 0.040). In contrast, when restricting attention to biased advisors, the higher compliance with bot advice becomes more pronounced ($\beta_{Bot} = 0.152$, SE = 0.067).

Taken together, these results suggest that both the use of a bot advisor and the elimination of concerns about advisor bias contribute to higher compliance with advice. In contrast, the presence of additional information sources has no evident negative effect. The mechanisms behind these patterns are explored in the following subsections.

4.2 Guessers' behavior

4.2.1 Resistance to flipped or repeated hints

The key theoretical mechanism is the dilemma in information transmission: neither consistency nor revision of advice is fully trustworthy, as the former signals potential bias on the part of the advisor, while the latter signals limited knowledge. The model predicts that resistance to repeated or reversed hints emerges only when the guesser has access to alternative information sources; in the absence of such sources, compliance is fully restored.

Motivated by this theoretical insight, Table 2 reports regression results in which the dependent variable is an indicator for following the advisor's hint. Because the main results do not differ meaningfully between the $p = 0.5$ and $p = 0.8$ treatments, we pool the data

for these $p < 1$ conditions in the main text.[10] The key independent variables include (i) a dummy indicating whether the current hint differs from the previous period ("Flipped Hint") and (ii) the length of the current streak of identical hints ("Length of Repeated Hint"). When the guesser has access to alternative information sources, the alignment between the advisor's hint and the guesser's private signal becomes relevant. We therefore additionally include (i) a dummy indicating whether the current hint matches the private signal observed in the current period ("Hint Match") and (ii) another dummy representing the current hint matches the cumulative number of previously observed private signals ("Cumulative Hint Match").

Table 2: Regression Results on Guesser's Reactions

	Dependent variable:				
	Follow Hint				
	Human50or80	Bot50or80	Human100	Bot100	Observable
	(1)	(2)	(3)	(4)	(5)
Flipped Hint	−0.846	−0.746	−0.929	−0.542	−0.288
	(0.387)	(0.306)	(0.445)	(0.567)	(0.360)
Length of Repeated Hint	−0.245	−0.066	0.186	0.252	−0.220
	(0.092)	(0.035)	(0.103)	(0.140)	(0.095)
Hint Match					0.019
					(0.274)
Cumulative Hint Match					1.818
					(0.303)
Period	−0.125	−0.084	−0.091	−0.049	0.026
	(0.059)	(0.023)	(0.031)	(0.040)	(0.059)
Constant	2.958	3.301	3.135	3.190	0.467
	(0.471)	(0.382)	(0.617)	(0.767)	(0.422)
Observations	378	760	361	342	369
Log Likelihood	−190.854	−324.572	−132.772	−87.697	−190.094
Akaike Inf. Crit.	391.707	659.144	275.545	185.395	394.188
Bayesian Inf. Crit.	411.382	682.310	294.989	204.569	421.564

Note: Standard errors in parentheses. Follow Hint: 1 if the guesser follows the hint, otherwise 0. Flipped Hint: 1 if the hint changes from the previous period, 0 otherwise. Length of Repeated Hint: the number of consecutive occurrences of the same hint as in the current period. Hint Match: 1 if the hint provided by the advisor matches the color of ball the guesser observed in this period, 0 otherwise. Cumulative Hint Match: 1 if the hint provided by the advisor matches the cumulative records of ball colors the guesser observed so far, 0.5 if there is a tie, 0 otherwise. All models are estimated by using the generalized linear mixed model where participants' characteristics are controlled by the random effects.

The regression results show that guessers resist both reversed and repeated hints, regard-

[10] Readers can refer to Tables A2 and A3 in the Appendix for the fully disaggregated results.

less of whether additional information sources are available. First, in the Human50or80 conditions—where advisor bias is possible but the guesser lacks alternative information—participants exhibit significant resistance to both flipped hints and repeated hints. In the regressions, the corresponding coefficients are negative and statistically significantly, indicating that both abrupt revisions and persistent repetition reduce the likelihood of compliance. In other words, deviations from moderate variability in advice trigger skepticism. This behavioral pattern suggests that the communication dilemma is salient for participants: they penalize advisors for revising their stance (interpreting it as limited knowledge) while also discounting advice that appears excessively consistent (interpreting it as strategic bias).

Importantly, the theory predicts that such resistance should arise only when guessers have access to additional information sources. However, the data do not support this prediction. In the Human50or80-OB conditions, both Flipped Hint and Length of Repeated Hint continue to have negative effects on compliance, although the effect of Flipped Hint is smaller and no longer statistically significant.[11] Thus, the presence of additional information sources is neither necessary nor sufficient for the emergence of the communication dilemma.

However, this resistance is highly sensitive to the perceived possibility of advisor bias and to whether the advisor is human.

First, in the Human100 condition—where bias is ruled out—Flipped Hint continues to have a negative and statistically significant effect, whereas Length of Repeated Hint no longer reduces compliance. This pattern indicates that resistance to persistent advice is driven primarily by concerns about advisor bias. When the possibility of bias is eliminated, repeated advice no longer triggers skepticism, while abrupt revisions continue to be penalized, consistent with concerns about limited knowledge.

Second, resistance to both repeated and flipped hints diminishes and effectively disappears when the advisor is identified as a computer bot. In the Bot100 condition, participants exhibit no significant resistance to either cue and follow machine-generated advice more consistently. This finding suggests that the removal of perceived strategic intentions attenuates the communication dilemma. Furthermore, even when the bot's signals are imperfect, the absence of suspected ulterior motives reduces skepticism toward both consistency and revision as seen in Bot50or80 conditions.

Finally, to investigate whether guessers attempted to infer the advisor's type from the sequence of hints, we analyze the changes in their posterior beliefs regarding their paired advisor's type elicited post-experiment (see Table A4). Using the deviation of the posterior belief from the prior probability as the dependent variable, we find that a higher frequency of hint reversals—that is, a lower degree of consistency—increases the guessers' belief that the advisor is unbiased (β = 1.540, SE = 0.488). This finding is highly consistent with our

[11] In this condition, the patterns observed in other treatments are largely replaced by the alignment between the participant's own observations and the advisor's hint. Specifically, the congruence between the observed ball and the advisor's cue becomes the primary determinant of participant behavior (see Column (5)), exerting a substantially stronger effect than the pattern of hints.

preceding discussion.

4.2.2 Response to correct and incorrect hints

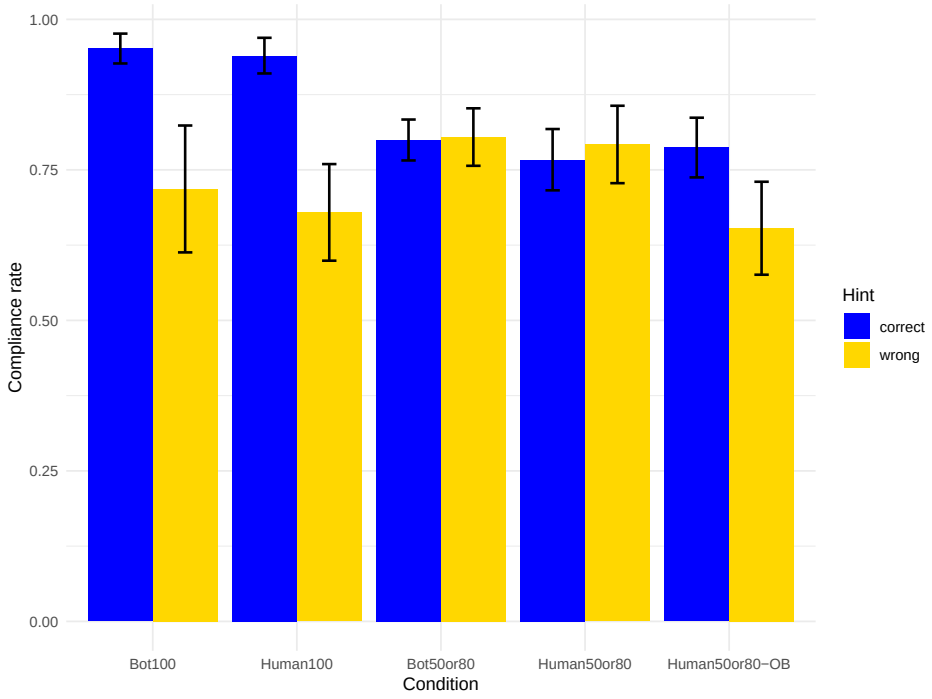

Figure 3: Proportion of compliance depending on whether hint is correct

Note: Error bars indicate 95% confidence intervals. Since each individual provides multiple observations, intervals are calculated based on individual-level clustering.

The experimental data further show to what extent guessers appropriately respond to hints. As shown in Figure 3, the compliance rate for "Correct Hints" is consistently higher than for "Wrong Hints" across all treatments, but the level of this gap highly depends on whether $p = 1$ or $p < 1$.

On the one hand, in the Bot100 and Human100 conditions—where the advisor is known to be unbiased—the gap between the rate of following correct hints and that of following incorrect hints is most pronounced. This pattern suggests that when concerns about strategic bias are eliminated, guessers are able to make effective use of the advisor's information while appropriately discounting hints that conflict with the probabilistic structure of the task.

On the other hand, introducing the possibility of bias ($p < 1$) fundamentally alters this pattern. In the Human50or80 treatments, the rate of following correct hints declines, whereas the rate of following incorrect hints remains relatively high compared to the $p = 1$ baseline. The resulting compression of the gap suggests that guessers have difficulty distinguishing between an honest error by an unbiased advisor and strategic manipulation by a biased one.

Interestingly, in the Human50or80-OB condition, the rate of following incorrect hints decreases again. This indicates that private information ($n = 1$) functions as a form of "sanity

check," enabling guessers to reject professional advice when it clearly conflicts with their own observations.

4.3 Advisors' behavior: Heuristic decision-making processes

In the previous subsection, we showed that guessers face a communication dilemma in information transmission. However, because this interaction is strategically complex, it remains unclear how advisors respond to the dilemma faced by guessers. This subsection turns to the behavior of advisors.

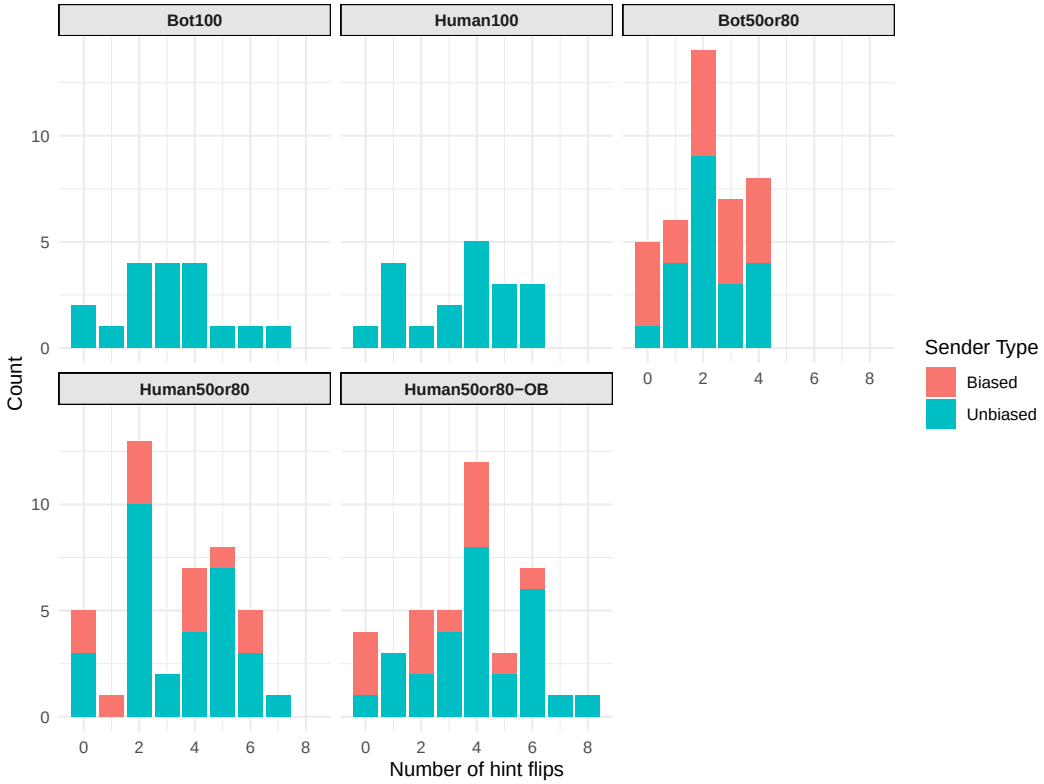

Figure 4: Histogram of the number of flips

Figure 4 presents histograms by experimental condition, showing the distribution of the number of hint reversals by advisors by the end of period 10.

In the absence of additional information sources, the equilibrium characterized in Proposition 1 implies that a human advisor reverses the hint only when the advisor is unbiased and the sign of Δ_t changes across periods. Biased advisors, by contrast, persistently recommend their preferred action and therefore never reverse their hints in equilibrium. Bots, however, make errors with a 10% probability. As a result, reversals may occur not only because of changes in Δ_t, but also due to mechanical errors. Therefore, the model predicts that flip frequencies in the Human conditions without alternative information sources should be weakly lower than—or at most equal to—those in the corresponding Bot conditions, even after accounting for potential human decision errors.

However, the behavior of human advisors departs from this prediction. Unbiased human advisors reverse their hints significantly more often than Unbiased Bot advisors ($M = 3.66$ vs. $M = 2.61$; Wilcoxon rank-sum test, $p = 0.005$). In contrast, no statistically significant difference is observed between biased human and biased bot advisors ($M = 2.88$ vs. $M = 2.11$; $p = 0.183$).

One possible interpretation is that unbiased human advisors flip hints more frequently in an attempt to signal impartiality by avoiding excessive consistency. Indeed, the previous subsection shows that guessers penalize persistent consistency even in the absence of alternative information sources. Anticipating such skepticism, unbiased advisors might therefore vary their recommendations strategically.

However, a direct comparison between unbiased and biased human advisors reveals no statistically significant difference in flip frequencies ($M = 3.66$ vs. $M = 2.88$; $p = 0.125$). This lack of statistical separation suggests that strategic differentiation is unlikely to be the primary mechanism underlying the observed pattern.

To explore this mechanism further, Table 3 presents a regression analysis where the advisor's choice in the current period is regressed on (i) the current-period signal (Number of Blue Balls) and (ii) the cumulative number of previously observed signals (Cumulative Blue minus Red). Column (1) reports results for human advisors, (2) presents results for human unbiased advisors, and (3) and (4) present the corresponding specification for bot advisors.

The results indicate that bots' decisions closely track the cumulative distribution of observed balls. In contrast, human advisors place substantially less weight on accumulated historical information and instead respond more strongly to the most recent observation.

Consistent with this interpretation, 89.5% of decisions by unbiased human advisors follow a simple rule: recommend Red if and only if the number of red signals observed in the current period exceeds the number of blue signals. The corresponding share for bot advisors is 63.1%.

This pattern suggests that the excess flipping behavior among human advisors reflects myopic updating that places disproportionate weight on the most recent observation, rather than deliberate attempts at strategic differentiation.

One might wonder whether this naive behavior by human advisors mechanically accounts for the excess noncompliance we observe. We argue that it does not. To see why, consider a guesser who correctly anticipates that the unbiased advisor follows a myopic rule, recommending R if and only if the majority of signals in the current period are red. Under this rule, a sequence of identical hints $h^t = (R, \cdots, R)$ arises either because (i) the advisor is red-biased, or (ii) the advisor is unbiased and the majority of current-period signals happened to be red in every period. Crucially, case (ii) is itself evidence that $\omega = R$: an unbiased advisor who repeatedly draws mostly red balls is more likely to be facing a red pot. A Bayesian guesser who correctly accounts for the advisor's myopic strategy should therefore still follow a consistent sequence of hints, even knowing that the advisor updates myopically rather than on cumulative evidence. The consistency penalty we document cannot be rationalized by the advisor's erratic behavior alone. It reflects instead a behavioral tendency on the guesser's

Table 3: Regression Results on Advisor's Behavior

	Dependent variable:			
	Send Blue Hint			
	Human	Human-Unbiased	Bot	Bot-Unbiased
	(1)	(2)	(3)	(4)
Blue Biased	2.221		2.505	
	(0.494)		(0.430)	
Red Biased	−3.813		−2.126	
	(0.542)		(0.385)	
Number of Blue Balls	2.696	4.034	0.028	0.076
	(0.182)	(0.311)	(0.099)	(0.118)
Cumulative Blue minus Red	0.050	0.091	0.160	0.230
	(0.019)	(0.026)	(0.016)	(0.019)
Constant	−4.041	−6.033	−0.040	−0.110
	(0.320)	(0.518)	(0.223)	(0.213)
Observations	1,210	960	1,160	780
Log Likelihood	−441.700	−284.067	−480.949	−324.432
Akaike Inf. Crit.	895.400	576.133	973.898	656.864
Bayesian Inf. Crit.	925.990	595.601	1,004.235	675.501

Note: Standard errors in parentheses. Send Blue Hint: 1 if the advisor sent a Blue hint, 0 otherwise. Blue/Red Biased: Indicators for the advisor's assigned bias type. Number of Blue Balls: Number of blue balls drawn in the current trial. Cumulative Blue minus Red: Cumulative difference between blue and red draws so far. All models are estimated by using the generalized linear mixed model where participants' characteristics are controlled by the random effects.

side.

4.4 Summary of experimental findings

Taken together, our experimental results document a robust behavioral departure from the theoretical benchmark of full compliance. First, the consistency-revision dilemma is highly salient: guessers exhibit significant resistance to both abrupt revisions and prolonged repetitions of advice, reducing overall compliance to approximately 75% even without alternative information sources. Second, this resistance is fundamentally driven by distrust in the advisor's intentions and competence; when advice is generated by an algorithmic bot or when bias is explicitly ruled out, the dilemma is mitigated and compliance substantially recovers. Third, while theory cautions that private information might undermine information transmission, the data reveal no such negative effect on guessers' welfare. Instead, private signals function effectively as a "sanity check," helping guessers reject objectively incorrect hints. Finally, we find that human advisors tend to update their recommendations myopically rather than based on cumulative evidence. However, this heuristic behavior alone cannot fully rationalize the guessers' noncompliance, confirming that the communication breakdown inherently stems from the guessers' profound skepticism toward human expert advice.

5 Conclusion

The rigorous analyses in this paper demonstrate that maintaining public compliance in expert communication requires considerations well beyond mere point-in-time "accuracy." We theoretically identify and experimentally validate a fundamental dilemma: citizens penalize experts both for maintaining excessive consistency, which they interpret as a signal of strategic bias, and for frequently revising their advice, which they view as a signal of incompetence or limited knowledge. Our results show that this dilemma reduces compliance to roughly 75% even if citizens have no alternative information sources.

Contrary to theoretical concerns, our empirical findings suggest that citizens' access to independent, private information does not inherently erode the authority of experts. Rather, private signals function as a screening device that filters out objectively poor advice while reinforcing robust recommendations. Policymakers, therefore, should not view alternative information sources as obstacles to overcome; instead, they must ensure that official guidance remains logically consistent with—and can be effectively explained in the context of—publicly observable evidence.

Furthermore, our results point to a potential pathway for restoring trust through institutional design: algorithmic advice. Compliance is significantly higher, and the communication dilemma less pronounced, when advice is delivered by a computer bot. Because bots are perceived to lack hidden agendas, they are insulated from suspicions of subjective bias and do not face the same degree of backlash when making errors. This highlights the potential for

expert committees to formalize and transparently disclose their algorithmic decision-making processes.

Ultimately, navigating this dilemma demands advice that strikes a delicate balance: it must be stable enough to signal competence, yet responsive enough to signal honesty. Experts must remain acutely aware of the longitudinal patterns their advice creates. Mitigating advisors' cognitive biases—particularly the "myopic updating" that overreacts to recent events is also important. These insights offer practical and institutional guidance on delivering scientific advice to the public in today's highly uncertain and skeptical policy arenas.

A Omitted Proofs

A.1 Proof of Proposition 1

Step 1. As the first step, we prove that this constitutes a perfect Bayesian equilibrium.

First, we examine the guesser's incentive. Let $\tilde{q}(h^t)$ be the guesser's subjective probability of $\omega = B$ and $\tilde{p}_k(h^t)$ be the guesser's subjective probability of the advisor being k-biased. Without notational abuse, we may use $\tilde{p}_k^{t-1}$ instead of $\tilde{p}_k(h^t)$. Given the strategies of biased and unbiased advisors, from the Bayes rule,

$$\tilde{q}(h^{t-1}, B) = \frac{(1 - \tilde{p}_B^{t-1} - \tilde{p}_R^{t-1})A_1 + (\tilde{p}_B^{t-1}\mathbf{1}\{s_B^* = B\} + \tilde{p}_R(h^{t-1})\mathbf{1}\{s_R^* = B\})\frac{1}{2}}{(1 - \tilde{p}_B^{t-1} - \tilde{p}_R^{t-1})(A_1 + A_2) + (\tilde{p}_B^{t-1}\mathbf{1}\{s_B^* = B\} + \tilde{p}_R(h^{t-1})\mathbf{1}\{s_R^* = B\})}, \tag{1}$$

where

$$A_1 = \Pr(h_t = B, \omega = B \mid \text{advisor} = \text{unbiased}, h^{t-1});$$

$$A_2 = \Pr(h_t = B, \omega = R \mid \text{advisor} = \text{unbiased}, h^{t-1}).$$

From the equilibrium strategy of the unbiased advisor, any h^{t-1} can be chosen by the unbiased advisor; that is, $1 - \tilde{p}_B^{t-1} - \tilde{p}_R^{t-1} > 0$. Furthermore, the equilibrium strategy of the unbiased advisor also implies that $A_1 > A_2$. Therefore, (1)$> \frac{1}{2}$, meaning that the guesser has an incentive to choose $a = B$ if $h_t = B$.

Similarly, it is also shown that the guesser has an incentive to choose $a = R$ if $h_t = R$.

Given the equilibrium strategy of the guesser, it is straightforward that the advisor follows the specified the strategy. Therefore, (i)-(iii) constitute a perfect Bayesian equilibrium.

Step 2. Next, we prove that there is no other equilibrium. Prove by contradiction. Suppose that the guesser ignores a hint by the advisor in period t for some h^t.

Assumption (i) implies that it is never the case that the guesser follows neither $h_t = B$ nor $h_t = R$. Without loss of generality, suppose that the guesser chooses $a = R$ for any h_t.

This implies that for the guesser,

$$\tilde{q}(h^{t-1}, R) \leq \frac{1}{2} \text{ and } \tilde{q}(h^{t-1}, B) \leq \frac{1}{2}$$

hold simultaneously. However, this is never the case from Assumption 1 (i), which is a contradiction.

Therefore, the guesser always follows the advisor's hint. Given this, the advisor takes the strategy (ii) and (iii). This completes the proof. □

A.2 Proof of Lemma 1

Let the number of $s_{it}^G = B$ observed by the guesser in period t be θ_t and let the sequence of θ_t be $\theta^t := (\theta_1, ..., \theta_t)$.

$h_t \neq h_{t-1}$ implies that the advisor is unbiased; that is, $\tilde{p}_B(h^t, \theta^t) = \tilde{p}_R(h^t, \theta^t) = 0$. Without loss of generality, we consider the case where $h_t = B$ and $h_{t-1} = R$.

Given the unbiased advisor's equilibrium strategy, hints imply that $\Delta_{t-1} \leq 0$ and $\Delta_t \geq 0$. This further implies that $0 \leq \Delta_t \leq m$.

Suppose that the number of blue balls minuns red balls until period t, privately observed by the guesser is $k < 0$. If $k \geq 0$, the guesser obviously follows the advisor's hint. Thus, we assume so.

Then, the above property implies that

$$\tilde{q}(h^t, \theta^t) \leq \frac{\frac{1}{2}\left(\frac{3}{5}\right)^m \left(\frac{2}{5}\right)^{-k}}{\frac{1}{2}\left(\frac{3}{5}\right)^m \left(\frac{2}{5}\right)^{-k} + \frac{1}{2}\left(\frac{2}{5}\right)^m \left(\frac{3}{5}\right)^{-k}}. \tag{2}$$

If the right-hand side of (2) is less than a half for some k, it implies that $\tilde{q}(h^t, \theta^t) < 1/2$ for some θ^t; that is, the guesser has an incentive not to follow the advisor's hint B. Therefore, it suffices to derive the condition under which the right-hand side of (2) is less than a half for some k.

This condition is given by

$$\frac{\frac{1}{2}\left(\frac{3}{5}\right)^m \left(\frac{2}{5}\right)^{nT}}{\frac{1}{2}\left(\frac{3}{5}\right)^m \left(\frac{2}{5}\right)^{nT} + \frac{1}{2}\left(\frac{2}{5}\right)^m \left(\frac{3}{5}\right)^{nT}} < \frac{1}{2} \Leftrightarrow nT > m.$$

In summary, the guesser has an incentive not to follow the advisor's hint B for some θ^t and h^t when $nT > m$. This completes the proof. □

A.3 Proof of Lemma 2

Without loss of generality, suppose that $h_\tau = B$ for any $\tau \in \{1, ..., t\}$. Consider an extreme case where $\theta_\tau = 0$ for any $\tau \in \{1, ..., t\}$; that is, every private information for the guesser indicates $\omega = R$, whereas every hint provided by the advisor indicates $\omega = B$, because the guesser is most likely to deviate from the hint in this case.

Then,

$$\tilde{q}(h^t, \theta^t) = \frac{\frac{1}{2}\left[\frac{1-p}{2} + pL_t(B)\right]\left(\frac{2}{5}\right)^{nt}}{\frac{1}{2}\left[\frac{1-p}{2} + pL_t(B)\right]\left(\frac{2}{5}\right)^{nt} + \frac{1}{2}\left[\frac{1-p}{2} + pL_t(R)\right]\left(\frac{3}{5}\right)^{nt}},$$

where $L_t(\omega) := \Pr(\Delta_\tau \geq 0 \; \forall \tau \leq t \mid \omega)$.

This is less than a half if and only if

$$\begin{aligned}
&\left[\frac{1-p}{2} + pL_t(B)\right]\left(\frac{2}{5}\right)^{nt} < \left[\frac{1-p}{2} + pL_t(R)\right]\left(\frac{3}{5}\right)^{nt} \\
\Leftrightarrow &\frac{1-p}{2} + pL_t(B) < \left[\frac{1-p}{2} + pL_t(R)\right]\left(\frac{3}{2}\right)^{nt} \\
\Leftrightarrow &\frac{1-p}{2}\left[\left(\frac{3}{2}\right)^{nt} - 1\right] > p\left[L_t(B) - L_t(R)\left(\frac{3}{2}\right)^{nt}\right]. \\
\Leftrightarrow &\frac{p}{1-p} < \frac{\left(\frac{3}{2}\right)^{nt} - 1}{2L_t(B) - 2L_t(R)\left(\frac{3}{2}\right)^{nt}} \text{ or } L_t(B) \leq L_t(R)\left(\frac{3}{2}\right)^{nt}.
\end{aligned}$$

This completes the proof. □

B Additional Analysis

Table A1: Regression Results on Performances

	Dependent variable:				
	Follow Hint	Correct Hint Pooled	Correct Guess	Follow Hint Unbiased Advisor	Follow Hint Biased Advisor
	(1)	(2)	(3)	(4)	(5)
Bot Advisor Dummy	0.061	0.089	0.098	0.030	0.152
	(0.035)	(0.052)	(0.052)	(0.039)	(0.067)
$p = 1$ Dummy	0.076	0.058	0.127	0.072	
	(0.035)	(0.052)	(0.052)	(0.040)	
Observable Dummy	−0.011	0.039	0.099	−0.020	0.033
	(0.043)	(0.064)	(0.063)	(0.051)	(0.072)
Constant	0.750	0.602	0.530	0.778	0.667
	(0.032)	(0.049)	(0.048)	(0.039)	(0.052)
Observations	160	160	160	116	44
R^2	0.068	0.023	0.045	0.063	0.131
Adjusted R^2	0.050	0.004	0.026	0.038	0.089
Residual Std. Error	0.179 (df = 156)	0.268 (df = 156)	0.264 (df = 156)	0.175 (df = 112)	0.181 (df = 41)
F Statistic	3.788 (df = 3; 156)	1.219 (df = 3; 156)	2.436 (df = 3; 156)	2.495 (df = 3; 112)	3.096 (df = 2; 41)

Note: Standard errors in parentheses. Follow Hint: 1 if the guesser followed the advisor's hint, and 0 otherwise. Correct Hint: 1 if the hint provided was accurate, and 0 otherwise. Correct Guess: 1 if the guesser's choice matched the true state, and 0 otherwise. Bot Advisor Dummy: 1 if the paired advisor is a bot, and 0 otherwise. $p = 1$ Dummy: 1 if $p = 1$, and 0 otherwise. Observable Dummy: 1 if the guesser observes a private signal ($n = 1$), and 0 otherwise. The 'Pooled' models include all advisor types, while the last two columns separate the results by advisors' types

Table A2: Regression Results on Guesser's Reactions to Human Advisors

	Dependent variable:				
	Follow Hint				
	Human50	Human80	Human100	Human50-OB	Human80-OB
	(1)	(2)	(3)	(4)	(5)
Flipped Hint	−0.743	−1.017	−0.929	0.385	−0.586
	(0.528)	(0.596)	(0.445)	(0.554)	(0.512)
Length of Repeated Hint	−0.266	−0.236	0.186	−0.157	−0.131
	(0.108)	(0.168)	(0.103)	(0.136)	(0.160)
Hint Match				−0.208	0.230
				(0.415)	(0.382)
Cumulative Hint Match				2.100	1.614
				(0.472)	(0.422)
Period	−0.026	−0.241	−0.091	−0.089	0.103
	(0.080)	(0.091)	(0.031)	(0.092)	(0.079)
Constant	2.221	3.934	3.135	0.731	0.109
	(0.564)	(0.810)	(0.617)	(0.640)	(0.590)
Observations	189	189	361	189	180
Log Likelihood	−100.666	−87.671	−132.772	−95.742	−89.935
Akaike Inf. Crit.	211.333	185.341	275.545	205.485	193.869
Bayesian Inf. Crit.	227.541	201.550	294.989	228.177	216.220

Note: Standard errors in parentheses. Follow Hint: 1 if the guesser follows the hint, otherwise 0. Flipped Hint: 1 if the hint changes from the previous period, 0 otherwise. Length of Repeated Hint: the number of consecutive occurrences of the same hint as in the current period. Hint Match: 1 if the hint provided by the advisor matches the color of ball the guesser observed in this period, 0 otherwise. Cumulative Hint Match: 1 if the hint provided by the advisor matches the cumulative records of ball colors the guesser observed so far, 0.5 if there is a tie, 0 otherwise. All models are estimated by using the generalized linear mixed model where participants' characteristics are controlled by the random effects.

Table A3: Regression Results on Guesser's Reactions to Algorithmic Advisors

	Dependent variable:		
	Follow Hint		
	Bot50	Bot80	Bot100
	(1)	(2)	(3)
Flipped Hint	−0.550	−0.550	−0.542
	(0.448)	(0.448)	(0.567)
Length of Repeated Hint	−0.021	−0.021	0.252
	(0.045)	(0.045)	(0.140)
Period	−0.082	−0.082	−0.049
	(0.033)	(0.033)	(0.040)
Constant	3.024	3.024	3.190
	(0.467)	(0.467)	(0.767)
Observations	380	380	342
Log Likelihood	−162.150	−162.150	−87.697
Akaike Inf. Crit.	334.301	334.301	185.395
Bayesian Inf. Crit.	354.001	354.001	204.569

Note: Standard errors in parentheses. Follow Hint: 1 if the guesser follows the hint, otherwise 0. Flipped Hint: 1 if the hint changes from the previous period, 0 otherwise. Length of Repeated Hint: the number of consecutive occurrences of the same hint as in the current period. All models are estimated by using the generalized linear mixed model where participants' characteristics are controlled by the random effects.

Table A4: Regression Results on Belief Change

	Dependent variable:
	Posterior minus Prior Probability
Number of Hint Flipped	1.540
	(0.488)
Biased Advisor Dummy	−3.476
	(2.174)
Bot Advisor Dummy	−5.623
	(2.558)
Observable Dummy	0.260
	(2.468)
Constant	−4.661
	(2.482)
Observations	119
R^2	0.157
Adjusted R^2	0.128
Residual Std. Error	11.076 (df = 114)
F Statistic	5.317 (df = 4; 114)

Note: Standard errors in parentheses. Number of Hint Flipped: The total number of hints flipped within the hint sequence. Biased Advisor Dummy: 1 if the paired advisor is a biased type, and 0 otherwise. Bot Advisor Dummy: 1 if the paired advisor is a bot, and 0 otherwise. Observable Dummy: 1 if the guesser observes a private signal ($n = 1$), and 0 otherwise.

References

Adjodah, Dhaval, Karthik Dinakar, Matteo Chinazzi, Samuel P Fraiberger, Alex Pentland, Samantha Bates, Kyle Staller, Alessandro Vespignani, and Deepak L Bhatt. 2021. "Association between COVID-19 outcomes and mask mandates, adherence, and attitudes." *PloS one*, 16(6): e0252315.

Algan, Yann, Daniel Cohen, Eva Davoine, Martial Foucault, and Stefanie Stantcheva. 2021. "Trust in scientists in times of pandemic: Panel evidence from 12 countries." *Proceedings of the National Academy of Sciences*, 118(40): e2108576118.

Bargain, Olivier, and Ulugbek Aminjonov. 2020. "Trust and compliance to public health policies in times of COVID-19." *Journal of Public Economics*, 192: 104316.

Benabou, Roland, and Guy Laroque. 1992. "Using privileged information to manipulate markets: Insiders, gurus, and credibility." *The Quarterly Journal of Economics*, 107(3): 921–958.

Bes-Rastrollo, Maira, Matthias B Schulze, Miguel Ruiz-Canela, and Miguel A Martinez-Gonzalez. 2013. "Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: A systematic review of systematic reviews." *PLoS Medicine*, 10(12): e1001578.

Blume, Andreas, Douglas V DeJong, Yong-Gwan Kim, and Geoffrey B Sprinkle. 1998. "Experimental evidence on the evolution of meaning of messages in sender-receiver games." *The American Economic Review*, 88(5): 1323–1340.

Blume, Andreas, Ernest K Lai, and Wooyoung Lim. 2020. "Strategic information transmission: A survey of experiments and theoretical foundations." *Handbook of Experimental Game Theory*, 311–347.

Cai, Hongbin, and Joseph Tao-Yi Wang. 2006. "Overcommunication in strategic information transmission games." *Games and Economic Behavior*, 56(1): 7–36.

Çelen, Boğaçhan, Shachar Kariv, and Andrew Schotter. 2010. "An experimental test of advice and social learning." *Management Science*, 56(10): 1687–1701.

Charness, Gary, and Brit Grosskopf. 2004. "What makes cheap talk effective? Experimental evidence." *Economics Letters*, 83(3): 383–389.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "oTree: An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Chen, Zhaowei, Jijia Hu, Zongwei Zhang, Shan Jiang, Shoumeng Han, Dandan Yan, Ruhong Zhuang, Ben Hu, and Zhan Zhang. 2020. "Efficacy of hydroxychloroquine in patients with COVID-19: Results of a randomized clinical trial." *medRxiv*.

Choi, Syngjoo, Chanjoo Lee, and Wooyoung Lim. 2025. "The Anatomy of Honesty: Lying Aversion vs. Deception Aversion." *Working Paper*.

Chung, Wonsuk, and Rick Harbaugh. 2019. "Biased recommendations from biased and unbiased experts." *Journal of Economics & Management Strategy*, 28(3): 520–540.

Commerford, Benjamin P, Sean A Dennis, Jennifer R Joe, and Jenny W Ulla. 2022. "Man versus machine: Complex estimates and auditor reliance on artificial intelligence." *Journal of Accounting Research*, 60(1): 171–201.

Crawford, Vincent P, and Joel Sobel. 1982. "Strategic information transmission." *Econometrica*, 1431–1451.

de Barreda, Ines Moreno. 2024. "Cheap talk with two-sided private information." *Games and Economic Behavior*, 148: 97–118.

Dickhaut, John W, Kevin A McCabe, and Arijit Mukherji. 1995. "An experimental study of strategic information transmission." *Economic Theory*, 6: 389–403.

Ettinger, David, and Philippe Jehiel. 2021. "An experiment on deception, reputation and trust." *Experimental Economics*, 24(3): 821–853.

Gautret, Philippe, Jean-Christophe Lagier, Philippe Parola, Line Meddeb, Morgane Mailhe, Barbara Doudier, Johan Courjon, Valérie Giordanengo, Vera Esteves Vieira, Hervé Tissot Dupont, et al. 2020. "Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial." *International Journal of Antimicrobial gents*, 56(1): 105949.

Harrison, Glenn W, Jimmy Martínez-Correa, and J Todd Swarthout. 2013. "Inducing risk neutral preferences with binary lotteries: A reconsideration." *Journal of Economic Behavior & Organization*, 94: 145–159.

Hossain, Tanjim, and Ryo Okui. 2013. "The binarized scoring rule." *Review of Economic Studies*, 80(3): 984–1001.

Huang, Jinyuqi, and Wooyoung Lim. 2025. "Role Uncertainty and Punishment Severity in Repeated Cheap Talk: Theory and Experiment." *Available at SSRN 5314097*.

Intemann, Kristen. 2023. "Science communication and public trust in science." *Interdisciplinary Science Reviews*, 48(2): 350–365.

Ishida, Junichiro, and Takashi Shimizu. 2016. "Cheap talk with an informed receiver." *Economic Theory Bulletin*, 4(1): 61–72.

Ishowo-Oloko, Fatimah, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. "Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation." *Nature Machine Intelligence*, 1(11): 517–521.

Jin, Ginger Zhe, Michael Luca, and Daniel Martin. 2021. "Is no news (perceived as) bad news? An experimental investigation of information disclosure." *American Economic Journal: Microeconomics*, 13(2): 141–173.

Kartal, Melis, and James Tremewan. 2018. "An offer you can refuse: the effect of transparency with endogenous conflict of interest." *Journal of Public Economics*, 161: 44–55.

Lai, Ernest K. 2014. "Expert advice for amateurs." *Journal of Economic Behavior & Organization*, 103: 1–16.

Li, Xiaolin, Özalp Özer, and Upender Subramanian. 2022. "Are we strategically naïve or guided by trust and trustworthiness in cheap-talk communication?" *Management Science*, 68(1): 376–398.

Loewenstein, George, Daylian M Cain, and Sunita Sah. 2011. "The limits of transparency: Pitfalls and potential of disclosing conflicts of interest." *American Economic Review*, 101(3): 423–428.

Meloso, Debrah, Salvatore Nunnari, and Marco Ottaviani. 2023. "Looking into crystal balls: A laboratory experiment on reputational cheap talk." *Management Science*, 69(9): 5112–5127.

Morris, Stephen. 2001. "Political correctness." *Journal of Political Economy*, 109(2): 231–265.

Oreskes, Naomi, and Erik M Conway. 2010. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.

Ottaviani, Marco, and Peter Norman Sørensen. 2006*a*. "Professional advice." *Journal of Economic Theory*, 126(1): 120–142.

Ottaviani, Marco, and Peter Norman Sørensen. 2006*b*. "Reputational cheap talk." *The Rand Journal of Economics*, 37(1): 155–175.

Schotter, Andrew. 2003. "Decision making with naive advice." *American Economic Review*, 93(2): 196–201.

Sobel, Joel. 1985. "A theory of credibility." *The Review of Economic Studies*, 52(4): 557–573.

Tang, Wei, Zhujun Cao, Mingfeng Han, Zhengyan Wang, Junwen Chen, Wenjin Sun, Yaojie Wu, Wei Xiao, Shengyong Liu, Erzhen Chen, et al. 2020. "Hydroxychloroquine in patients with mainly mild to moderate coronavirus disease 2019: Open label, randomised controlled trial." *BMJ*, 369.

Wang, Amy T, Christopher P McCoy, Mohammad Hassan Murad, and Victor M Montori. 2010. "Association between industry affiliation and position on cardiovascular risk with rosiglitazone: Cross sectional systematic review." *BMJ*, 340.

Wilson, Alistair J, and Emanuel Vespa. 2020. "Information transmission under the shadow of the future: An experiment." *American Economic Journal: Microeconomics*, 12(4): 75–98.

Online Appendix (Not for Publication)

Contents

C English Translation of Instructions

The following are instructions and decision screens in Human50-Observable(OB) condition ($n = 1$). In Human80-OB conditions, only the explanation about the probability of being biased and unbiased advisors (Role A) is different. In conditions of $n = 0$, the explanation of Role B observing a ball from the pot is removed. In Human100 condition, the explanation about the biased type of Role A is removed. In Bot advisor conditions, the explanations are a bit different and thus are shown after the instructions decision screens of Human50-OB condition.

The instructions, confirmation quiz, and experimental tasks were all conducted through oTree. Participants in the experiment were able to progress through the instructions and the confirmation quiz at their own pace.

C.1 Instruction Page 1

First, pairs of participants are formed randomly, with one assigned Role A and the other Role B. These roles do not change during the task. The two participants will work on a task where they need to guess the color of a pot.

There are two types of pots, as shown in the diagram below.

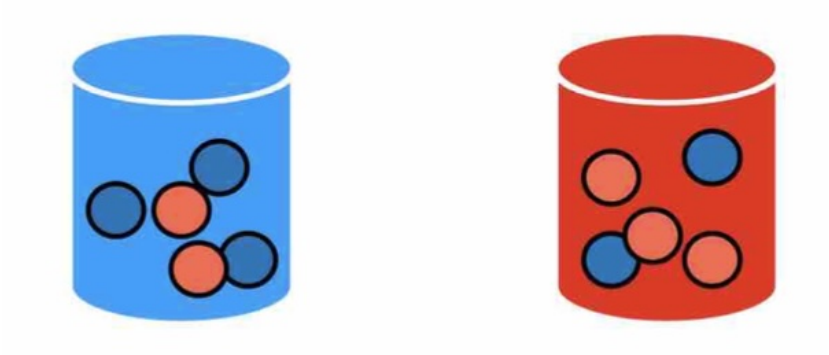

Figure 5: Two types of pots

The blue pot contains three blue balls and two red balls (the pot on the left in the diagram). Therefore, the blue pot has a higher number of blue balls, with a 60% probability of drawing a blue ball and a 40% probability of drawing a red ball.

The red pot contains three red balls and two blue balls (the pot on the right in the diagram). Thus, the red pot has a higher number of red balls, with a 60% probability of drawing a red ball and a 40% probability of drawing a blue ball.

The color of the pot is determined with a 50% probability of being blue and a 50% probability of being red. The color of the pot is decided at the start and does not change during the task. However, you will not know the color of the pot. Your task is to look at the color of the ball drawn from the pot and guess the color of the pot.

The person in Role A will draw one ball from the pot, check its color, and then return it to the pot. This process will be repeated three times. Since the ball is returned to the pot each time, the probabilities of drawing red or blue balls remain the same for all three draws. After observing the colors of the three balls, Role A will provide a hint to Role B regarding the color of the pot. The hint will be either blue or red.

The person in Role B will first check the hint from Role A. Then, they will draw one ball from the pot, check its color, and return it. After that, they will guess the color of the pot, which can be either blue or red.

In summary, the sequence of actions is as follows:

Role A observes the color of the ball three times. Role A communicates a hint to Role B. Role B observes the color of the ball once. Role B guesses the color of the pot.

This entire sequence constitutes one transaction, which can be illustrated as shown in the diagram below.

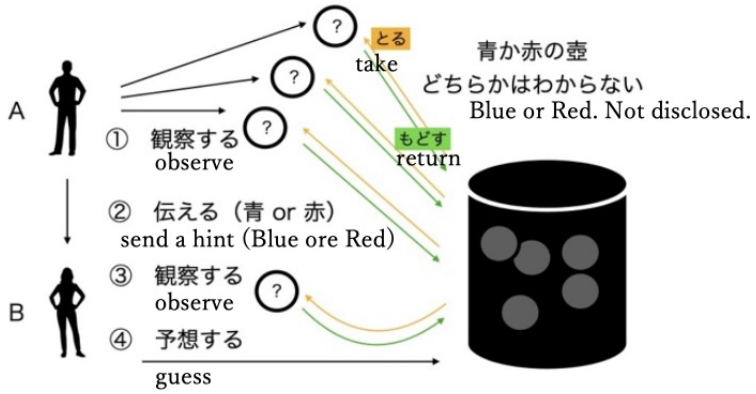

Figure 6: Entire sequence of a transaction

After each transaction, the choices (the hint and the guess) made by both participants will be disclosed to each other. However, the color of the pot will not be revealed, so it will remain unknown whether the hint or the guess was correct.

This transaction will be repeated 10 times. The color of the pot does not change during this process. The color of the pot will remain the same for rounds 1 to 10.

On the next page, we will explain the additional rewards you can earn.

C.2 Instruction Page 2

As previously described, the person in Role A will observe the color of the balls three times and then send a hint to the person in Role B regarding the color of the pot.

There are three types of Role A (Blue, Red, White), and the rules for earning rewards differ by type.

For Type Blue: In this case, if Role B predicts Blue, Role A earns 10 points; otherwise, they earn 0 points. Since there are 10 decisions, the maximum points Role A can earn is 100.

For Type Red: In this case, if Role B predicts Red, Role A earns 10 points; otherwise, they earn 0 points. Again, with 10 decisions, the maximum points Role A can earn is 100.

For Type White: In this case, if Role B predicts the correct pot, Role A earns 10 points; otherwise, they earn 0 points. With 10 decisions, the maximum points Role A can earn is 100.

The type of Role A will be notified at the time of role assignment. The probability of being Type Blue is 25%, Type Red is 25%, and Type White is 50%. The type does not change during the process. Additionally, Role B will not be able to know Role A's type.

Role B will confirm the hint from Role A and then make a prediction about the color of the pot.

The reward rules for Role B are as follows:

If Role B predicts the correct pot, they earn 10 points; if they choose the wrong pot, they earn 0 points. Since there are 10 decisions, the maximum points Role B can earn is 100.

After the 10 exchanges, you will draw a lottery to potentially earn a bonus of 500 yen. The score you earn will determine your probability of winning the 500 yen bonus. For example, if your score is 50 points, your probability of winning the 500 yen is 50%.

The quiz will begin on the next page.

C.3 Confirmation Quiz

Each page displayed one quiz question, and participants could check their results immediately after answering. Feedback regarding the results was provided. On the feedback screen, in addition to indicating whether the choice was correct or incorrect, a brief explanation was also displayed. Since the content overlaps with the instructions, the feedback screen is omitted here.

We will now conduct a pre-quiz consisting of four questions to assess your understanding of the rules. You will receive an additional 50 yen for each correct answer, so please take your time to answer carefully. If needed, you can refer back to the explanation of the experiment at the bottom of the page.

Quiz 1: You two will be divided into Role A and Role B, and you will repeatedly engage in the task of guessing the color of the pot. Please choose one incorrect option from the following statements.

- Option 1: You will repeat the task of guessing the pot's color 10 times. (*Correct*)
- Option 2: The roles will not change during the task. (*Correct*)
- Option 3: The color of the pot changes randomly to blue or red each time. (*Wrong*)

Quiz 2: Role A will take one ball from the pot, check its color, and return it to the pot. This process will be repeated three times. Assuming the pot is blue, choose the correct statement regarding the probabilities of the ball colors.

- Option 1: The probability of the ball being blue is 60%. This remains constant regardless of the number of draws. (*Correct*)
- Option 2: The probability of the ball being blue is 80%. This remains constant regardless of the number of draws. (*Wrong*)
- Option 3: The probability of the ball being blue changes with each draw. The probability for the first draw is 80%, the second is 60%, and the third is 40%. (*Wrong*)

Quiz 3: The scoring rules differ between Role A and Role B. Choose one incorrect option regarding the explanations of each scoring rule.

- Option 1: There are multiple types of Role A, and the scoring rules differ for each type. (*Correct*)
- Option 2: Role B earns 10 points for each correct prediction of the pot. (*Correct*)
- Option 3: Role B cannot know the type of Role A (blue, red, white). (*Correct*)
- Option 4: The probability of Role A being of type blue, red, or white is 33.3% for each. (*Wrong*)

Quiz 4: Select the correct option regarding the scores you will earn.

- Option 1: If your score after 10 transactions is 70 points, you will earn a bonus of 700 yen. (*Wrong*)
- Option 2: If your score after 10 transactions is 50 points, the probability of earning a bonus of 500 yen is 50
- Option 3: If your score after 10 transactions is 100 points, you will earn 1000 yen as a reward for answering all questions correctly. (*Wrong*)

C.4 Experimental Implementation Screen

The following screens will repeat every period:Role Notification, Hint Provision (Advisor's Decision Screen), Guess the color of pot (Guesser's Decision Screen), Display of the choice result (Feedback Screen).

During the task, the participants can refer back to the explanation of the experiment at the bottom of each page.

C.5 Role Notification

Either of the following two are displayed:

You are Role A. Your type is Red. Therefore, you earn 10 points each time B predicts Red. (*This part will be changed according to the actual type of Role A. Hereafter, the explanation will be done for the case of Type Red Advisor*)

Your task is to observe the colors of the balls three times and send a hint about the color of the pot to Role B. However, your screen will only display the counts of blue and red balls you observed.

or

You are Role B. You earn 10 points each time you correctly predict the pot.

Your task is to confirm the hint from A and the colors of the balls once, then make a prediction about the color of the pot. The hint will only show information indicating either Blue or Red.

C.6 Hint Provision

Your type is Red. Therefore, you earn 10 points every time B predicts Red.

You have performed the action of ”taking one ball out of the pot, checking the color of that ball, and putting it back into the pot” three times.

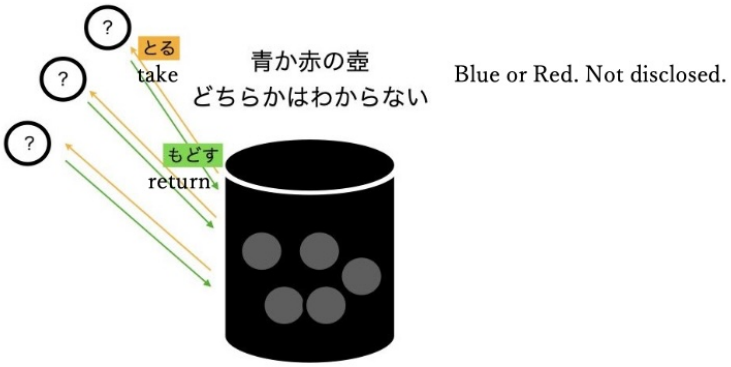

Figure 7: Your observe a ball three times

As a result, you observed:
1 blue ball and 2 red balls.
Please send a hint to B.
Send Hint: Blue or Red
Next

C.7 Guess the Color of the Pot

You received a hint from A indicating Red.

You took out one ball from the pot. The color of the ball was red.

Please guess the color of the pot.

Send Answer

Blue or Red

Next

C.8 Display of Choice Result

Role A sent the hint ”Red.” Role B predicted ”Red.”

Proceeding to the next round.

Next

C.9 Instruction Page 2 of Bot conditions

Explanations about the Role A's type is replaced by the following.

Role A will be performed by AI (computer).

There are three types of AI (Blue, Red, and White). The rule for sending a hint differs depending on the type.

Type Blue: This type of AI basically sends a Blue hint, regardless of the colors of the balls it has observed. However, there is a 10% chance that it will send the opposite color from what it originally intended (if it intended Blue, it will send Red, and vice versa).

Type Red: This type of AI basically sends a Red hint, regardless of the colors of the balls it has observed. However, there is a 10% chance that it will send the opposite color from what it originally intended (if it intended Blue, it will send Red, and vice versa).

Type White: This type of AI will consider all the colors of the balls it has observed so far and inform you of the color of the pot that is more probable at that point. If the probabilities are the same, it will randomly choose either blue or red. However, there is a 10% chance that it will send the opposite color from what it originally intended (if it intended Blue, it will send Red, and vice versa)