Rational vigilance of intentions and incentives guides learning from advice

[Anonymized for Submission]

Author Note

We have no known conflicts of interest to disclose.

All pre-registrations, materials, analysis scripts, and data are available at https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3.

Abstract

People *vigilantly* evaluate advice: they seek to learn from competent and well-intentioned informants, while ignoring incompetent or manipulative ones. Yet the mechanisms that support vigilance of informants' motivations remain poorly-understood. For instance, how do people combine informants' incentives and intentions when evaluating testimony—through rational inference, or simple heuristics? To answer such questions, here we develop a rational model of vigilance and test its predictions in two pre-registered online experiments with participants in the U.S (Ns = 626, 555). Across financial, real estate, and medical decision scenarios, participants' inferences closely tracked our model's predictions: participants discounted advice when selfish speakers benefited from manipulating them, and this discounting was attenuated by perceived benevolence. When pay-offs were shared, incentives increased trust, as our rational model—but not simple heuristics—predict. Our rational analysis thus establishes a formal framework for vigilance of motivations that also sheds light on phenomena from persuasion to polarization.

Keywords: vigilance, social learning, Bayesian inference, disagreement, belief, judgment

Research Transparency Statement

General Disclosures

Conflicts of interest: We are not aware of any conflicts of interest. Funding: This research was supported by [anonymized author grant]. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research received approval from [author institution].

Study One

Preregistration: The hypotheses, methods, and analyses were preregistered (https://aspredicted.org/zp8r-3tbr.pdf) prior to data collection. All deviations from the preregistration in the analyses are noted as exploratory. Materials: All study materials are publicly available (https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3). Data: All primary data are publicly available (https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3). Analysis scripts: All analysis scripts are publicly available (https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3).

Study Two

The hypotheses, methods, and analyses were preregistered (https://aspredicted.org/rdyf-c4vx.pdf) prior to data collection. All deviations from the preregistration in the analyses are noted as exploratory. Materials: All study materials are publicly available (https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3). Data: All primary data are publicly available (https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3). Analysis scripts: All analysis scripts are publicly available (https://osf.io/a3vzc/?view_only=cd6ff959753044a881eecf07da07a3a3).

Rational vigilance of intentions and incentives guides learning from advice

Why do many trust teachers and scientists, but few trust lobbyists or politicians (Nicholas, 2022)? Why do juries believe the testimony of eyewitnesses more than defendants (Loftus, 1984)? And why do product reviews from fellow consumers carry more weight than ads (Nielsen, 2012)? *Inferred motivations* provide a unifying explanation across these cases: Some people (e.g., teachers) do not typically benefit from manipulating our beliefs and actions, while others (e.g., politicians) stake their careers on social influence. Accordingly, research has shown that people are highly vigilant of their informants' motivations: For example, patients who learn about payments from the pharmaceutical industry to physicians grow skeptical of professional medical advice (Kanter et al., 2019). But *why* and precisely *how* do motivations influence inferences from testimony? For instance, do people always lose faith in communication when informants are incentivized or selfish? Or do they jointly evaluate their informants' context, incentives, and intentions in a more complex calculus? If people consider many factors, do they do so heuristically or rationally? Answering these questions requires fulfilling two aims: first, identifying the key ingredients of vigilant inference (where much progress has been made), and second, outlining how an ideal reasoner would combine these ingredients to draw inferences from others' advice (where crucial gaps remain).

Sperber et al. (2010) distill decades of research on the components of vigilant inference to two factors: "(...) the communicator's *competence* on the topic of her assertions, and her *motivation* for communicating." Accordingly, research has extensively studied the psychological, evolutionary, and sociological principles of vigilance of competence as a pillar of social learning (for reviews, see Mercier, 2017; Sobel & Kushnir, 2013). This research has shown, for instance, that people are skilled at tracking informant accuracy (Landrum et al., 2015; Soll & Larrick, 2009) and that this skill develops remarkably early in children (P. L. Harris, 2012).

Vigilance of motivations is a second pillar of social learning, and it entails sensitivity

to two interacting components: the *intentions* and *incentives* guiding testimony (Locke et al., 1968). Research suggests that people are just as sensitive to these considerations as competence: even young children often prioritize intentions over competence when forced to choose between informants (Mascaro & Sperber, 2009; but see Johnston et al., 2015). Research on adults also suggests that people can reasonably discount testimony on the basis of basic intentional information (Handley-Miner et al., 2023; A. J. Harris et al., 2016; Shafto et al., 2012). Research across fields has shown that people are also vigilant of others' incentives (Bond Jr et al., 2013; Loewenstein et al., 2014; Shaw et al., 2023). For instance, the perceived credibility of scientific research decreases when people learn that it has been privately funded (Critchley & Nicol, 2009).

Yet large gaps pervade our understanding of the *mechanisms* of vigilance: for instance, we do not know whether people are optimally vigilant in realistic communicative settings. This is because there are no rational models that shed light on how people *should* evaluate key inputs of vigilant inference, such as others' incentives, intentions, and context. Here we address this gap through a formal model of vigilant inference, and test the predictions of this model through two experiments across medical, financial, and consumer decisions.

A Formal Model of Vigilance of Motivations

Vigilance of motivations requires understanding *why* speakers say what they do. Such pragmatic language understanding has been productively formalized in the Rational Speech Acts (RSA) framework (Frank & Goodman, 2012). In RSA, speakers choose utterances (u) to update listeners' beliefs about the true world state (w). Reasoning about the speaker's aim to inform allows 'pragmatic listeners' to move beyond literal interpretation. Classical RSA assumes that the speaker is trying to inform listeners' *beliefs* about the world state, while recent work has extended RSA to capture cases in which speakers are trying to guide listener's *actions* (a) instead (Sumers et al., 2024). Importantly, these models assume that speakers and listeners are purely cooperative

(Degen, 2023) or purely adversarial (Oey et al., 2023), which limit their ability to account for vigilance. We thus extend this work by considering a *mixed motives* setting, where the speaker's cooperative stance can interpolate between these two extremes of full cooperation and pure deception.

Rational Vigilance of Motivations

We model vigilance of motivations by considering how a listener interprets a speaker's recommendation given mixed incentives and intentions (see Supplementary Materials SMA for a detailed derivation). We assume that the speaker and listener have instrumental reward functions, $R_S(a)$ and $R_L(a)$, that reflect the value that each party receives if the listener takes an action ($a \in A$). The speaker's 'joint' reward function can then be modeled as a convex combination of their own reward and the listener's reward,

$$R_{\text{Joint}}(R_L, R_S, \lambda, a) = \lambda R_L(a) + (1 - \lambda) R_S(a), \tag{1}$$

where $\lambda \in [0, 1]$ parametrizes benevolence. As $\lambda \to 0$, the speaker approaches pure selfishness, and considers only their personal instrumental reward; as $\lambda \to 1$, the speaker approaches pure altruism.

In deciding what to say, the speaker chooses from all possible action-reward pairs. The speaker's probability of choosing an utterance (P_S) depends on the expected utility, which is given by the reward associated with each action (see Equation 1 above), times the probability of a naive listener (L_0) taking it (formally, their policy, $\pi_{L0}(a \mid u)$):

$$P_S(u \mid R_S, R_L, \lambda, A) \propto \exp\{\beta_S \cdot \sum_{a \in A} R_{\text{Joint}}(R_L, R_S, \lambda, a) \pi_{L0}(a \mid u)\}, \tag{2}$$

where β parametrizes the speaker's optimality. Finally, a pragmatic listener (L_1) who is given a recommendation formalizes *vigilance* by reasoning about both the true reward of the prescribed action to themselves (R_L) and the speaker's motivations (their instrumental reward, R_S and helpfulness, λ):

$$P_{L1}(R_L \mid u) \propto P(u \mid R_S, R_L, \lambda, A) P(R_s) P(R_L) P(\lambda). \tag{3}$$

In the simplest setting, listeners do not know how good different actions will be for them (i.e., they are uninformed about R_L), but they do know how their actions will benefit the speaker (R_S), and how benevolent the speaker is (λ). These prior beliefs can account for context-specific dependencies between speakers' and listeners' rewards, and allow listeners to exercise vigilance.

When listeners and speakers have independently distributed rewards, our model predicts that listeners should grow increasingly skeptical of testimony as speakers receive higher incentives. This is because learning that speakers have high incentives 'explains away' the possibility that the action was also good for the listener (Wellman & Henrion, 1993). Importantly, this skepticism should be moderated by the benevolence of the speaker, as benevolent speakers' recommendations are primarily driven by listener outcomes.

Experiment 1: Rational Inferences from Incentives

To test these predictions of our model experimentally, we examined responses to vignettes across three domains in which informants with varying intentions and incentives provided advice to participants. Participants either made inferences about the quality of a card they could sign up for (*Credit Card*), the quality of a house they might purchase (*Real Estate*), or the quality of a medical prescription (*Medical*).

Methods

Participants

Participants were 626 adults (284 male, 337 female, 4 other, mean age = 38.9) recruited on Prolific in exchange for monetary compensation ($1.60 for a 8-minute study). As pre-registered, an additional 34 participants were excluded for completing the study too quickly (less than 3 minutes in an 8 minute study), 36 were excluded for failing either of our two attention checks, and 44 were excluded for failing to recognize that a clearly better reward (e.g., 5% commission rather than 0% commission for a real estate agent) is better for the informants. All experiments were approved by the institutional review board of [author institution], and all participants gave informed consent. Participation across pilots

and experiments was restricted using the Prolific platform such that participants could only participate in one study. Participation across all studies was restricted to users currently residing in the United States with an approval rating $\geq 98\%$ on at least 100 tasks. The sample size was determined on the basis of prior pilot testing to provide $\geq 98\%$ power to detect a key interaction effect at $\alpha = .001$ (see Supplementary Materials SMB). The design, sample size, and analyses were pre-registered (see https://aspredicted.org/zp8r-3tbr.pdf), and any deviations from the pre-registered analyses are noted as exploratory.

Materials and Procedure

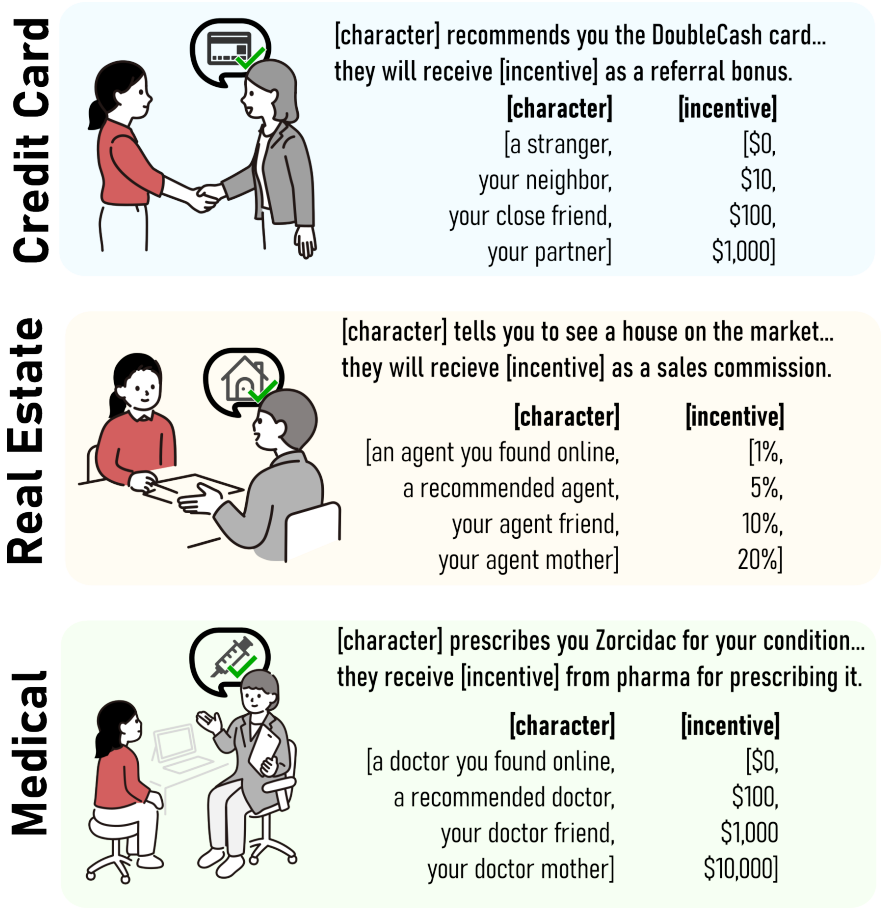
Participants evaluated 16 vignettes in each scenario: they evaluated recommendations made by four different informants—chosen to vary in perceived intentions—across four levels of incentives (see Figure 1). As an example, in the credit card vignette, participants read:

> You are interested in getting a credit card. One day, as you are having a conversation with [informant], the topic of credit cards come up. [The informant] tells you that they have done a lot of research and they think the new DoubleCash card is the best.
>
> Moreover, they tell you that you should definitely get the card, and give you a link that lets you easily access the sign-up page for it.

Where the informant was either a stranger, a neighbor, a close friend, or a romantic partner. Participants then learned about referral bonuses ("A referral bonus is a cash reward someone may get for convincing another person to sign up for a card"). They then considered four cases in which the informant either received i) no referral bonus, ii) a $10 referral bonus, iii) a $100 referral bonus, iv) and a $1,000 referral bonus. In each case, participants indicated how good they thought the card was by indicating their agreement with the following claim: "The DoubleCash card is a much better choice for me than other cards." They rated their agreement using a 7-point Likert scale from "Strongly Disagree"

Figure 1

Vignettes Used in Study



Note. A diagram illustrating the set of vignettes used in each scenario. Participants responded to all 16 combinations of incentives and informants in each scenario, with informant order randomized.

[1] to "Strongly Agree" [7], with "Neither Agree Nor Disagree" as a neutral midpoint [4]. Participants completed 16 key trials in which they evaluated the quality of the card for each combination of informants and incentives.

After completing these key trials, participants provided an open-ended explanation of their strategy during the experiment, and completed several other ratings. First, they rated their informants' communicative intentions ("When others provide us with information, they can be self-interested, care about what's best for us, or both care about their own well-being and ours. Consider the people you had in mind in the previous

question. Please rate how much they care about themselves vs. your well-being.") on a slider from 'entirely self-interested' [0] to 'only about you' [100] with 'both self-interested and cares about you' [50] as a midpoint. They then rated the informants' incentives—in Credit Card, for instance, they rated how good they thought the differing referral bonuses were, from 'getting this bonus would not matter at all' [0] to 'would be extremely good to get this bonus' [100]. The slider handles were hidden until participants clicked on the slider to indicate their preferences; the informants and incentives were all displayed in random order. Finally, participants provided demographics (age, sex, education), and answered to two attention checks. Note that for model fitting purposes, we rounded the intention and incentive scores to deciles and quartiles from 0 to 1, respectively (e.g., an initial score of 84% would be rounded to 0.8 for the intention score, and .75 for the incentive score). The vignettes and questions used in the Real Estate and Medical scenarios were conceptually equivalent to those shown above for Credit Card (See Supplementary Materials SMC for the exact wording in each scenario).

Given that our data contained multiple measurements from each participant, we used linear mixed-effects regressions implemented through the `lme4` package (Bates et al., 2014) in `R` across all of our analyses. In keeping with recent recommendations, we first fit 'maximal models' including random slopes and intercepts, and iteratively simplified models if they did not converge (Barr et al., 2013). To fit our Bayesian model, we used a probabilistic programming language, `webPPL` (Goodman & Stuhlmüller, 2014).

The Supplementary Materials (SMD) contains a detailed description of how model predictions were generated, and our open access repository contains code, data, and instructions for executing the model. A key detail is that our implementation of the model evaluates the full space of possibilities present in the inference problem (i.e., every possible combination of incentives and intentions is represented as a distinct possibility). As our key behavioral measure asks for agreement with the claim that the recommended option is better than alternatives, we compare human inferences and model predictions by using the

probability that the recommended option is better than alternatives (i.e., $R_{L_R} = 1$ & $R_{L_A} = 0$).

In fitting our model, we use three free parameters that we fit to participant data. These are the noisiness of speakers, β_S, the noisiness of the literal listener, β_{L0}, and a parameter κ that rescales the benevolence parameter, λ, according to λ^κ. The noisiness parameters are fit as they are difficult to measure explicitly. The rescaling of benevolence allows us to account for nonlinearities in the use of the trust scale (see Supplementary Materials SMD for details).

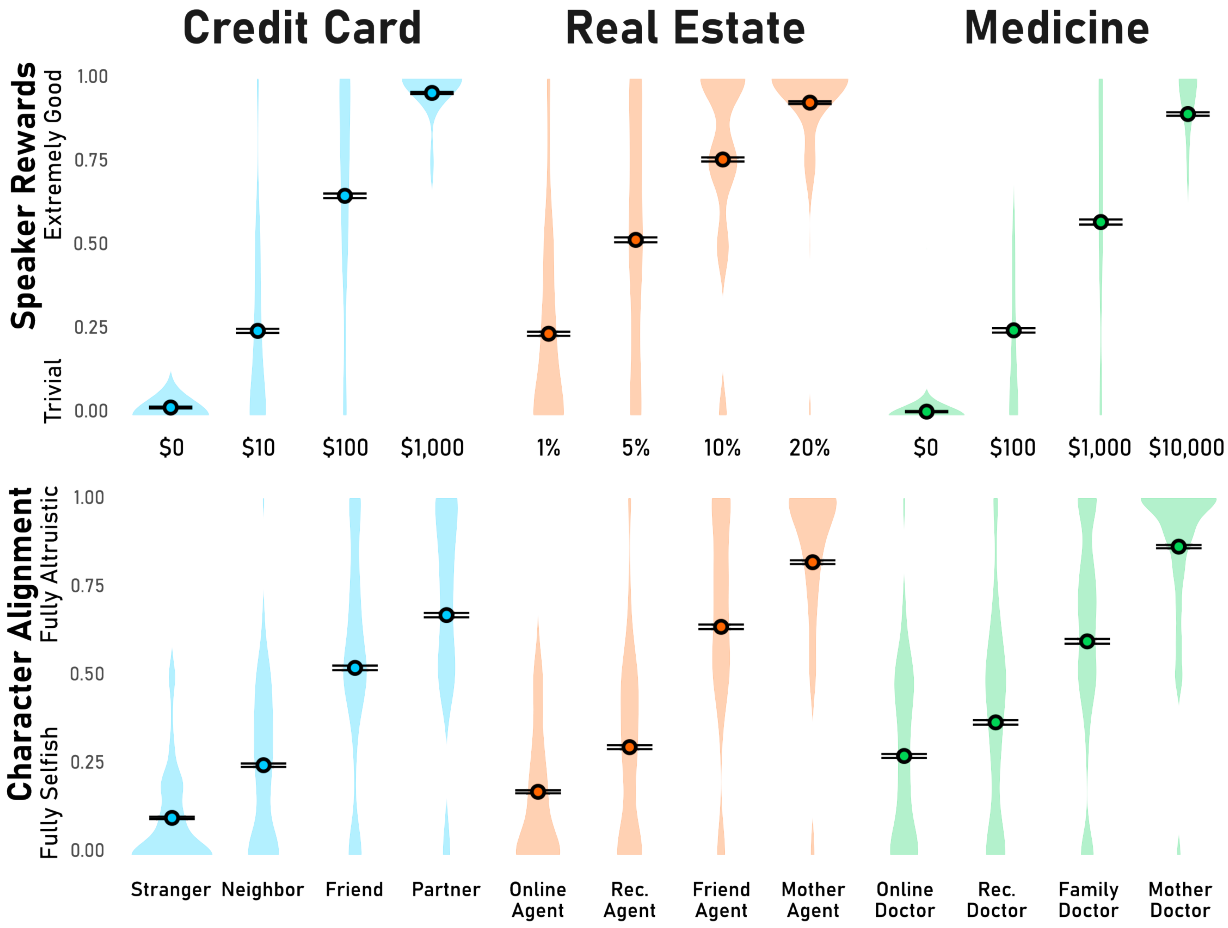
Results

We first investigated whether our manipulations of speakers' intentions and incentives were successful. As expected, participants inferred that informants with closer relationships to themselves to be increasingly benevolent, $\hat{\beta} = 0.18$, 95% CI $[0.16, 0.20]$, $t(1.99) = 19.32$, $p = .003$. Participants also perceived larger referral bonuses to be increasingly desirable for speakers, $\hat{\beta} = 0.27$, 95% CI $[0.20, 0.33]$, $t(2.01) = 7.90$, $p = .015$. For these analyses, we operationalized the effect of different relationships and incentive magnitudes through an ordinal mapping (with the least close informant and smallest incentive receiving 1, and the closest informant and largest incentive receiving 4). These results were obtained from a mixed-effects regression across the three vignettes with random slopes and intercepts for each participant.

Having established that our manipulations effectively altered perceived intentions and incentives, we investigated the effects of these factors on participant inferences. Our key pre-registered analysis regressed participants' inductive inferences—their belief in the quality of the recommended option—on (i) the perceived self-interestedness of the informant, (ii) the informant's incentives for their recommendation, (iii) the interaction of these factors, as well as (iv) random intercepts and slopes for each participant. For parsimony, we present the parameters of a single model that includes data across all scenarios. Regression tables for all models reported in this section—including regressions

Figure 2

Manipulation Checks for Intentions and Incentives



Note. Plot shows that the manipulations of incentives and intentions worked as intended across all conditions, with participants recognizing that larger financial incentives were more beneficial for speakers, and that closer relationships indicate greater benevolence. The violin plots show the distribution of data, the black circle shows the mean response, and the lines around the circle show ± 1 standard error of the mean.

fitted to individual scenarios—can be found in Supplementary Materials (SME).

In line with past research in social cognition, we find that participants are likelier to form beliefs in line with testimony provided by benevolent others, $\hat{\beta} = 0.18$, 95% CI $[0.17, 0.19]$, $t(624.99) = 52.09$, $p < .001$ (i.e., increasing benevolence, λ, increases participants' beliefs that the recommended option is in fact better than alternatives). In

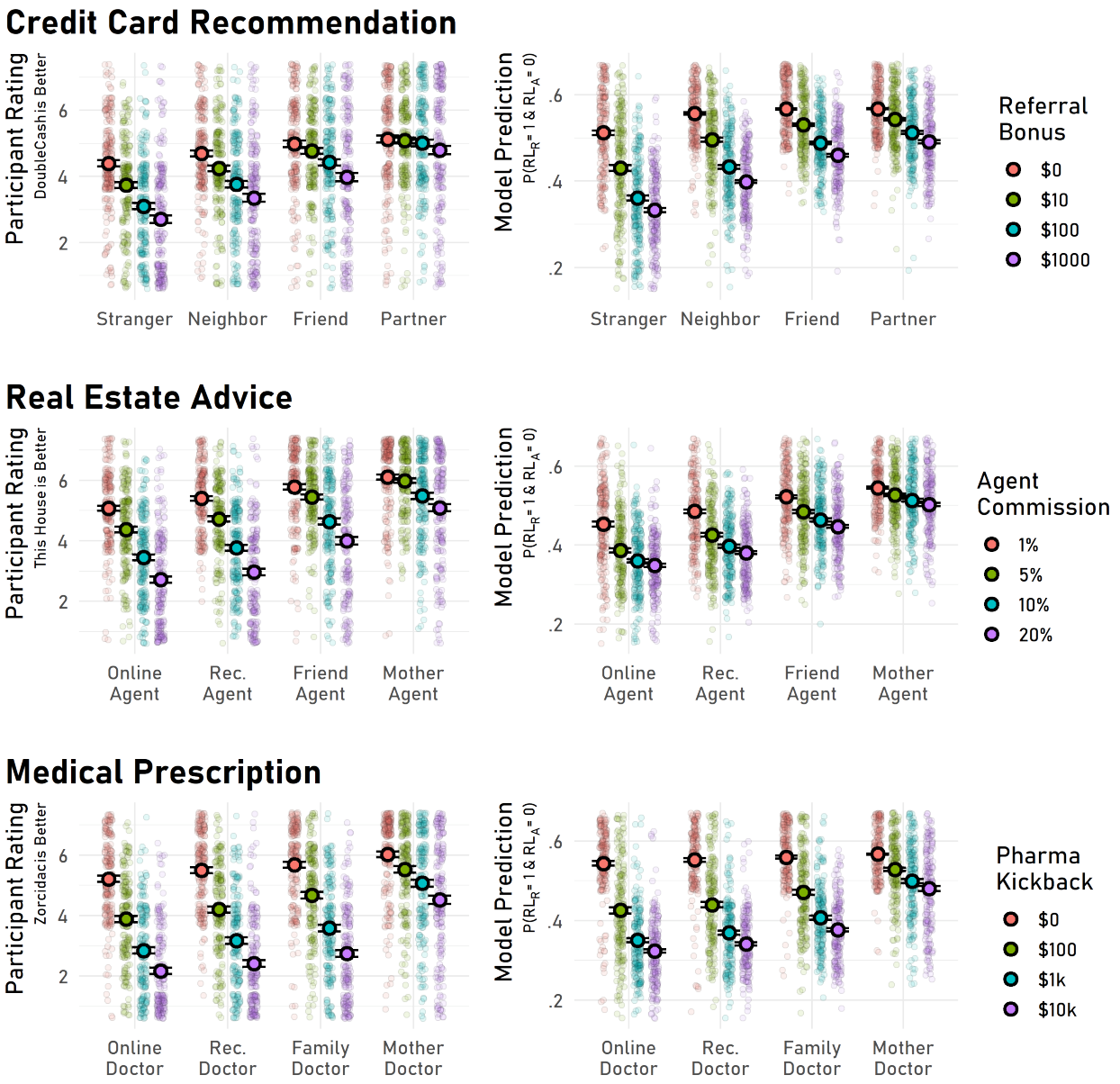
line with past research in applied psychology, we find that increasing incentives for the speaker (R_S) have the opposite effect, lowering the likelihood that participants form beliefs aligned with the speaker's recommendation, $\hat{\beta} = 0.27$, 95% CI $[0.26, 0.28]$, $t(624.98) = 68.01$, $p < .001$. Our findings also shed new light on how motivations influence inferences, by revealing a substantial role for the interaction of incentives and intentions, $\hat{\beta} = 2.18$, 95% CI $[1.92, 2.45]$, $t(451.15) = 16.09$, $p < .001$ (see the left column of Figure 3). This interaction captures the intuitive insight that listeners expect benevolent informants (such as a mother prescribing a medication to their children) to be less sensitive to incentives (such as a pharmaceutical kickback) than informants who are also self-interested.

Importantly, these dynamics are also present in our model's predictions. We conducted the regression described above on the probability assigned by the model to the recommended option being better than alternatives, $P(R_{L_R} = 1 \;\&\; R_{L_A} = 0)$. This analysis reveals that benevolent intentions increase the informativeness of testimony, $\hat{\beta} = 0.11$, 95% CI $[0.10, 0.11]$, $t(1167.83) = 37.15$, $p < .001$, while incentives explain away the informativeness of testimony, $\hat{\beta} = -0.27$, 95% CI $[-0.27, -0.26]$, $t(1056.91) = -98.95$, $p < .001$, with an interaction across these two factors, $\hat{\beta} = 0.21$, 95% CI $[0.21, 0.22]$, $t(1204.13) = 65.84$, $p < .001$ (see the right column of Figure 3). Note that this interaction is a direct reflection of the incentive structure that the optimal vigilant listener relies on to make inferences (see Equation 1).

These analyses show that our model qualitatively replicates patterns of vigilant inference in humans. To assess the quantitative fit between predictions and inferences, we correlated mean model predictions and participant inferences across our 16 experimental conditions. This pre-registered analysis reveals high alignment between model and participant responses (see Figure 4). We observed significant and substantial correlations across Credit Card, $r = .96$, 95% CI $[.89, .99]$, $t(14) = 13.35$, $p < .001$, Real Estate, $r = .94$, 95% CI $[.82, .98]$, $t(14) = 9.89$, $p < .001$, and Medical, $r = .99$, 95% CI $[.97, > .99]$, $t(14) = 24.80$, $p < .001$. In an exploratory analysis, we additionally examined whether our

Figure 3

Human and Model Inferences Across Intentions and Incentives



Note. Figure shows the distribution of participant inferences (left column) and model predictions (right column) across the three scenarios on each row. The faint jittered points plot the distribution of individual participant inferences and the corresponding model predictions, whereas the larger outlined dots show the mean estimate, and the bars around the mean show ± 1 standard error of the mean. 'Rec.' corresponds to the recommended agent and doctor in the latter scenarios.

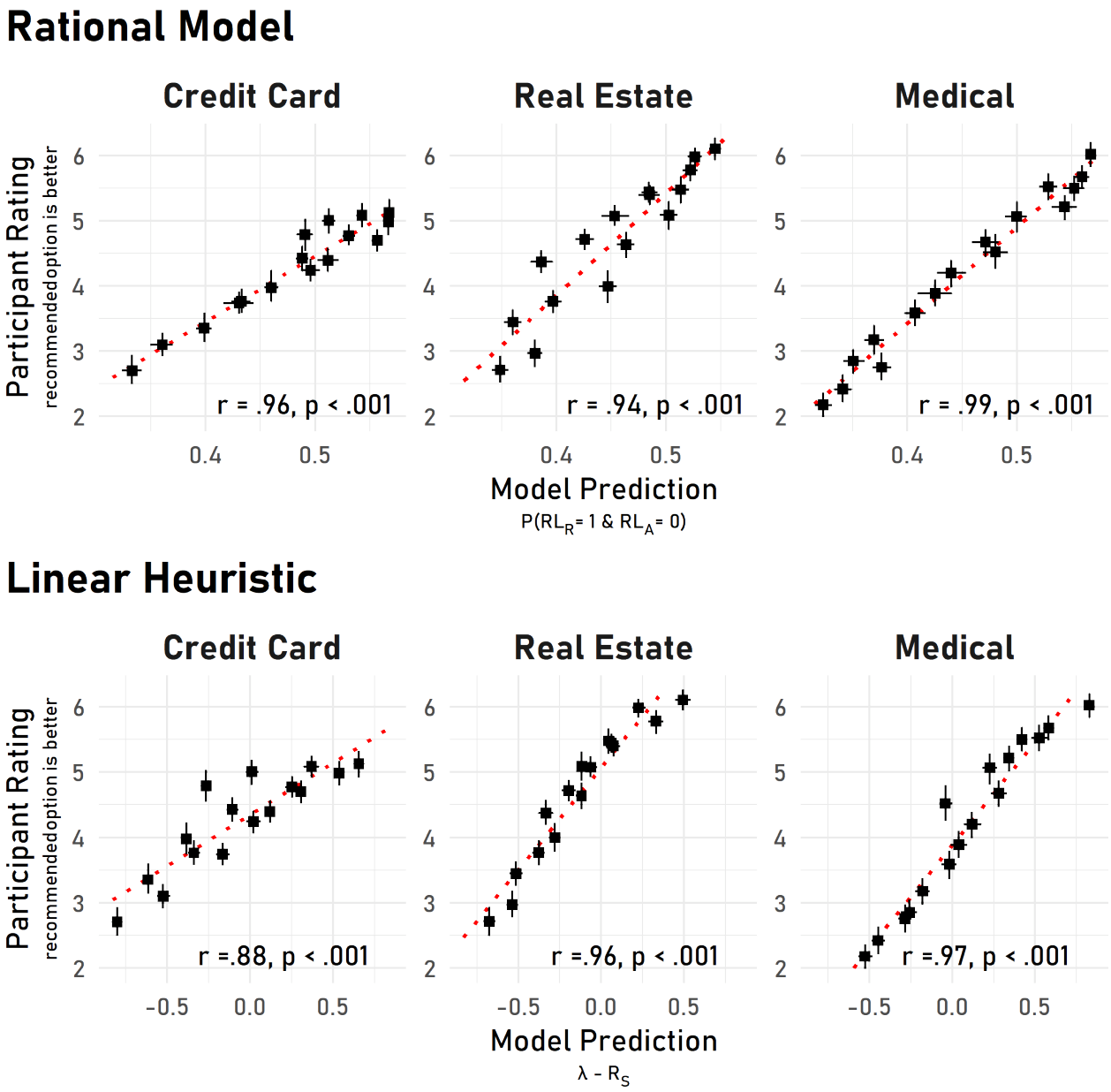
model characterized individual participants' responses to experimental manipulations (vs. the pattern of average responses). We found that the model was positively correlated with 85% of individual participants, with a mean correlation of .60, 95% CI [.57, .63], and median correlation of .74; see Supplementary Materials SMF for these additional analyses), suggesting that our model captures both aggregate and individual inferences. Lower individual-level correlations are commonly observed in such models due to a combination of noisy responding and probabilistic inference mechanisms at the individual level (Eberhardt & Danks, 2011; Vul et al., 2014).

While these results show that our model can capture people's inferences, they do not show that our model is the best or only predictor of people's inferences. Eyeballing the data in Figure 3, for instance, it is plausible that a simple, linear heuristic can capture much of the variance in people's responses. People may simple track the difference between speaker incentives and perceived benevolence (i.e., $P(R_{L_1} = 1, R_{L_0} = 0) \approx \lambda - R_S$). In fact, classic work in judgment and decision-making has shown that a family of models that make similar predictions to this linear heuristic predict people's inferences when it comes to vigilance of competence and credibility (Birnbaum et al., 1976).

This heuristic captures the 'explaining away' effect we observe for informants of ambivalent benevolence (e.g., strangers). Exploratory analyses using the Akaike Information Criterion (AIC)—an information-theoretic measure of predictive fit that automatically penalizes models for complexity—support this intuition. This linear heuristic fits participants overall inferences better than our Bayesian model (AIC_{Bayes} = 32,431.55, $\text{AIC}_{Heuristic}$ = 31,772.38, ΔAIC = 659.17; lower scores indicate better performance). This analysis computes the likelihood assigned by the model to each data-point and penalizes model complexity in the aggregate (with the Bayesian model having 3 free parameters and the linear heuristic having none). We can also more intuitively quantify the variance explained by each model using a measure of R^2 applicable to mixed-effects models (Nakagawa & Schielzeth, 2013), which reveals that the Bayesian model explains essentially

Figure 4

Model-Human Alignment Across Experimental Conditions



Note. Black points show the mean model prediction (horizontal axis) and participant inference (vertical axis) across our 16 conditions, with error bars showing bootstrapped 95% confidence intervals on each axis. The line of best fit is shown in red. Perfect alignment between model predictions and human inferences corresponds all points being on the red line.

the same amount of variance in the dataset as the heuristic ($R^2_{Bayes} = .317, R^2_{Heuristic} = .319, \Delta R^2 = .002$). This suggests that the substantial difference in AIC across models is a consequence of the accumulation of minute differences in predictive accuracy across our large dataset ($\approx 10,000$ observations) that do not amount to a substantive difference in overall variance explained.

Processing the data to isolate the effect of the interaction shows that the heuristic performs worse than the Bayesian model at characterizing how these factors relate to each other. Instead of regressing model predictions directly to participant inferences, we can compute difference scores for each incentive condition in our study. For instance, we can compute the difference in inferences across \$0 and \$10 referral bonuses in the credit card case. If benevolence and incentives interact, then we would expect the size of this difference to shrink across informants of increasing benevolence (e.g., your friend likely cares enough about you to not mislead you for \$10, but a stranger might). As expected, this exploratory analysis shows that our model performs better in predicting these difference scores than the linear model ($\text{AIC}_{Bayes} = 12{,}489.22$, $\text{AIC}_{Heuristic} = 12{,}671.64$, $\Delta\text{AIC} = 182.42$).

Nuances aside, these models ultimately make very similar predictions across the data from this study (the raw correlation of model predictions across the entire dataset is $r = .85$, 95% CI $[.85, .86]$, $t(10,014) = 162.85$, $p < .001$). Theoretically, this is because the 'explaining away' phenomenon that causes incentives to decrease inferred quality is the primary effect in Experiment 1. This effect itself is a consequence of the speaker and listener rewards being independently distributed—under such circumstances, the heuristic and the rational model are hard to distinguish empirically. When this independence does not hold, however, the predictions of the linear heuristic and our model diverge sharply. For instance, when rewards are positively correlated, our model predicts that higher speaker incentives substantially *increase* the probability that testimony is accurate. The linear heuristic, on the other hand, would continue to predict that increasing incentives will explain away testimony.

Experiment 2: Positive Correlation in Incentives

In Experiment 2, we therefore introduced minimal changes to our previous three vignettes that coupled speakers' and listeners' rewards. If people flexibly adjust to such reward structures (and our model continues to capture their inferences) there would be a stronger case that they are rationally vigilant. If people dogmatically discount incentive-based testimony, the case for rational vigilance would be weaker.

Methods

Participants

Participants were 555 adults (221 male, 326 female, 8 other, mean age = 40.6) recruited on Prolific in exchange for monetary compensation ($1.60 for a 8-minute study). As pre-registered, an additional 1 participant was excluded for completing the study too quickly (less than 3 minutes), 47 were excluded for failing either of our two attention checks, and 65 were excluded for failing to recognize that a clearly better reward (e.g., 5% commission rather than 0% commission for a real estate agent) is better for the characters. An additional pre-registered criterion excluded 83 participants who did not comprehend that the scenario entailed positively correlated rewards. Since this study was structurally identical to the previous study and used the same analyses, we recruited the same number of participants. The design, sample size, and analysis for this experiment were pre-registered (see https://aspredicted.org/rdyf-c4vx.pdf).

Materials and Procedure

As in Experiment 1, participants evaluated 16 vignettes in each of three scenarios: they evaluated recommendations made by four different characters—chosen to vary in perceived intentions—across four levels of incentives. However, the vignettes in this study coupled speaker and listener rewards: For the credit card vignette, participants learned that the referral bonus would be split; for real estate, participants learned that high-quality houses that need to be sold urgently both have high commission rates and are also sold at

steep discounts; and for medicine, participants learned that pharmaceutical payments to doctors were provided on the basis of the doctor's previous success in treating patients.

Prior to evaluating speakers' recommendations, participants read a high-level summary of the vignettes that introduced the idea that rewards are correlated. As an example, for the credit card vignette, participants first read:

> Credit cards sometimes offer referral bonuses of varying amounts. A shared referral bonus is a cash reward someone may get for getting others to sign up for a card they have—moreover, the person who signs up also receives a portion of the bonus.
>
> The DoubleCash card offers shared referral bonuses that are *split equally* between the person who recommends the card and the person who receives it.

Participants were then asked to evaluate the correlation between listener and speaker rewards through the following question:

> Consider the following situation. Someone is advising you to sign up for the DoubleCash card. In some situations, it benefits us when the people we are engaging with are rewarded. (...) In other situations, it costs us when the people we are engaging with are rewarded. (...) In this scenario, would it benefit you if the person advising you received a large referral bonus?

Participants provided their responses on a 9-point Likert scale, from high certainty of positive correlation (The person receiving a large referral bonus '...is absolutely certain to benefit me') to high certainty of negative correlation ('...is absolutely certain to come at a personal cost'), with independent rewards as a midpoint ('...is irrelevant to my outcomes'). Participants then received feedback about whether they were correct, and were all provided with another high-level summary of why rewards are positively correlated, before re-evaluating the same question. The rest of the study design was identical to Experiment

1 (see Supplementary Materials SMB for the exact wording in all three scenarios).[1]

For our model to generate predictions that take these inferred reward correlations into account, we need to define a joint reward distribution, R, that specifies how R_L and R_S are related. We did so through the following equation, which parametrizes the extent of dependence through $\rho \in [-1, 1]$,

$$P(R_L = 1 \mid R_S = x) = \begin{cases} (1-\rho) \cdot 0.5 + \rho \cdot x, & \text{if } \rho \geq 0 \\ (1-|\rho|) \cdot 0.5 + |\rho| \cdot (1-x), & \text{if } \rho < 0. \end{cases} \tag{4}$$

This allows us to generate mixture reward distributions that interpolate between independent reward distributions, $\rho = 0$, perfect positive correlation, $\rho = 1$, and perfect anti-correlation, $\rho = -1$. The vigilant listener, speaker, and literal listener all use this correlated reward distribution in their inferences.

We fit our model using largely the same procedure as in Experiment 1. The primary difference in modeling was that, for this study, we needed to find an optimal mapping between participant's responses to our reward-dependency question and the correlation parameter, ρ. To find this mapping, we performed separate grid searches in each vignette, and found that a simple mapping that assigns a relatively low correlation to participants that infer low likelihoods of reward dependence (1 or 2 on the likert scale), and a high correlation to those that infer high likelihood of dependence (3 or 4 on the likert scale), performed best (see Supplementary Materials SMD for details and plots).

Results

As pre-registered, we analyzed whether our context-sensitive Bayesian model better predicted people's inferences than the linear heuristic through AIC comparisons. We found that the Bayesian model performed better in predicting participant's inferences than the linear heuristic (AIC_{Bayes} = 30,333.51, $\text{AIC}_{Heuristic}$ = 30,744.46, ΔAIC = 410.95; these

[1] With the minor exception that participants evaluated whether houses were a 'good deal' in this experiment, as opposed to 'a good fit' for their preferences in Experiment 1.

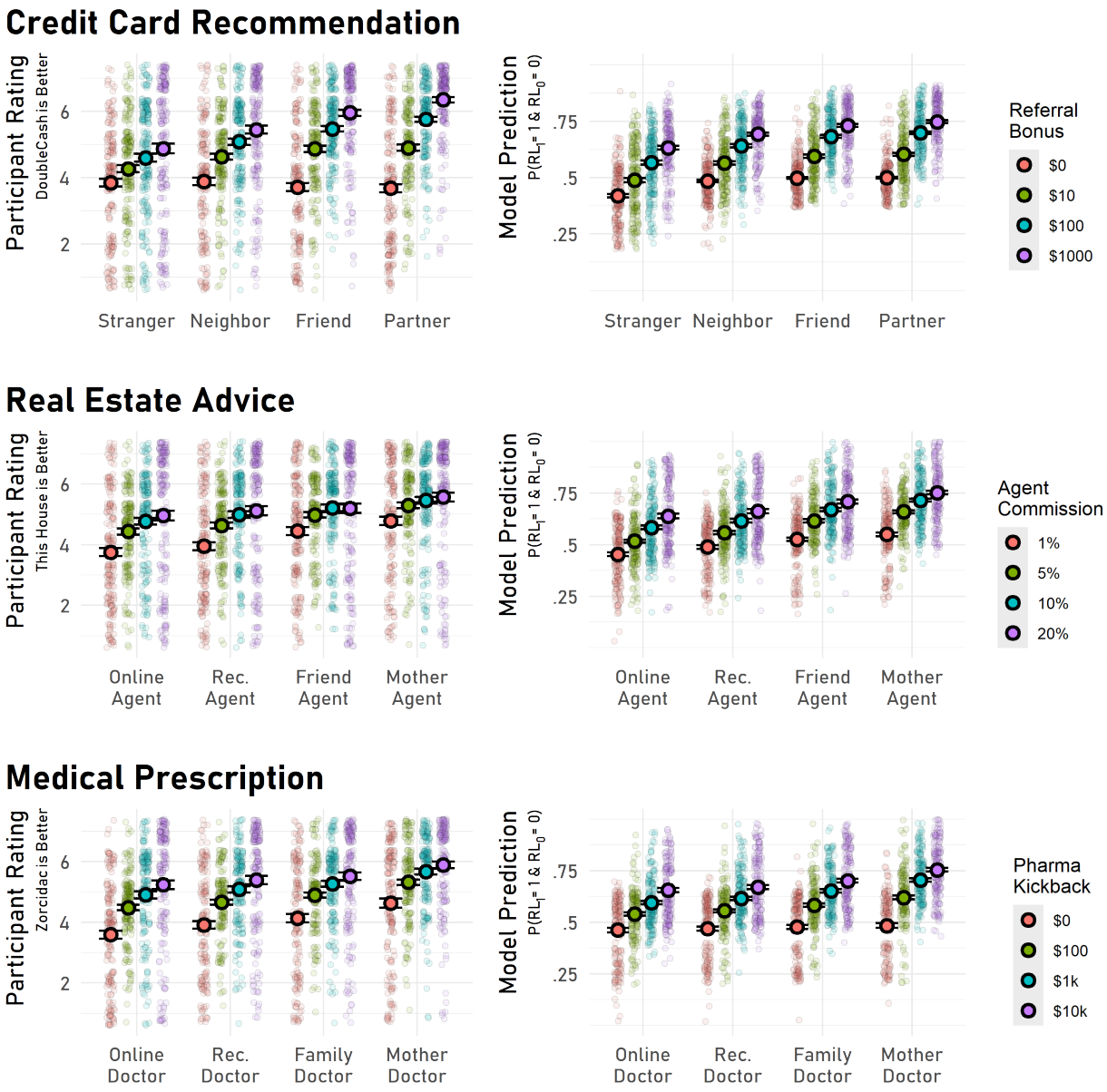
analyses penalize the Bayesian model for the additional 2 free parameters used to estimate the correlation structure). This is because people's inferences align with the modified reward structure we examined in this study, with increasing incentives increasing people's tendency to infer a higher likelihood of truth from testimony (see Figure 5).

To assess the quantitative fit between predictions and inferences, we again correlated mean model predictions and participant inferences across our 16 experimental conditions. This analysis reveals that the average model-participant correlations are high in each vignette (in *Credit Card*, $r = .96$, 95% CI $[.89, .99]$, $t(14) = 13.31$, $p < .001$; in *Real Estate*, $r = .96$, 95% CI $[.89, .99]$, $t(14) = 13.00$, $p < .001$; in *Medical*, $r = .96$, 95% CI $[.88, .99]$, $t(14) = 12.54$, $p < .001$). This suggests that our model succeeds in capturing the latent function relating inferences across conditions in this study as well (see Figure 6). One nuance that our model does not capture is an interaction in *Credit Card*, whereby increasing levels of benevolence leading to increasing positive sensitivity to rewards. This interaction seems to be absent in the other two vignettes. One explanation for this trend is that there may be cross-character variance in the certainty of perceived reward correlation, which we cannot account for, as we only measured reward correlation once for each participant. For instance, one's friends or partners may be seen as much less likely to lie about how referral bonuses work, resulting in changes in the certainty of the reward corelation. In *Real Estate* and *Medical*, on the other hand, speakers are rendering professional services. This may stabilize the certainty of the reward structure, as there are costs to deceit in these contexts.

The linear heuristic performs considerably worse than the rational model—in fact, it is systematically misaligned with participant inferences in some cases (for *Credit Card*, $r = -.59$, 95% CI $[-.84, -.14]$, $t(14) = -2.76$, $p = .015$; for *Real Estate*, $r = -.21$, 95% CI $[-.64, .32]$, $t(14) = -0.79$, $p = .445$; for *Medical*, $r = -.49$, 95% CI $[-.80, .00]$, $t(14) = -2.13$, $p = .051$). This is because participants showed appropriate sensitivity to the reward structure, and did not continue to dogmatically discount incentivized testimony,

Figure 5

Human and Model Inferences Across Intentions and Incentives



Note. Figure shows the distribution of participant inferences (left column) and model predictions (right column) across the three scenarios on each row. The faint jittered points plot the distribution of individual participant inferences and the corresponding model predictions, whereas the larger outlined dots show the mean estimate, and the bars around the mean show ± 1 standard error of the mean. 'Rec.' corresponds to the recommended agent and doctor in the latter scenarios.

which is what the linear heuristic would have predicted. We also see these trends in a participant-level analysis. Our model is positively correlated with 78% of participant's inferences, with an average correlation of .41 (bootstrapped 95% CI [.37, .44]), whereas the linear heuristic is only positively correlated with 35% of participant's inferences, with an average correlation of -.16 (bootstrapped 95% CI [-.21, -.12]).

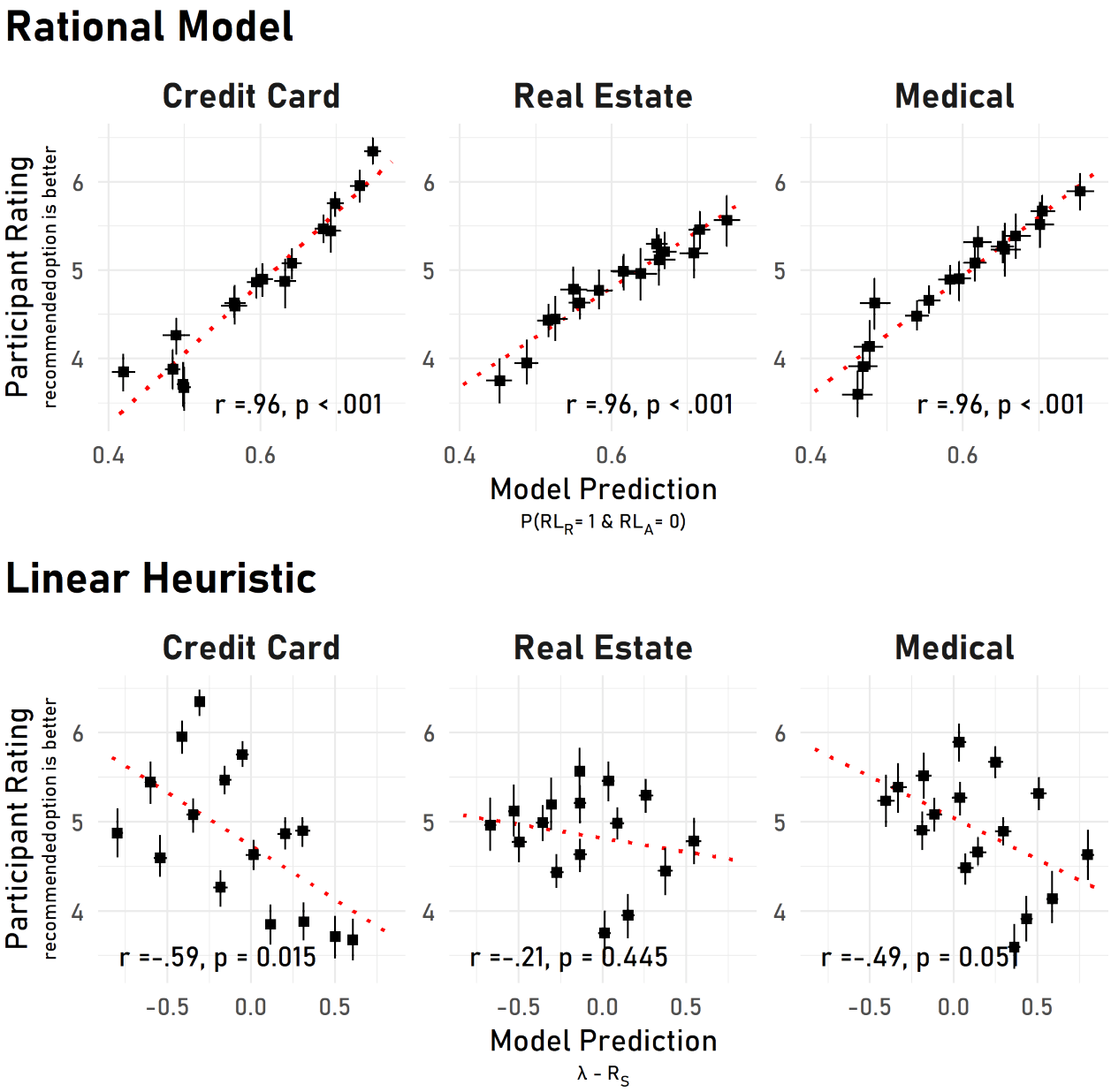
General Discussion

Testimony is powerful: people can convince juries to free the innocent (or convict them), persuade the public to vaccinate (or to drink bleach), and influence consumers to sign up for advantageous credit cards (or fall for scams). Learning from testimony thus requires inference mechanisms that incorporate both trust and vigilance. In this paper, we shed new light on these mechanisms through a Bayesian model of instrumental communication. This model explains how prior beliefs and evidence about others' motivations rationally guide inferences from testimony. We also investigated key empirical predictions of this model through two experiments, and found that participants' responses closely replicate the dynamics of the model across diverse decision scenarios. Below, we first situate these findings within the broader literature on vigilance, then outline fruitful directions for future research, and conclude by considering important theoretical implications of this work for our understanding polarization and societal disagreement.

Much work across different disciplines has examined the mechanisms underlying vigilance of competence. Yet those who deceive us are often not incompetently misinformative, but instead deliberately shape their messages to manipulate our behavior, because they have a stake in our judgments and decisions. Yet existing work on motivational vigilance does not provide rational explanations for *why* we draw the inferences that we do. Our Bayesian model addresses these gaps: We propose that people rationally consider (i) their prior beliefs about how selfish others are, and (ii) the evidence that they obtain about others' and their own outcomes, (iii) in light of their communicative context which relates these outcomes (iv) through recursive social reasoning. This model

Figure 6

Model-Human Alignment Across Experimental Conditions in Experiment 2



Note. Black points show the mean model prediction (horizontal axis) and participant inference (vertical axis) across our 16 conditions, with error bars showing bootstrapped 95% confidence intervals on each axis. The line of best fit is shown in red. The top row shows the predictions of our rational model, and the bottom row shots the predictions of the linear heuristic.

leads to surprising predictions that are reflected in participants' behavior: Learning about others' motivations can 'explain away' the informativeness of testimony—even when potential rewards are known to be independently distributed. Yet such explaining away can be blocked by prior knowledge of informant's intentions (Experiment 1), or context-driven coupling between speaker and listener outcomes (Experiment 2).

Though our models' predictions broadly align with participant behavior across the three scenarios we investigated, further research is necessary to examine the extent to which the model generalizes to other cases. A domain of particular interest is political polarization. Recent work highlights identities (Van Bavel & Pereira, 2018), rational mechanisms of evidential inference (Jern et al., 2014) and other mechanisms (Jost et al., 2022) as causing belief polarization. Our work demonstrates a novel and complementary explanation for how polarization occurs even when partisans are exposed to the same testimony: People may have differing prior beliefs about others' instrumental motivations; with partisans believing informants aligned with the other side to be systematically selfish. Beyond polarization, addressing inferences of selfish or nefarious intent may allow people to re-evaluate their views on entrenched societal disagreements (Oktar & Lombrozo, 2024).

In evaluating the evidence for our rational model that our studies provide, it is important to consider the possibility that other linear heuristics could capture people's inferences in Experiment 2. For instance, simply inverting the linear heuristic we examined, such that rewards are added to the benevolence parameter (rather than subtracted) would result in well-calibrated predictions. Importantly, however, no single linear heuristic can account for inferences in both studies (as the slope for incentives has to be negative in the first study and positive in the second study). Hence, our results conclusively show that people do not dogmatically rely on contextually-inappropriate heuristics for vigilant inference in our studies. While it is possible that people may adaptively deploy a mixture of heuristics in a context-sensitive manner (Lieder & Griffiths, 2017), such a strategy would only be adaptive insofar as it approximates the vigilance our

model prescribes. That is, this would be a different algorithm that aims to realize the rational, computational-level analysis we have presented here (Marr, 1982).

Our results show we can use rational analysis to make sense of inferences across a relatively small set of contexts with participants in the U.S in an online study. Much exciting work remains to be done examining vigilance across broader communicative, cultural, and experimental contexts. Our theory and experiments lay the groundwork for this future research on the nature and mechanisms of vigilance.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Birnbaum, M. H., Wong, R., & Wong, L. K. (1976). Combining information from sources that vary in credibility. *Memory & Cognition*, *4*, 330–336.

Bond Jr, C. F., Howard, A. R., Hutchison, J. L., & Masip, J. (2013). Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology*, *35*(2), 212–221.

Critchley, C. R., & Nicol, D. (2009). Understanding the impact of commercialization on public support for scientific research: Is it about the funding source or the organization conducting the research? *Public Understanding of Science*, *20*(3), 347–366.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, *21*, 389–410.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Stuhlmüller, A. (2014). The Design and Implementation of Probabilistic Programming Languages.

Handley-Miner, I. J., Pope, M., Atkins, R. K., Jones-Jang, S. M., McKaughan, D. J., Phillips, J., & Young, L. (2023). The intentions of information sources can affect what information people think qualifies as true [Article 7718, 10 pages]. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-34806-4

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a bayesian network approach. *Cognitive Science*, *40*(6), 1496–1533.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Jern, A., Chang, K.-m. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Johnston, A. M., Mills, C. M., & Landrum, A. R. (2015). How do children weigh competence and benevolence when deciding whom to trust? *Cognition*, *144*, 76–90.

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, *1*(10), 560–576.

Kanter, G. P., Carpenter, D., Lehmann, L. S., & Mello, M. M. (2019). Us nationwide disclosure of industry payments and public trust in physicians. *JAMA network open*, *2*(4), e191947–e191947.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794.

Locke, E. A., Bryan, J. F., & Kendall, L. M. (1968). Goals and intentions as mediators of the effects of monetary incentives on behavior. *Journal of Applied Psychology*, *52*(2), 104–121.

Loewenstein, G., Sunstein, C. R., & Golman, R. (2014). Disclosure: Psychology Changes Everything. *Annual Review of Economics*, *6*, 391–419. https://doi.org/10.1146/annurev-economics-080213-041341

Loftus, E. F. (1984). Expert testimony on the eyewitness. In G. Wells & E. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 273–283). Cambridge University Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380.

Mercier, H. (2017). How gullible are we? a review of the evidence from psychology and social science. *Review of General Psychology*, *21*(2), 103–122.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Nicholas, B. (2022). The IPSOS global trustworthiness index press release. *IPSOS Report*, 1–6.

Nielsen. (2012). Global trust in advertising and brand messages [Retrieved Online from Nielsen.com on January 12, 2024].

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346–362.

Oktar, K., & Lombrozo, T. (2024). How beliefs persist amid controversy: The Paths to Persistence Model. https://doi.org/10.31234/osf.io/9t6va

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science*, *15*(3), 436–447.

Shaw, E., Lynch, M., & Scurich, N. (2023). Juror perceptions of incentivized informant testimony. *Psychology, Crime & Law*, *30*(1), 1–19.

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*, 779–797.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393.

Sumers, T. R., Ho, M. K., Griffiths, T. L., & Hawkins, R. D. (2024). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological review*, *131*(1), 194–230.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Wellman, M. P., & Henrion, M. (1993). Explaining'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(3), 287–292.