A Rational Model of Vigilance in Motivated Communication

Kerem Oktar (oktar@princeton.edu)
Department of Psychology, Princeton University

Theodore Sumers (sumers@princeton.edu)
Department of Computer Science, Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Department of Psychology and Computer Science, Princeton University

Abstract

We are able to learn from others through a combination of trust and vigilance: we trust and believe people who are reliable and have our interests at heart; we ignore those who are incompetent or self-interested. While past work has studied how others' competence influences social learning, relatively little attention has been paid to how others' motivations influence such processes. To address this gap, we develop a Bayesian model of vigilance that considers the speaker's instrumental self-interest, and test predictions of this model through an experiment. In accordance with our model, participants become more vigilant when informants stand to benefit from influencing their actions. When perceived self-interest is maximal, testimony can be discounted wholesale, rendering middle ground increasingly difficult, if not impossible, to find. Our results have implications for research on polarization, misinformation, and societal disagreement.
Keywords: vigilance; disagreement; communication; motivation; inference;

Introduction

Why do most trust doctors and teachers, but few trust lobbyists or politicians? Why do we believe the testimony of eyewitnesses more than defendants? And why do product reviews from fellow consumers carry more weight than ads?[1] *Inferred motivations* provide a common explanation across these cases: Some people—like doctors and eyewitnesses—typically do not benefit from manipulating our beliefs and actions, while others—like politicians and defendants—stake their careers and lives on social influence. Despite starkly influencing who we listen to, believe, and rely on, inferred motivations have received little direct attention from the cognitive science community. In particular, we lack formal models that can make precise predictions about how people adjust their beliefs in others' testimony based on evidence of instrumental motivations. Here, we address this gap through a novel computational model of these inferences, and present an experiment that tests key predictions of our model. Beyond explaining consequential inferences, our findings pave the way towards a novel account of polarization and have implications for the design of interventions that increase alignment in trust.

[1] According to a global public opinion poll by IPSOS, 50% trust teachers when only 12% trust politicians (Nicholas, 2022). Similarly, when the conflicting testimonies of a defendant and an eyewitness are the primary evidence in a trial, 75% of defendants are convicted (Loftus, 1984). Finally, 92% indicated trusting word-of-mouth recommendations more than advertisements, according to a global consumer poll (Nielsen, 2012).

Prior Work: Social Learning and Vigilance

Selective Learning from Testimony

Others' testimony extends the reach of our beliefs beyond the horizon of our experiences (Lippmann, 1922). For instance, we can learn that the climate is warming without feeling hot—and that vaccines work without falling ill—because we trust the testimony of scientists. Yet learning from testimony is not as simple as merely trusting what we are told. Other people may be incompetent or ignorant, or even manipulative, aiming to sway our judgments and decisions in service of their own interests (Harris, 2012). Philosophers have long recognized the importance of critical, reflective inquiry in shielding one's beliefs from undue manipulation by others (Fricker, 1995). Such *epistemic vigilance* is thought to undergird the very possibility of cooperative communication and social knowledge transmission, as it allows people to selectively trust reliable testimony (Sperber et al., 2010).

The Need for Research on Vigilance of Motivation

Research in psychology has empirically investigated the mechanisms underlying vigilance (Mercier, 2017; Clément, 2010), and recent cross-disciplinary work aims to investigate its macro-scale implications (Reisinger, Kogler, & Jäger, 2023; Flache et al., 2017). Yet some aspects of vigilance remain elusive. As Sperber et al. (2010) noted in a landmark review of vigilance: "What is most urgently needed is (...) research on how trust and mistrust are calibrated to the situation, the interlocutors and the topic of communication. Here, two distinct types of consideration should be taken into account: the communicator's *competence* on the topic of her assertions, and her *motivation* for communicating." Since this review, much work has focused on the vigilance of competence, whereas vigilance of motivation has been relatively understudied. We outline this research and the gaps in our understanding of vigilance before describing how our rational analysis addresses these gaps.

Vigilance of Competence. Epistemologists have provided formal Bayesian models of source credibility (Merdes, Von Sydow, & Hahn, 2021; Olsson, 2011), psychologists have shown that sophisticated mechanisms for reliability inference are present both in adults and children (Aboody, Yousif, Sheskin, & Keil, 2022; Langenhoff, Engelmann, & Srinivasan, 2023), and computational cognitive scientists

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

have shown how such inferences can be formalized (Shafto, Goodman, & Frank, 2012; Landrum, Eaves, & Shafto, 2015). Bayesian analyses have been recently been expanded to explain people's inferences from persuasive messages (Bhui & Gershman, 2020), and how people can detect the difference between systematically biased and noisy information sources (Schulz, Schulz, Bhui, & Dayan, 2023). The literature on inferences of reliability or competence has thus received extensive, cross-disciplinary interest (Harris, Koenig, Corriveau, & Jaswal, 2018).

Vigilance of Motivation. Despite this progress, the question of how others' *motivations* influence vigilance and inference has received much less direct attention. Most of the relevant work comes from public policy and science communication, which has examined how views of scientists change in response to evidence about their funding sources. This research has found that the credibility of research decreases when people learn that it has been privately funded (Critchley, 2008; Critchley & Nicol, 2009), and that scientists working as lobbyists are perceived as having less integrity and benevolence than publicly funded scientists (König & Jucks, 2019; Johnson & Dieckmann, 2020). Motivation-induced vigilance has consequences beyond lower trust: Revealing a lobbying-association causes people to interpret misleading evidence provided by a scientist more accurately (Gierth & Bromme, 2020). More generally, the literatures on persuasion (Cialdini & Goldstein, 2004; O'Keefe, 2018) and negotiations (Mnookin, 1992; Reynolds, 2020) have shown that revealing the motivations underlying a message can make them more or less convincing, depending on whether the motivations are self-interested, altruistic, or reciprocal (O'Keefe, 2013; Whatley, Webster, Smith, & Rhodes, 1999). Despite this cross-discipline body of evidence that inferred motivations influence the persuasive appeal of messages, explanations for *why* or *how* they do so remain informal.

Research has thus shown that we lose trust in communication when it could be driven by self-interest (Mercier, 2020). But why does this happen? We seek to answer this question by developing a rational model of communicative inference amid instrumental motivation.

A Formal Model of Vigilance

The Rational Speech Acts framework

Our model builds on the Rational Speech Acts (RSA) framework (Frank & Goodman, 2012). RSA is a formal model of language understanding, modeling communication as a recursive reasoning process between a speaker and a listener. Speakers choose utterances u to update the listener's beliefs about the true world state w. Formally, speakers choose an utterance according to a utility function $U(u,w)$:

$$P_S(u \mid w) \propto \exp\{\, \beta_S \cdot U(u,w) \,\}, \tag{1}$$

where β_S is a soft-max parameter controlling speaker optimality. The speaker is assumed to address a so-called "literal" listener typically hold a uniform prior over possible world states. L_0 denotes their posterior beliefs after hearing the utterance. The speaker thus reasons about the literal listener's beliefs after hearing the utterance:

$$P_{L0}(w \mid u) \propto \delta_{[\![u]\!](w)}\, P(w), \tag{2}$$

where $\delta_{[\![u]\!](w)}$ denotes the meaning of u, returning one when utterance u is true of w and zero otherwise. L_0 is *literal* because it uses only lexical meanings. To formalize pragmatic language understanding, RSA defines a *pragmatic* listener, L_1. The pragmatic listener embeds a speaker model (which in turn embeds a literal listener, L_0):

$$P_{L1}(w \mid u) \propto P_S(u \mid w) P(w). \tag{3}$$

Reasoning about the speaker's intent allows the pragmatic listener to move beyond a literal interpretation and consider the speaker's communicative intentions (see Figure 1).

While the classical RSA speaker objective assumes the speaker is trying to reduce the listener's uncertainty about the world state (Frank & Goodman, 2012), recent work has extended RSA to consider decision-theoretic utility (Sumers, Ho, Griffiths, & Hawkins, 2023). This approach models the listener as a rational agent who will act in the world. The speaker's communication influences their beliefs, and thus their subsequent actions. Listeners are modeled as noisy-rational actors who choose an action a from a set of available actions A. The listener updates their beliefs after hearing an utterance (Eq. 2), then marginalizes over them to estimate the reward for action a:

$$R_L(a,u) = \sum_{w \in W} R(a,w) P_{L_0}(w \mid u). \tag{4}$$

The scalar reward value for an action is defined by the world state: $R : \mathcal{A} \times W \to \mathbb{R}$. R_L represents the listener's posterior beliefs about their decision problem: it specifies the listener's expected reward for action a after hearing utterance u. The listener then forms a decision policy π_L as a softmax (Savage, 1954), choosing an action a according to:

$$\pi_L(a \mid u, A) \propto \exp\{\, \beta_L \cdot R_L(a,u) \,\}, \tag{5}$$

where β_L is the listener's softmax optimality. Then, rather than seeking to be informative, speakers can instead maximize the listener's expected rewards:

$$U_{\text{Speaker}}(u \mid w, A) = \sum_{a \in A} \pi_L(a \mid u, A) R(a,w). \tag{6}$$

While RSA has been used to model a wide range of linguistic phenomena (Degen, 2023), it rests on a fundamental—and fundamentally unrealistic—assumption: that speakers and listeners are purely cooperative.[2] In our model, we consider a

[2]Notably, studies of deception use similar recursive reasoning, but make the opposite assumption: that communication is purely adversarial (Oey, Schachner, & Vul, 2023; Alon, Schulz, Rosenschein, & Dayan, 2023).

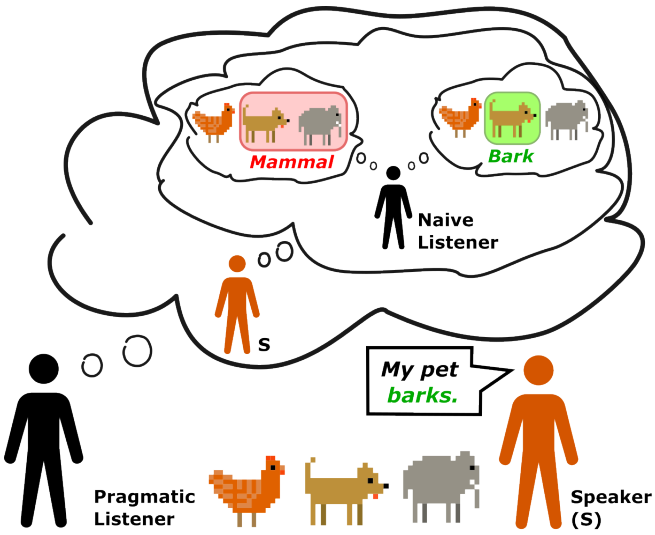

Figure 1: A graphic illustrating the recursive reasoning process in the Rational Speech Acts framework. In this case, we can explain why a dog owner might say "my pet barks" instead of "my pet is a mammal," even though both statements are true. This is because "mammal" is under-informative for a naive listener, since it could refer to either the elephant or the dog.

mixed motives setting where the speaker's cooperative stance can interpolate between these two extremes (full cooperation and pure deception), and extend RSA to enable inference over the speaker's latent motivational state. In our empirical tests we focus on how knowledge of this cooperativeness parameter influences listener's inferences.

Computational Model

We extend prior work on instrumental communication (Sumers et al., 2023; Sumers, Hawkins, Ho, Griffiths, & Hadfield-Menell, 2022; Van Rooy, 2003; Benz, 2006; Franke, 2009) by relaxing the assumption that the speaker is cooperative, instead allowing the speaker and listener's reward functions to diverge. A pragmatic listener then faces an additional challenge: they must infer the speaker's *motivations* for producing an utterance, in order to determine whether or not to believe it.

Formally, we assume that the speaker and listener have *independent* instrumental reward functions, $R_S(a)$ and $R_L(a)$ respectively. These reflect the value that each party receives if the listener takes an action a.[3] For example, in a trial setting, we can imagine that the listener is a juror whose actions are to vote to *convict* or *acquit*. In this case, we expect that the defendant benefits from a vote to acquit regardless of their guilt ($R_S(\text{acquit}) > 0$). In contrast, an eyewitness is a neutral third party with minimal stake in the outcome ($R_S(\text{acquit}) \approx 0$).

[3]For notational simplicity, we drop the dependence on w from prior work, writing $R(a)$ instead of $R(a|w)$. This assumes that the world state is synonymous with the rewards of actions.

The speaker's reward function can then be modeled as a convex combination of their instrumental reward and the listener's instrumental reward,

$$R_{\text{Joint}}(R_L, R_S, \lambda, a) = \lambda R_L(a) + (1-\lambda) R_S(a), \qquad (7)$$

where $\lambda \in [0,1]$. When $\lambda = 0$ the speaker is purely self-interested, and considers only their personal instrumental reward. When $\lambda = 1$ the speaker is purely altruistic.

Then the speaker's probability of choosing an utterance is a softmax over the expected utility – the "joint" reward associated with each action, times the probability of the listener taking it:

$$P_S(u \mid R_S, R_L, \lambda, A) \propto \exp \sum_A R_{\text{Joint}}(R_L, R_S, \lambda, a) * \pi_L(a \mid u). \quad (8)$$

A purely self-interested speaker manipulates the listener for their own benefit, choosing utterances that lead the listener to maximize R_S regardless of the consequences for themselves (R_L). In contrast, an altruistic speaker guides the listener to maximize R_L. Intermediate values lead to a compromise between speaker and listener's benefit.

Finally, a pragmatic listener formalizes *epistemic vigilance* by reasoning about both the true reward of an action (R_L) and the speaker's motivations (their instrumental reward, R_S and helpfulness, λ). The vigilant listener marginalizes over the speaker's motivations to perform inference over R_L, using their priors over R_S, R_L and λ to decide whether or not to believe the speaker:

$$P(R_L \mid u) \propto P(u \mid R_S, R_L, \lambda, A) P(R_s) P(R_L) P(\lambda). \qquad (9)$$

In the next section, we use simulations to illustrate how a vigilant listener's beliefs are shaped by these priors.

Simulations

We consider a minimal setting where a listener must choose from three actions, $A = (a_0, a_1, a_2)$. Each action is worth 0 or 1 to the speaker and listener and their reward functions are independent of each other. We assume $\beta_S = \beta_L = 10$ and place a uniform prior over possible reward functions. Finally, the set of utterances consists of all action-value pairs, i.e. "a_0 is worth 0", "a_0 is worth 1", "a_1 is worth 0", and so on.

To demonstrate the effects of epistemic vigilance, we vary the listener's priors over the speaker's motivations. We assume the speaker says "a_0 is worth 1" and then plot the listener's resulting posterior (Equation 9) in Figure 2B. The leftmost plot shows how the speaker's altruism affects the listener's inference. When the listener believes the speaker is self-interested ($\lambda = 0$) they discount testimony entirely; their posterior remains uniform over [0,1]. As λ increases, the listener is increasingly willing to believe the speaker, and their posterior shifts towards the stated value ("a_0 is worth 1"). The middle plot shows how the speaker's reward influences the listener's beliefs. When the speaker does not benefit from the action ($R_S = 0$) the listener is willing to believe them. However, as R_S increases, the listener is more skeptical. Finally,

the rightmost plot varies both the speaker's altruism and rewards. Altruistic speakers are trusted regardless of whether they stand to benefit, whereas self-interested speakers are met with increasing skepticism as their potential rewards increase.

Testing the Model's Predictions

We presented participants with a simple vignette where different people recommend a credit card, with differing amounts of referral bonuses motivating their recommendations. Participants provided their judgments of the quality of the card across conditions, and then provided relevant social judgments (e.g., how self-interested they think the recommenders tend to be).

Our primary aim in this study was to investigate three foundational questions regarding the psychology of vigilance: First, do people in fact systematically decrease their trust in others' testimony when they learn that they have independent instrumental motivations? Second, do the ways in which they adjust their credences accord with the predictions of our computational model? And finally, do people's prior levels of trust in other individuals moderate the effect of instrumental motivations? Given that the recursive inferences underlying instrumental reasoning are quite complex, people could be relatively insensitive to further considerations, such as the identity of the informant.

Methods

Participants Participants were 77 adults (48 male, 29 female, 0 other, mean age = 37.4) recruited on Prolific in exchange for monetary compensation ($1.20 for a 6-minute 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.

Materials and Procedure Participants read four sets of vignettes in random order. Each set involved the same scenario, but different characters. As an example, here is the vignette involving a close friend:

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

Participants then indicated how good they thought the card was initially, using a 7-point Likert scale from "Much worse than other cards" [1] to "Much better than other cards" [7], with "Neither worse nor better" as a neutral midpoint [4]. Participants next 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 scenarios, and re-rated the same card-quality scale for each. In these scenarios, they considered learning that the informant either received i) no referral bonus, ii) a $10 referral bonus, iii) a $100 referral bonus, iv) and a $1000 referral bonus. Participants completed these judgments separately for four characters: a stranger, a neighbor, a close friend, and a romantic partner.

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 how self-interested the characters are ("When others interact with us, they can be self-interested, care about us, or both care about their own and our well-being. Consider the people you had in mind in the previous question. Please rate how much they care about themselves vs. you.") on a slider from 'entirely self-interested' [0] to 'only wants what is best for you' [100] with 'both self-interested and cares about you' [50] as a midpoint. They then rated how good they thought the differing referral bonuses were, on average, from 'getting this bonus would not matter at all' [0] to 'would be extremely good to get this bonus' [100]. Finally, they provided ratings of how competent the characters are, from 'totally incompetent at knowing whether a card is good' [0] to 'extremely competent at knowing whether a card is good' [100]. Participants finally provided demographics (age, gender, education), as well as a self-reported attention check.

Results

Given that our data contained multiple measurements from each participant, we used linear mixed-effects regressions implemented through the `lme4` package (Bates, Mächler, Bolker, & Walker, 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, Levy, Scheepers, & Tily, 2013).

Manipulation Checks. We first investigated whether our manipulations worked as intended (see Figure 2A). As expected, participants inferred strangers, neighbors, close friends, and lovers to be increasingly altruistic, $\hat{\beta} = 21.27$, 95% CI $[19.11, 23.42]$, $t(74.01) = 19.37$, $p < .001$. Across analyses of different characters, we operationalized the effect of different relationships through an ordinal mapping.

Participants also perceived larger referral bonuses to be increasingly desirable, $\hat{\beta} = 29.07$, 95% CI $[26.62, 31.51]$, $t(74.00) = 23.30$, $p < .001$; though reported utility showed diminishing sensitivity to increasing dollar amounts, replicating a finding that extends back to Bernoulli (1738).

Altruism moderates trust. Having established that participants perceived differences in altruism across characters, we investigated whether this altruism translated to increased trust in the characters' testimony. As predicted by our computational model (see Figure 2B, left panel), participants generally drew stronger inferences from the testimony of characters perceived as more altruistic, $\hat{\beta} = 3.07$, 95% CI $[2.81, 3.34]$,

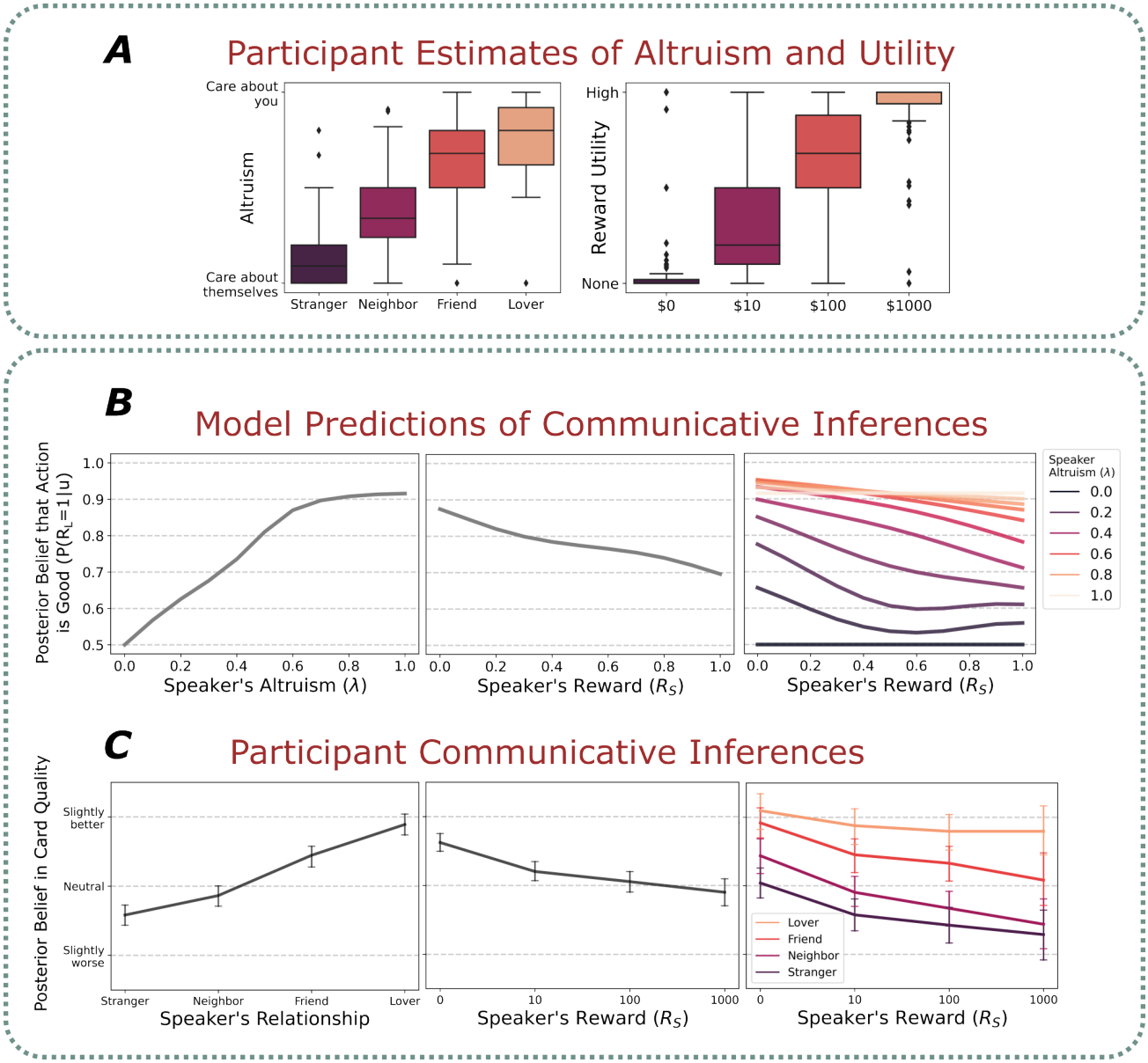

Figure 2: Results from simulations and the behavioral experiment. **A.** Box plots of altruism and reward utility for the experimental conditions. As expected, participants perceive closer others as increasingly altruistic in their motivations, and they perceive higher dollar amounts as better (though with diminishing marginal returns, such that exponential increases in dollar amounts linearly increase utility). **B.** Simulations illustrating how the speaker's motivations affect the listener's inference. Each plot shows the listener's posterior estimate after hearing an action is worth 1. Left: when the listener believes the speaker is self interested ($\lambda = 0$) they discount their testimony entirely. As λ increases, the listener is increasingly willing to believe the speaker. Center: When the speaker does not benefit from the action ($R_S = 0$) the listener is willing to believe them. As R_S increases, the listener is more skeptical. Right: combining both altruism and reward. Altruistic speakers are trusted regardless of whether they stand to benefit, whereas self-interested speakers are met with increasing skepticism as their own instrumental rewards increase. **C.** Participants' inferences about how good the recommendation is for them, as a function of the same features explored in the simulations. Note that participant inferences go below neutral in the case of the stranger, indicating that participants are considering the possibility that the card might be bad for them, a possibility we did not explore in the simulation. The line connects mean reward inferences and 95% confidence intervals across conditions.

$t(74.00) = 22.52$, $p < .001$ (see Figure 2C, left panel).

Incentives moderate trust. Across all characters, our model predicts that increased speaker incentives (i.e., referral bonuses) should explain away listener rewards, leading to lower trust (see Figure 2B, middle panel). We also observe this decreasing marginal trust with increasing rewards in our data, $\hat{\beta} = -0.23$, 95% CI $[-0.35, -0.11]$, $t(74.00) = -3.70$, $p < .001$ (see Figure 2C, middle panel).

Interaction between altruism and incentives.

Beyond the intuitive direct relationships between altruism, incentives, and trust, the key prediction of our computational model was an *interaction* between these factors. In particular, our model predicts that high prior beliefs about altruism should protect trust from the effect of incentives. For instance, at the limit of pure altruism, knowledge of incentives should have practically no effect on trust (see Figure 2B, right panel). We also observe this key interaction in our data, $\hat{\beta} = 0.05$, 95% CI $[0.01, 0.09]$, $t(74.00) = 2.34$, $p < .05$ (see Figure 2C, right panel). The more altruistic the character, the weaker the decrease in trust from incentives.

Discussion

Testimony is powerful: we 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 an online experiment, and found that participants' responses closely replicate the dynamics of the model. 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 of misinformation, polarization, and societal disagreement.

Much work across different disciplines has examined the mechanisms underlying vigilance of *competence*—how we come to selectively learn from the reliable and knowledgeable, and discard the testimony of noisy or overconfident sources (Harris et al., 2018). 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. Psychologists have tested how such personal stakes influence trust through intuitive, qualitative predictions—for instance, investigating whether lobbying reduces trust in scientists (König & Jucks, 2019). Critically, however, the existing work on motivational vigilance does not provide rational explanations for *why* we draw *precisely* the inferences that we do. Our Bayesian model addresses these gaps: We propose that people rationally consider (i) their prior beliefs about how self-interested others are, and (ii) the evidence that they obtain about others' and their own outcomes, (iii) through recursive social reasoning. This model 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 (i.e., without assuming zero-sum rewards; or that others profit when we suffer).[4]

Though our models' predictions broadly align with participant behavior in the specific case of inferences from credit cards, two additional steps need to be taken to broadly validate it as an account of epistemic vigilance. First, the computational model needs to be fit to participant data to enable quantitative comparisons between model predictions and human inferences. Second, further experiments are necessary to investigate how well our model generalizes to other cases. Given that people's judgment and decision-making is remarkably sensitive to the decision domain in question (Oktar & Lombrozo, 2022a), our model might need additional contextual information to be able to predict inferences across contexts. In particular, people's assumptions about how others' and their own outcomes tend to be related in a given domain will substantially influence the predictions of the model.

A domain of particular interest is politics. With partisan animosity and misalignment increasing in the U.S and globally, recent research has focused on elucidating the psychological mechanisms underlying polarization (Iyengar, Lelkes, Levendusky, Malhotra, & Westwood, 2019). This work has uncovered identities (Van Bavel & Pereira, 2018), cognitive limitations (Pennycook, McPhetres, Bago, & Rand, 2022), and other mechanisms (Jost, Baldassarri, & Druckman, 2022) as causing polarization. Our work demonstrates a complementary explanation for how polarization persists even when partisans are exposed to the same testimony: People may have differing prior beliefs about others' instrumental motivations; with partisans believing sources aligned with the other side to be systematically self-interested. This mechanism may explain why certain interventions—such as exchanging personal narratives (Kalla & Broockman, 2020)—can reliably shift views amid polarized conversations: These interventions help listeners calibrate their understanding of the motivations underlying political or scientific testimony. Beyond polarization, addressing inferences of selfish or nefarious intent may allow people to re-evaluate their views on entrenched societal disagreements (Oktar & Lombrozo, 2022b), and reconsider their intuitive judgments through communication and deliberation (Oktar, Lerner, Malaviya, & Lombrozo, 2023).

[4]Such explaining away happens commonly in causal reasoning: If we see rain on our window, and notice that the sprinkler was on, we infer that it is unlikely to be raining, even if sprinklers are turned on randomly. This is because both rain and sprinkling are unlikely to occur at the same time, compared to just one of them occurring (Wellman & Henrion, 1993).

Acknowledgments

We thank members of the Computational Cognitive Science Lab for their feedback. This work and related research were made possible by the support of the NOMIS Foundation.

References

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2022). Says who? Children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General*.

Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023). A (dis-) information theory of revealed and unrevealed preferences: emerging deception and skepticism via theory of mind. *Open Mind*, *7*, 608–624.

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*.

Benz, A. (2006). Utility and relevance of answers. In *Game theory and pragmatics* (pp. 195–219). Springer.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, *5*, 175–192. (Mem. for 1730-31)

Bhui, R., & Gershman, S. J. (2020). Paradoxical effects of persuasive messages. *Decision*, *7*(4), 239–258.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*, 591–621.

Clément, F. (2010). To trust or not to trust? children's social epistemology. *Review of philosophy and psychology*, *1*, 531–549.

Critchley, C. R. (2008). Public opinion and trust in scientists: the role of the research context, and the perceived motivation of stem cell researchers. *Public Understanding of Science*, *17*(3), 309–327.

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.

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of Social Influence: Towards the Next Frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4), 2.

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

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. University of Amsterdam.

Fricker, E. (1995). Telling and trusting: Reductionism and anti-reductionism in the epistemology of testimony. *Mind*, *104*, 393-411.

Gierth, L., & Bromme, R. (2020). Beware of vested interests: Epistemic vigilance improves reasoning about scientific evidence (for some people). *PLoS One*, *15*(4), e0231387.

Harris, P. L. (2012). *Trusting What You're Told: How Children Learn from Others*. Harvard University Press.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive Foundations of Learning from Testimony. *Annual Review of Psychology*, *69*(1), 251–273.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129–146.

Johnson, B. B., & Dieckmann, N. F. (2020). Americans' views of scientists' motivations for scientific work. *Public Understanding of Science*, *29*(1), 2–20.

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.

Kalla, J. L., & Broockman, D. E. (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, *114*(2), 410–425.

König, L., & Jucks, R. (2019). Hot topics in science communication: Aggressive language decreases trustworthiness and credibility in scientific debates. *Public Understanding of Science*, *28*(4), 401–416.

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.

Langenhoff, A. F., Engelmann, J. M., & Srinivasan, M. (2023). Children's developing ability to adjust their beliefs reasonably in light of disagreement. *Child Development*, *94*(1), 44–59.

Lippmann, W. (1922). *Public opinion*. MacMillan.

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.

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.

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Merdes, C., Von Sydow, M., & Hahn, U. (2021). Formal models of source reliability. *Synthese*, *198*, 5773–5801.

Mnookin, R. H. (1992). Why negotiations fail: An exploration of barriers to the resolution of conflict. *Ohio St. J. on Disp. Resol.*, *8*, 235.

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

Nielsen. (2012). *Global trust in advertising and brand mes-*

sages. (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.

Oktar, K., Lerner, A., Malaviya, M., & Lombrozo, T. (2023). Philosophy instruction changes views on moral controversies by decreasing reliance on intuition. *Cognition*, *236*, 105434.

Oktar, K., & Lombrozo, T. (2022a). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. *Cognition*, *223*, 105021.

Oktar, K., & Lombrozo, T. (2022b). Mechanisms of belief persistence in the face of societal disagreement. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme*, *8*(2), 127–143.

O'Keefe, D. J. (2013). The relative persuasiveness of different forms of arguments-from-consequences: A review and integration. *Annals of the International Communication Association*, *36*(1), 109–135.

O'Keefe, D. J. (2018). Persuasion. In *The handbook of communication skills* (pp. 319–335). Routledge.

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. G. (2022). Beliefs about COVID-19 in Canada, the United Kingdom, and the United States: A novel test of political polarization and motivated reasoning. *Personality and Social Psychology Bulletin*, *48*(5), 750–765.

Reisinger, D., Kogler, M. L., & Jäger, G. (2023). On the interplay of gullibility, plausibility, and criticism: A computational model of epistemic vigilance. *Journal of Artificial Societies and Social Simulation*, *26*(3), 2–20.

Reynolds, J. W. (2020). Talking about abortion (listening optional). *Texas A&M Law Review*, *8*, 141.

Savage, L. J. (1954). *The foundations of statistics*. John Wiley & Sons.

Schulz, L., Schulz, E., Bhui, R., & Dayan, P. (2023, Oct). *Mechanisms of mistrust: A Bayesian account of misinformation learning.* PsyArXiv. Retrieved from `osf.io/preprints/psyarxiv/8egxh` doi: 10.31234/osf.io/8egxh

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning From Others: The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

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., Hawkins, R., Ho, M. K., Griffiths, T., & Hadfield-Menell, D. (2022). How to talk so ai will learn: Instructions, descriptions, and autonomy. *Advances in Neural Information Processing Systems*, *35*, 34762–34775.

Sumers, T. R., Ho, M. K., Griffiths, T. L., & Hawkins, R. D. (2023). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review*.

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.

Van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*, 727–763.

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

Whatley, M. A., Webster, J. M., Smith, R. H., & Rhodes, A. (1999). The effect of a favor on public and private compliance: How internalized is the norm of reciprocity? *Basic and Applied Social Psychology*, *21*(3), 251–259.