

BRIEF REPORT

Paradoxical Effects of Persuasive Messages

Rahul Bhui and Samuel J. Gershman
Harvard University



The same persuasive message can be interpreted in a positive or negative way, challenging our ability to predict its effectiveness. Here, we analyze theoretically and experimentally how causal reasoning contributes to this process of interpretation and can produce attitude reversals due to the network structure of beliefs. We conduct two vignette experiments, one based on the famous slogan of the car rental agency Avis (“We’re No. 2—that means we try harder”), and the other based on online product reviews. When participants’ contextual beliefs about the economic environment are manipulated, message effectiveness changes as predicted by a Bayesian mechanism in which seemingly negative information is “explained away” in a more positive light, or vice versa. Thus, causal reasoning may help account for certain counterintuitive kinds of high-level attitude change.


Keywords: causal reasoning, Bayesian inference, judgment, attitude change

Supplemental materials: <http://dx.doi.org/10.1037/dec0000123.supp>


Many public health or advertising campaigns are based on persuasive messages that attempt to shift the audience’s preferences. However, these campaigns may have counterintuitive effects because people can interpret the same message in different ways. Clever senders can harness the power of reinterpretation to turn superficially negative information to their benefit. For example, in 1962 the car rental agency Avis fought back against their more popular

competitor Hertz using the now-famous slogan “We’re No. 2—that means we try harder” (Eriksson & Holmgren, 1995). This reconception of second place as “underdog” rather than “loser” was considered tremendously successful, and the Avis campaign has been regarded as one of the most iconic ad campaigns of the 20th century. What makes messages like this convincing? When can seemingly unfavorable information be viewed in a more flattering light? We show that such phenomena can be naturally accounted for in a Bayesian framework. Due to the structure of causal representations, information which is overtly negative can have indirect positive implications, or vice versa.

Abundant work in social and consumer psychology has studied how content and context influence persuasion (e.g., Petty & Cacioppo, 1996). A long-standing line of research analyzes attitude change using theories of probabilistic and causal reasoning (e.g., Hahn & Oakford, 2007; McGuire, 1960a, 1960b; Wyer & Goldberg, 1970). Models of Bayesian argumentation provide a computational framework that

 Rahul Bhui, Department of Psychology and Center for Brain Science, and Department of Economics, Harvard University; Samuel J. Gershman, Department of Psychology and Center for Brain Science, Harvard University.

We thank Andrei Shleifer for many helpful discussions. This work was sponsored in part by the Harvard Mind/Brain/Behavior Initiative.

 The data and code used for analysis are available at <https://github.com/r-bhui/persuasive-messages>

Correspondence concerning this article should be addressed to Rahul Bhui, Department of Psychology and Center for Brain Science, Harvard University, 52 Oxford Street, Cambridge, MA 02138. E-mail: rbhui@g.harvard.edu

can make specific predictions about how message content and context interact nonadditively, complementing more qualitative approaches. This formalization enables greater theoretical rigor, helping to clarify the cognitive mechanisms underlying persuasion and to sharpen predictions about when messages will succeed or fail. The Bayesian approach has been shown to account for variation in judged argument strength (Hahn & Oaksford, 2006, 2007; Hahn & Hornikx, 2016; Zenker, 2013), and Bayesian networks capture more sophisticated levels of reasoning that contribute to persuasiveness (Hahn, Oaksford, Harris, 2013; Harris, Hahn, Madsen, & Hsu, 2016).

Circumstances under which seemingly negative messages can be interpreted positively (or vice versa) have been studied in various fields, such as the boomerang effect in social psychology (Hovland, Janis, & Kelley, 1953) and two-sided communication in marketing (Crowley & Hoyer, 1994). The boomerang effect refers to situations in which weak positive messages produce negative changes in beliefs, and two-sided messages are those which admit negative information to raise persuasive power. However, definitions of argument strength tend to be vague or even circular. For example, Petty and Cacioppo (1986, emphasis removed) consider a “strong message” to be “one containing arguments . . . such that when subjects are instructed to think about the message, the thoughts that they generate are predominantly favorable” (p. 133) which begs the question. The Bayesian framework provides a clearer characterization of argument strength based on probabilistic reasoning, and permits a more rigorous depiction of the cognitive computations involved.

Explanations of these phenomena based on Bayesian argumentation have lagged behind, however. Analysis of reversals has been restricted to the “faint praise effect” whereby weak positive information has a negative effect on beliefs. Harris, Corner, and Hahn (2013) explain this by a rational argument from ignorance: if the (informed) source had stronger positive evidence, they would have provided it, and thus the absence of evidence implies evidence of absence. While this mechanism is a possible source of reversals, it is distinct from our focus in this article.

Here, we consider how attitude reversals can occur due to positive or negative information being “explained away” (Pearl, 1988), a mechanism

which has not been previously identified in studies of Bayesian argumentation. Messages can indicate value-irrelevant reasons for apparently good or bad observations. This could cause a reversal if the alternative explanation is convincing enough to account for the information by Bayesian standards. We investigate this mechanism more concretely in two different situations.

First, in the Avis scenario, we examine how the seemingly negative attribute of being the second-place competitor can be interpreted positively. Normally, being first place is good because it is a sign of quality; however, this implication can be explained away when popularity may be caused by quality-irrelevant factors like advertising or customer habit. According to our novel Bayesian account, the same slogan should be more convincing in environments where the first-place company does advertise more or where being first place is not a strong sign of quality.

Second, in the setting of online product reviews, we examine how the seemingly positive attribute of a higher rating can be interpreted negatively. Higher ratings are typically a sign of better product quality, but this is not the case when the reviews are fake or biased. This account predicts that the most glowing reviews will be most heavily discounted when fake reviews are sufficiently likely to exist (because fraud will generally produce high ratings)—they become considered “too good to be true.”

We find some evidence supporting these predictions in vignette experiments based on each scenario. The impact of the message depends on the sophisticated network of audience beliefs. We thus provide novel demonstrations of how certain kinds of counterintuitive attitude change such as a famed real-world example can be understood in terms of Bayesian argumentation. This helps us to more precisely systematize our understanding of persuasion and preference reversals.

Theory

We first discuss our theoretical framework. To transparently illustrate the computational cognitive principles involved, we construct a simple model of the mental representations of message recipients. Our approach uses a Bayesian network as depicted in Figure 1, which comprises two types of components: nodes, which capture the relevant attributes in a scenario, and edges, which capture the causal rela-

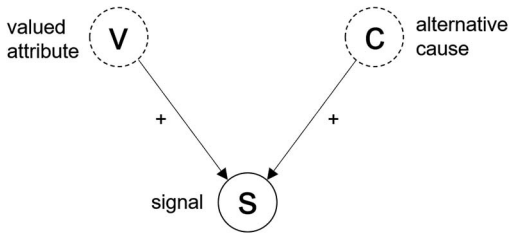


Figure 1. A belief network that can lead to attitude reversals via “explaining away”. Nodes represent relevant attributes, and arrows represent the causal relationships between them. Dashed nodes denote potentially unobserved attributes which may be inferred.

tionships that connect attributes to each other. Individuals reason about what the magnitudes of different attributes collectively imply according to the structure of their beliefs. This reasoning could be consciously accessible at high levels, although it does not need to be. Note that these models are intended to capture the representations that people hold in their minds. The belief networks may not necessarily reflect the truth, only how message recipients view the world.

In our setup, people value one of the attributes v (of uncertain magnitude) which they attempt to estimate based on other attributes, in particular, the signal s . This signal can be thought of as a piece of information that is typically considered good due to its association with value. However, it can also be influenced by the alternative cause c , and hence is not an unambiguous sign of value. Messages can alter people’s preferences by changing their perceptions of various elements in this network, directly or indirectly (via inference). When this leads people to attribute the signal more to the alternative cause, its connection to value is dampened. This mechanism attenuates the negative impact of low signals and the positive impact of high signals. As a result, even when the signal is lower, the receiver’s estimate of value may be higher if this is counteracted by the attribution of the signal to the alternative cause.

There are multiple ways to mathematically specify this mechanism, and, in what follows, we lay out two such cases (with the recognition that these are illustrative examples and not the only specifications possible).¹ In the Avis scenario, we suppose the network follows a linear

Gaussian structure, and also includes an extra edge between the valued attribute and the alternative cause (which is further assumed to be observed for simplicity). In the product review setting, we suppose the signal comes from a mixture model.

Avis Car Rental

We consider the relevant attributes in the Avis example to be the quality of a company’s product or service (q), its level of popularity (ℓ), and its level of resources (r). Quality covers all aspects of dealing with a company, from the condition of their cars to the excellence of their service. This is the valued attribute that potential customers do not observe and are trying to infer based on their belief network. Popularity refers generally to the prevalence of the company’s product or service among the market, measured along dimensions like market share or number of clients. Resources consist of money as well as equipment, experience, reputation, and other such forms of capital.

Figure 2 depicts a belief network that can account for the reversal in preferences caused by the Avis message. This network reflects three possible relationships. First, if a company provides high quality service, this will likely attract clients and make the company more popular; so there is a positive link from quality to popularity, with strength γ . Second, popularity can stem from sources that may not closely reflect quality, like marketing or existing customer habits. Well-off established incumbents are at an advantage here because, for instance, they have more capital to facilitate marketing; so there is a positive link from resources to popularity, with strength α . Third, resource-rich companies can use their capital to deliver better quality to their clients. They have more available funds or accumulated experience with which to properly maintain their equipment; so there is a positive link from resources to quality with strength β .

What might the Avis slogan of “We’re No. 2—that means we try harder” entail? It clarifies

¹ Due to the degrees of freedom that would be needed to fit the theoretical constructs from the data, we instead present simulations that mimic the qualitative features of participant judgments, and leave more direct model fits for future work.

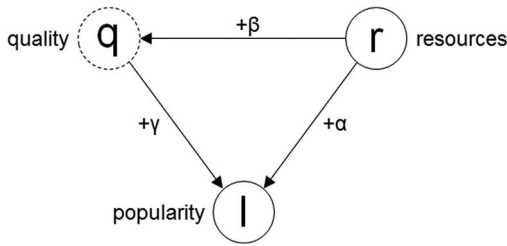


Figure 2. A belief network that can lead to the Avis phenomenon. Greek letters reflect the strengths of the causal relationships. The link from quality (q) to popularity (ℓ) reflects the direct negative implication of popularity levels. The link from resources (r) to popularity (ℓ) enables popularity to be explained away by quality-irrelevant sources. The link from resources (r) to quality (q) entails that resources can serve as a cue of value separate from popularity.

that they are indeed the smaller, second-place competitor. Customers may not have been so keenly aware of or attentive to this information before seeing the ad, and hence their perceptions of ℓ and r are downgraded. By this account, the key to Avis’s reversal lies in the link between resources and popularity. If richer companies are more able to buy popularity (i.e., if α is large), then being worse off can actually become an advantage. Second-place companies do not have as many resources with which to buy popularity; therefore any popularity that they do have must be a result of their quality. The extent to which customers find this reasoning convincing depends on the magnitudes of the relationships between variables.

Let us formalize this analysis. For transparency, we assume the network follows a linear Gaussian structure as follows:

$$q = \beta r + \varepsilon_q, \quad (1)$$

$$\ell = \alpha r + \gamma q + \varepsilon_\ell, \quad (2)$$

with the noise terms $\varepsilon_q, \varepsilon_\ell \sim \mathcal{N}(0, 1)$. Customers reason about the attributes by estimating $E(q | \ell, r)$. Rearranging Equations 1 and 2 gives us $q = (\ell - \alpha r - \varepsilon_\ell)/\gamma$ and $\varepsilon_\ell + \gamma\varepsilon_q = \ell - (\alpha + \beta\gamma)r$, and so we have that

$$E(q | \ell, r) = (\ell - \alpha r - E[\varepsilon_\ell | \ell, r])/\gamma \quad (3)$$

$$= (\ell - \alpha r - E[\varepsilon_\ell | \varepsilon_\ell + \gamma\varepsilon_q] + \gamma\varepsilon_q) / \gamma = \ell - (\alpha + \beta\gamma)r / \gamma \quad (4)$$

$$= \left(\ell - \alpha r - \frac{\ell - (\alpha + \beta\gamma)r}{\gamma^2 + 1} \right) / \gamma \quad (5)$$

$$= \frac{\gamma\ell + (\beta - \gamma\alpha)r}{\gamma^2 + 1}. \quad (6)$$

In the present account, the ad makes second place feel even less popular and poorer relative to first place. Formally, suppose the ad decreases perceptions of popularity and resources for the second-place company by magnitudes $\Delta\ell > 0$ and $\Delta r > 0$. This changes its inferred quality by

$$\Delta E(q | \ell, r) = \frac{(\gamma\alpha - \beta)\Delta r - \gamma\Delta\ell}{\gamma^2 + 1}. \quad (7)$$

The ad benefits second place if

$$\gamma\alpha\Delta r > \gamma\Delta\ell + \beta\Delta r. \quad (8)$$

All else being equal, this model predicts that the ad will be most effective when α is high, because the greater popularity of the first-place company is then clearly due to their resources rather than their quality. This is the key pathway through which “explaining away” operates. It also predicts the ad will be most effective when β is low, because then the first-place company’s greater resources do not translate into quality, which is what customers ultimately care about. This effect reflects a different Bayesian mechanism, known as *cue combination*. The γ term has an ambiguous effect that depends on the link between resources and popularity, and the degree to which perceptions of each are modified.² To test the simple predictions of this model, we conduct an experiment that varies the α and β parameters by changing the contextual information provided to participants.

Experiment

Participants. Two thousand participants from the United States were recruited from Am-

² Expression 8 also reveals a property of the ad’s beneficial effect that may not be obvious. It is modulated by the relationship between quality and popularity, and is equal to zero if this relationship is weak (i.e., if $\gamma = 0$). This occurs because resource level serves to explain away popularity—but this is pointless if low popularity does not carry negative implications needing to be explained away.

azon Mechanical Turk; each was paid \$0.85 for their participation.³ The study was approved by the Harvard Committee on the Use of Human Subjects.

Procedure. Our experiment consisted of a vignette synthesized from the Avis scenario, to tap the rich real-world thought processes of participants while maintaining experimental control. Participants read the vignette depicted in Figure 3, in which they were choosing a fictitious company to rent a car from based on information provided about the rental market. In order to focus on our mechanism of interest, several aspects of the choice problem like price, brand name, and physical appearance were held constant. Participants were then presented a message based on the “we try harder” slogan, and indicated whether the message made them “*more likely*,” “*equally likely*,” or “*less likely*” to pick the second-place company. Finally, participants were asked to write down any reasons for their answer as a free-text response.⁴ The focal context in all conditions consisted of two sentences that related to advertising inequality and maintenance difficulty. These were varied separately (with order counter-balanced) producing a control condition and two treatments. In the control condition, one sentence indicated that richer companies could advertise more, and the other indicated that maintaining a fleet of cars was inexpensive. The former was modified to create the “advertising cap” treatment, in which both companies were said to advertise about the same amount. The latter was modified to create the “difficult maintenance” treatment, in which maintaining a fleet of cars was said to be expensive.

Both of these treatments are predicted to make participants respond less favorably to the message. Intuitively, the advertising cap treatment should weaken the link between resources and popularity, corresponding in our model to a decrease in the value of α . This should reduce the audience’s ability to explain away first place with quality-irrelevant factors, and constitutes the most direct test of our hypothesis. The difficult maintenance treatment strengthens the link between resources and quality, corresponding to an increase in the value of β . This should enhance the contribution of the resource cue, increas-

ing the audience’s direct inference of quality for the more well-off company.

Results

The observed preference changes due to the message are displayed in Figure 4. Analyses exclude individuals who spent less than 30 s reading the vignette or who wrote less than 10 characters in their free response, as a rough attention check, which is standard when using the Amazon Mechanical Turk pool, leaving 1,711 participants in the sample.⁵ The proportion of participants *more likely* to pick No. 2 dropped by 10 percentage points in the advertising cap treatment and by 4.5 percentage points in the difficult maintenance treatment.

We analyze the data using a Bayesian ordered logistic regression, which assumes the discrete ordered responses (*less/equally/more likely* to pick No. 2) occur when the output of an underlying linear model falls between various cutpoints (the linear model and cutpoints are estimated). That is, for predictors X (experimental condition) and slope parameters B (intercept and effects of condition), *less likely* is reported when $BX + \epsilon$ is less than some threshold, *more likely* when it is greater than a higher threshold, and *equally likely* when it is between the lower and higher thresholds. Bayesian analyses depict uncertainty in a nuanced way without the use of arbitrary statistical cutoffs, helping to ensure that the information contained in the data is neither overstated nor understated (Amrhein, Trafimow, & Greenland, 2019; Gelman et al., 2013; Hurlbert, Levine, & Utts, 2019; McShane, Gal, Gelman, Robert, & Tackett, 2019; Wasserstein, Schirm, & Lazar, 2019).⁶

³ This sample was based on a power calculation (for detecting a difference in proportions of roughly 5 pp at 80% power) combined with budgetary constraints.

⁴ This was done on a separate page only after they selected one of the three options, and without prior warning, to avoid biasing the choice process.

⁵ The conditions had 586 (control), 563 (advertising cap), and 562 (difficult maintenance) participants. The exclusion criterion was informed by pilot data.

⁶ This regression and those below were computed with the *brms* package in R, and used the weakly informative default priors for coefficients recommended by Gelman et al. (2008), a Cauchy distribution with center 0 and scale 2.5 (though alternative assumptions do not appreciably change the results).

Imagine you're a businessperson on a trip, and in order to get around town you need to rent a car. Upon entering the airport plaza, you see kiosks set up for each of the two car rental agencies in town: Freshline and Rightway. Both have the standard model you're looking for, and charge about the same price for it. Wanting the trip to go smoothly, you consider what you know about their companies.

Control:

[As a result of town laws regarding advertising, richer companies are able to advertise more. Because of the well-paved streets and good weather, maintaining a fleet of cars is easy and relatively inexpensive.]

Advertising Cap Treatment:

[As a result of town laws regarding advertising, both companies advertise about the same amount. Because of the well-paved streets and good weather, maintaining a fleet of cars is easy and relatively inexpensive.]

Difficult Maintenance Treatment:

[As a result of town laws regarding advertising, richer companies are able to advertise more. Because of the poorly-paved streets and wet weather, maintaining a fleet of cars is hard and relatively expensive.]

Freshline is the largest agency, holding 82% of the town's market share. Rightway is second in popularity, holding 18% of business.

Think about which company would provide you the best quality.

While walking around and making your final decision, you see a poster from Rightway that says:

"We're No. 2 – That means we try harder.
We can't afford to make you wait
We can't afford to give you old cars
-Rightway"

Compared to before, does this make you feel more likely to pick Rightway, less likely to pick Rightway, or do you feel the same?

Figure 3. Text used in the experimental vignette. The bolded and bracketed text varied across conditions (this text was formatted the same as surrounding text when shown to participants). An image of a vehicle on the beach (not displayed) was also shown in all conditions to split up the first block of text from the second block and to increase participant engagement with the task.

The results are reported in Table 1, and indicate that the advertising cap treatment had a negative effect (odds ratio [OR] = 0.677, posterior 95% CI [0.542, 0.845], $P(OR < 1) > 99.9\%$). The same conclusion is obtained from model comparison of a regression that includes experimental condition as a predictor (using data from only the control and advertising cap conditions) against a null regression that ignores it, according to the one-standard-error rule (Hastie, Tibshirani, & Friedman, 2009; $\Delta WAIC$ [widely applicable information criterion] = -9.31 , $SE = 6.68$). The Bayesian analysis suggests that the diffi-

cult maintenance treatment may have also had a negative effect ($OR = 0.841$, posterior 95% CI [0.674, 1.060], $P(OR < 1) = 92.8\%$), although the analogous model comparison does not yield evidence of this ($\Delta WAIC = -0.13$, $SE = 2.86$).⁷

To see whether the treatments influenced participants' reasoning as anticipated, we hand-coded the free response data according to the explicit reasoning provided. In particular, we

⁷ We note for completeness that if all participants are included, the values of $P(OR < 1)$ are 99.6% (advertising cap) and 72.1% (difficult maintenance).

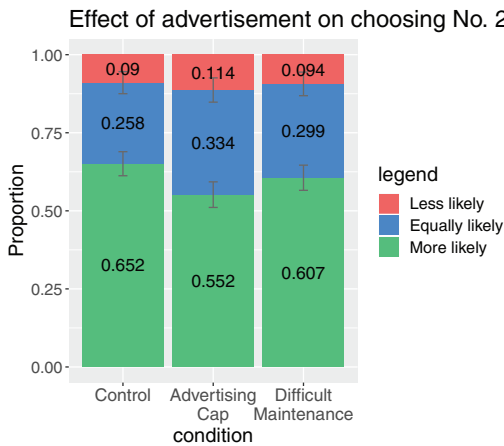


Figure 4. Preference changes due to message. Stacked bars depict proportions of participants in all response categories. Bayesian 95% credible intervals shown based on a Dirichlet-multinomial distribution with Jeffreys prior. See the online article for the color version of this figure.

counted the frequency of responses which pertained to No. 2 having less advertising (and thus having a quality-irrelevant disadvantage) or fewer resources (and thus being less able to afford high quality). Most responses were vague about the precise thought process involved—and, as noted earlier, the reasoning need not be consciously accessible. However, in a number of cases participants were quite transparent. Examples of the types of reasoning provided are depicted in Figure 5. The former reasoning (which speaks in favor of No. 2) should be least frequent in the advertising cap treatment in which the advertising difference is negated. The latter reasoning (which speaks against No. 2) should be most frequent in the difficult maintenance treatment in which the cost of maintenance is raised.

Empirical stated reasoning patterns are shown in Figure 6, and corresponding Bayesian logistic regressions are reported in Table 2. The posterior probability that “less advertising” is mentioned less in the advertising cap condition is 96.9% compared to the control condition and 99.8% compared to the difficult maintenance condition. This is recapitulated in the model comparisons of regressions that include experimental condition as a predictor against null regressions that ignore it (considering the pairwise contrasts separately: $\Delta\text{WAIC}_{\text{AC-Control}} =$

-4.74 , $SE = 2.46$; $\Delta\text{WAIC}_{\text{AC-DM}} = -8.82$, $SE = 3.29$). The posterior probability that “fewer resources” is mentioned more in the difficult maintenance condition compared to the control condition is 88.5% and compared to the advertising cap condition is 99.6%, but the analogous model comparisons do not lead to the same conclusion ($\Delta\text{WAIC}_{\text{DM-Control}} = 0.23$, $SE = 2.58$; $\Delta\text{WAIC}_{\text{DM-AC}} = -4.53$, $SE = 5.37$). Thus, stated reasoning patterns appear to reflect variation according to the explaining away mechanism, though not necessarily according to cue combination.

We note that explaining away could modulate message strength in multiple ways. For instance, persuasiveness should be affected by the perceived honesty of the source (Hahn et al., 2013), which may itself be influenced by the message. Self-disclosure of negative information can make the source seem more credible (Crowley & Hoyer, 1994; Eisend, 2006; Pechmann, 1990). While credibility from self-disclosure does not by itself explain the crucial connection between being second place and trying harder, it may modulate message strength here. This might have occurred in our experiment; the proportion of participants who cited honesty in their free response reasoning was highest in the difficult maintenance condition,⁸ where the message should be considered most negative, and this would have cut against the treatment effect. To better explore phenomena like this, in the following section we provide a fuller examination of how perceived deception can lead the audience to discount ostensibly better signals.

Online Product Reviews

We consider how higher ratings for a product could make potential customers *less* inclined to buy it. Generally speaking, good reviews with higher star ratings are supposed to reflect better underlying quality of a product. However, they could instead be fake reviews which are high

⁸ These responses were defined based on whether they contained the text string “honest” (for example, in the word “honesty”), excluding by hand responses that did not indicate the participant felt the company to be honest. The proportion in the difficult maintenance condition was 5.9% vs 3.8% in the control condition, greater with posterior probability 96.3% according to a Bayesian logistic regression of response type on condition.

Table 1
 Posterior Estimates From Bayesian Ordered Logistic Regression of Message-Induced Preference Change on Experimental Condition

Coefficient	Attitude change		
	Posterior mean	Posterior 95% CI	$P(B < 0)$
Advertising cap	-0.390	[-0.613, -0.169]	>.999
Difficult maintenance	-0.174	[-0.394, -0.059]	0.928
Cutpoint <i>less</i> <i>equally</i>	-2.406	[-2.612, -2.195]	
Cutpoint <i>equally</i> <i>more</i>	-0.612	[-0.772, -0.448]	
N	1711		
WAIC	3,056.41		

Note. $P(B < 0)$ denotes the posterior probability that the coefficient is less than 0. WAIC = widely applicable information criterion.

regardless of quality. This issue seems to be of growing concern, as retailers have increasingly come under scrutiny for enabling fake reviews to persist, and multiple websites have appeared to combat their spread (such as ReviewMeta and Fakespot). Excessively high ratings can thus make people suspicious that they are too good to be true (Gunn et al., 2016). Consequently, a four-star rating may be more convincing than a five-star rating if the latter is a strong enough indicator of fraud.

To capture this formally, as depicted in Figure 7, assume that the star rating (r) of a product depends on its true quality (q) and whether or not the reviews are fake (f), both of which are unobserved. We suppose both the ratings and quality take on integer values from 1 to 5, and the status of the reviews is binary, with $f = 0$ when they are real and $f = 1$ when they are fake. We further suppose that the ratings are generated by a mixture model, such that they are equal to the true quality when reviews are real,

but are four to five stars (most likely 5) regardless of quality when reviews are fake. For simplicity, we let the prior distribution of quality be uniform. We also leave aside the natural possibility that fake reviews tend to be generated by companies with low quality products (a negative link from q to f), but this can be incorporated into the model and would amplify potential reversals.

Under the above assumptions, two useful quantities can be computed: the probability that the reviews are fake given the rating, $P(f = 1 | r)$, and the expected quality given the rating, $E(q | r)$. This model makes key predictions regarding these quantities, visualized by the simulations in Figure 8. First, people will believe higher ratings are more likely to be fake, and especially so when the prior probability of fake reviews is high.

$$P(f = 1 | r) = \frac{P(r | f = 1)P(f = 1)}{P(r)} \quad (9)$$

No. 2 has less advertising: *"I would have to agree with the advertisement. The other car rental agency has a competitive advantage by being able to advertise more than the other. So, in order to stay competitive, the second car company DOES have to provide better service and better vehicles. Seeing that sign would put that idea into my head and I would probably choose the second car rental agency over the larger one just for that reason. I do think viewing that ad would have a positive effect on me."* ("More likely" response)

No. 2 has fewer resources: *"Being number 2 and much smaller makes me feel they may have limited resources and actually may not be able to provide me better service. I'd rather use the largest provider since they have a large amount of resources to assure me zero issues."* ("Less likely" response)

Figure 5. Examples of free-text responses providing reasons for the participant's decision.

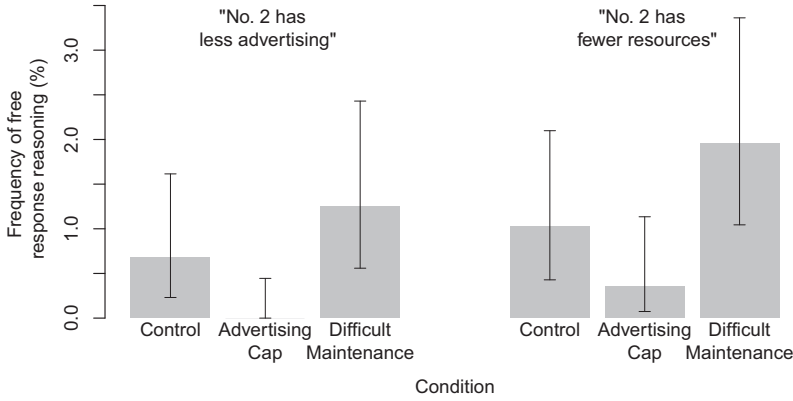


Figure 6. Proportions of participants providing various types of free response reasoning across experimental conditions. Bayesian 95% credible intervals shown based on a beta-binomial distribution with Jeffreys prior.

$$= \frac{P(r|f=1)P(f=1)}{P(r|f=1)P(f=1) + P(r|f=0)P(f=0)} \tag{10}$$

$$= \frac{P(r|f=1)P(f=1)}{P(r|f=1)P(f=1) + \sum_q P(r|f=0, q)P(q)P(f=0)} \tag{11}$$

$$= \frac{P(r|f=1)P(f=1)}{P(r|f=1)P(f=1) + \sum_q \mathbf{1}\{r=q\}P(q)P(f=0)} \tag{12}$$

$$= \frac{P(r|f=1)P(f=1)}{P(r|f=1)P(f=1) + P(q=r)P(f=0)} \tag{13}$$

This expression is increasing in r since $P(r | f = 1)$ is increasing in r . For example, letting $P(r = 5 | f = 1) = 0.8$, $P(r = 4 | f = 1) = 0.2$ and $P(f = 1) = 0.1$, a five-star rating is fake with 31% probability, a four-star rating is judged fake with 10% probability, while a three-star rating is never fake. This effect is moreover increasing in $P(f = 1)$. For example, if $P(f = 1) = 0.25$ instead, the respective ratings are judged fake with probabilities 57%, 25%, and 0%.

Table 2
Posterior Estimates From Bayesian Logistic Regression of Free Response Reasoning on Experimental Condition

Coefficient	Reasoning ("less advertising")		
	Posterior mean	Posterior 95% CI	$P(B > 0)$
Control	2.488	[-0.077, 7.713]	0.969
Difficult maintenance	3.170	[0.615, 8.356]	0.998
Intercept	-7.661	[-12.749, -5.261]	
N	1711		
WAIC	129.08		
Coefficient	Reasoning ("fewer resources")		
	Posterior mean	Posterior 95% CI	$P(B < 0)$
Control	-0.610	[-1.687, 0.406]	0.885
Advertising cap	-1.680	[-3.434, -0.381]	0.996
Intercept	-3.999	[-4.670, -3.435]	
N	1711		
WAIC	208.00		

Note. WAIC = widely applicable information criterion; CI = confidence interval.

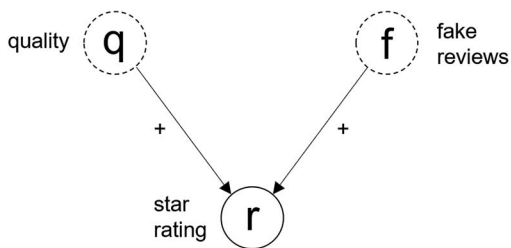


Figure 7. A belief network that can lead to attitude reversals with product reviews.

Second, the positive effect of higher ratings on expected quality will be reduced when the prior probability of fake reviews is high. Intuitively, a five-star rating could occur either because the true quality is five or because the reviews are fake (in which case the true quality is likely lower). Accounting for the latter possibility dampens the impact of the signal.

$$P(q|r) = \frac{P(r|q)P(q)}{P(r)} \quad (14)$$

$$= \frac{[P(r|q, f=0)P(f=0) + P(r|q, f=1)P(f=1)]P(q)}{\sum_q [P(r|q, f=0)P(f=0) + P(r|q, f=1)P(f=1)]P(q)} \quad (15)$$

$$= \frac{\mathbf{1}\{r=q\}P(f=0) + P(r|f=1)P(f=1)}{P(f=0) + 5P(r|f=1)P(f=1)} \quad (16)$$

$$E(q|r) = \sum_q qP(q|r) \quad (17)$$

$$= \frac{rP(f=0) + 15P(r|f=1)P(f=1)}{P(f=0) + 5P(r|f=1)P(f=1)} \quad (18)$$

Expected quality is thus roughly a combination of r (when $P(f=0) = 1$) and the prior mean of 3 (when $P(f=1) = 1$). The fundamental tension arises because when r increases, so does $P(r|f=1)$, pulling the quality estimate back toward its prior mean. The balance between these forces determines whether rating has a positive or negative effect on net, and is modulated by $P(f=1)$.

Third, these two effects should occur in tandem. In the present account, the attenuation of value directly results from the increased attribution of ratings to fraud. Expressions 13 and 18 reveal that both changes are modulated by the term $P(f=1)$, the prior probability that reviews are fake. In the following experiment, we test the main predictions of the theory by varying this probability.

Experiment

Participants. Nine hundred and six participants from the United States were recruited from Amazon Mechanical Turk using TurkPrime (Litman, Robinson, & Abberbock, 2017); each was paid \$0.25 for their participation. This sample size was determined by a rule recommended by Kruschke (2018), based on when the 95% highest density intervals (HDI)

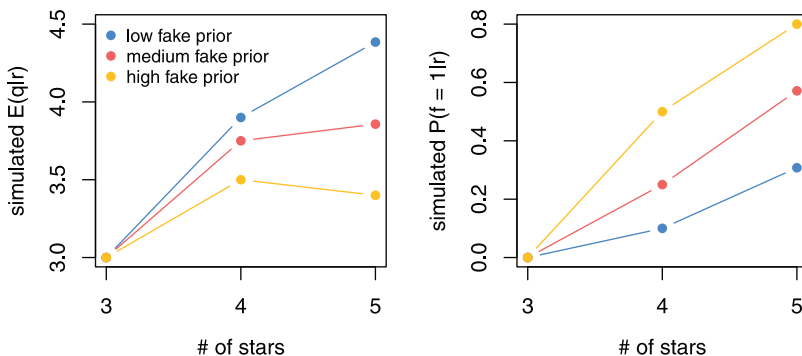


Figure 8. Model simulation of inferences from review star ratings, with priors of fake reviews set as 0.1 (low), 0.25 (medium), and 0.5 (high), and $P(r=5|f=1) = 1 - P(r=4|f=1) = 0.8$. See the online article for the color version of this figure.

for parameters of interest (described below) are entirely contained inside or outside a region of practical equivalence (ROPE). Subjects were recruited in batches of 100 until either the HDI + ROPE criterion was met, or 900 participants were reached (the latter occurred first, with minor discrepancies due to technical limitations). The study was approved by the Harvard Committee on the Use of Human Subjects.

Procedure. In this experiment, participants read the following text in which they were asked to imagine buying a product (the example review was taken from an actual review of earbuds on a retail website):

Suppose you are browsing a large online retailer, looking for a pair of earbuds. You know the average rating for earbuds sold on this website is 3 stars. You see 20 reviews for an “Extra Bass Earbud Headset”, some of which say things like “#/5 stars: Really good, I like it very much.”

Participants were subsequently asked about whether they would be more or less likely to buy the product depending on whether these reviews mostly gave the product three, four, or five stars (every participant gave a response for each star level). They described this attitude change on a scale from -100 to $+100$, with positive numbers meaning more likely and negative numbers meaning less likely, and big numbers reflecting how strongly they felt. They were also asked how likely they thought the reviews were fake or biased, depending on the star ratings of the reviews, responding on a scale from 0 to 100 to reflect the likelihood.

We used a within-subjects design in which all participants were then told, “[s]uppose that the

manufacturer of these earbuds was the retailer running the website.” This information was meant to increase their prior belief of fake reviews, $P(f = 1)$, by conveying that the retailer would be more willing and able to tolerate deceit. In light of the information, they were asked the same questions regarding their likelihood of buying the product and their beliefs about whether the reviews were fake, depending on star rating. Thus, each participant provided six attitude change responses and six fake belief responses in total.

Based on the theory, we hypothesized that:

1. The effect of star rating on attitude change would be lowered when the retailer was the manufacturer.
2. The probability that the reviews were fake would be increasing in star rating especially when the retailer was the manufacturer.
3. The magnitudes of these effects would be positively correlated with each other (i.e., their values would be negatively correlated).

Results

The effects of star rating and manufacturer identity on attitude change and fake review beliefs are depicted in Figure 9. The results exclude people who provided any response that was not a number within the permissible bounds (-100 to $+100$ for attitude change, 0 to 100 for beliefs), leaving 852 participants. We statistically tested the three key predictions using a Bayesian multilevel regression (including max-

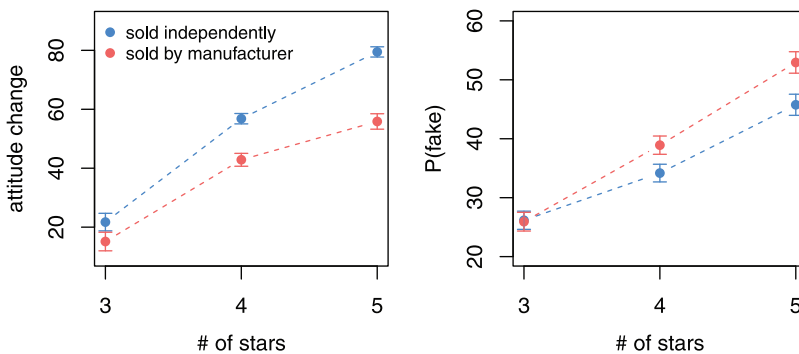


Figure 9. Effect of star rating and manufacturer identity on attitude change and beliefs about fake reviews. Bayesian 95% credible intervals shown based on Cauchy (0, 10) prior. See the online article for the color version of this figure.

imal random effects structure at the participant level). The two dependent variables of interest were the degree to which participants were more or less likely to buy the product, and their beliefs about whether the reviews were fake. The independent variables were the star ratings of the reviews (with baseline set at 3 stars) and the identity of the manufacturer, as well as their interaction.⁹

The results are shown in Table 3. In line with Prediction 1, higher star ratings do not boost one's preference as strongly when the retailer is the manufacturer, as revealed by the negative interaction term in the attitude change regression. Consistent with Prediction 2, higher star ratings particularly increase the perceived chances of the reviews being fake in the same scenario, as revealed by the positive interaction term in the belief regression. Moreover, as Prediction 3 implies, these effect magnitudes are positively correlated across individuals, meaning the interaction coefficient values are negatively correlated ($r = -0.235$, 95% Bayesian¹⁰ credible interval $[-0.301, -0.170]$). This last observation can be seen in Figure 10 which displays the results among participant subgroups defined based on the magnitude of the treatment effect on attitude change (in the 5-star rating case); those who exhibit the steepest drop in attitude change when informed that the retailer is the manufacturer also exhibit the strongest increase in their beliefs that the reviews are fake.

In some cases higher star ratings even have negative effects on attitude change, and this occurs for 142 participants (16.7% of the sample). In the Appendix, we split participants up into different subgroups based on the exact pattern of this effect, and show how such data can be accommodated by the model or an extension that incorporates a distaste for fraudulent reviews.

Discussion

In this article, we analyze the effects of persuasive messages using Bayesian networks to model the structure of mental representations, in order to help understand how the same attribute can be considered good or bad. The Bayesian framework captures the ability of individuals to reason about the value implications of causal relationships between attributes. As a result,

messages that alter perceptions of attributes can have sophisticated indirect implications that may produce counterintuitive attitude changes. We focus on how apparently negative information could be *explained away* in a more flattering light, and vice versa.

We conducted two vignette experiments to test the implications of this theorized mechanism. The first was styled after Avis Car Rental and their famous advertisement proclaiming “We’re No. 2—that means we try harder.” Consistent with our focal mechanism, the positive effect of the Avis message was reduced when high popularity was more difficult to explain away by company resources (due to a cap on advertising). Stated reasoning patterns followed suit, as participants appeared least likely to cite the gap in advertising as an explanatory factor in this condition (where there was no gap). However, there was little support for another mechanism, cue combination, which would have weakened the message when resources contributed more strongly to quality (due to high vehicle maintenance costs), and this was accompanied by limited evidence that participants were most likely to cite the resource difference in that condition (where the importance of the differential was enhanced).

The second experiment was styled after the setting of online product reviews. In line with Bayesian principles, the positive effect of higher ratings on attitude was diminished when the online retailer was also the product’s manufacturer (and hence more willing and able to facilitate fake reviews). This attitude change was linked to stated beliefs; discounting of particularly high ratings was coupled (across participants) with an enhanced belief that the highest reviews were fake or biased.

We do not claim that probabilistic reasoning explains all possible elements of attitude change (or that our specific account is the only one at

⁹ The sampling rule was based on the HDIs in both regressions for the coefficients on manufacturer identity and its interaction with star rating. We conservatively used a ROPE spanning -1 to $+1$ and would have stopped only when the HDIs for all of these coefficients fell into the ROPE, absent the preset limit.

¹⁰ This was computed using the *BayesFactor* package in R, assuming noninformative priors for the means and variances of the two populations and a shifted, scaled Beta(3,3) prior for the correlation magnitude, though alternative assumptions do not change the result.

Table 3
Posterior Estimates From Bayesian Multilevel Regressions of Attitude Change (Top) and Fake Review Belief (Bottom)

Coefficient	Attitude change		
	Posterior mean	Posterior 95% CI	$P(B < 0)$
Intercept	23.77	[20.93, 26.64]	<.001
Star rating	28.87 (0.79)	[27.24, 30.44]	<.001
Retailer/Manufacturer	-6.22 (-0.17)	[-8.16, -4.34]	>.999
Star Rating \times Retailer/Manufacturer	-8.49 (-0.23)	[-9.85, -7.12]	>.999
<i>N</i>	852		
Coefficient	<i>P</i> (fake)		
	Posterior mean	Posterior 95% CI	$P(B < 0)$
Intercept	25.55	[24.03, 27.15]	<.001
Star rating	9.79 (0.40)	[8.86, 10.74]	<.001
Retailer/Manufacturer	0.20 (0.01)	[-1.05, 1.49]	0.377
Star Rating \times Retailer/Manufacturer	3.71 (0.15)	[2.80, 4.61]	<.001
<i>N</i>	852		

Note. Standardized regression coefficients in parentheses describe effect size in units of the dependent variable’s standard deviation. CI = confidence interval.

play), merely that it helps to elucidate more precisely some of the key mechanisms involved. Petty and Cacioppo (1996) identify seven theoretical approaches to understanding attitude change. These approaches either do not naturally account for our results, or else overlap

with the Bayesian approach to the extent that they may apply.

1. According to theories of conditioning, attitudes toward a cue are more favorable when the cue has been repeatedly paired

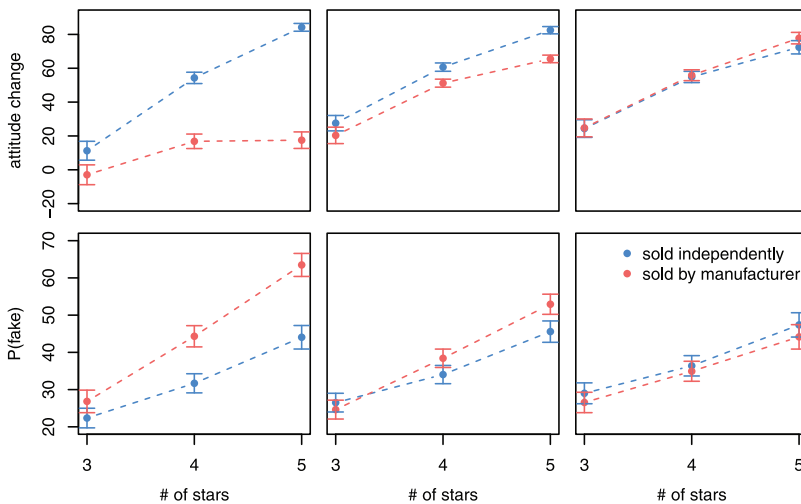


Figure 10. Results based on splitting subjects into three groups based on the magnitude of the treatment effect in terms of attitude change (the gap between red and blue points in the top row) when star rating is five. Left = highest third, middle = middle third, right = lowest third. Bayesian 95% credible intervals shown based on Cauchy (0, 10) prior. See the online article for the color version of this figure.

- with something positive (e.g., Staats & Staats, 1958). However, the experimental messages are identical across treatments, and there is negligible crude contextual variation to trigger different associations.
2. The message-learning approach asserts that properties of the source, the message, the recipient, and the channel influence the degree to which the message is comprehended and retained (e.g., Hovland et al., 1953). Through this lens, the only variation in our experiments is in the source, and the notions of credibility and trustworthiness that modulate its influence are transparently represented in the Bayesian framework.
 3. Judgmental theories entail that items which fall below some reference point are generally evaluated less favorably (e.g., Helson, 1964). In our first experiment, the advertising cap treatment does not obviously change expectations of quality, while the difficult maintenance treatment might decrease the expectation of quality which would actually benefit second place. It is also not clear how reference dependence would produce the intricate pattern of results observed in our second experiment.
 4. Motivational approaches argue that people attempt to maintain congruency in their network of associated attitudes (e.g., Osgood & Tannenbaum, 1955). Bayesian models are not only able to naturally encode this congruency via the rational combination of cues (Gershman, 2019), but also account for the effects of variation in prior beliefs as in our second experiment.
 5. Attributional theories posit that attitude changes are based on inferences that people draw about the behavior of others and themselves (e.g., Kelley, 1973). Such inference is typically referred to as causal, and its properties are readily cast in Bayesian terms (Ajzen & Fishbein, 1975).
 6. Theories of self-persuasion focus on the role of internally generated information (e.g., Petty & Cacioppo, 1979). Yet there is no obvious reason why any of our treatments would influence the degree of self-involvement without ap-

pealing to variation in external information.

7. In combinatory approaches, evaluation is described as the integration (such as the averaging) of different pieces of information available about a stimulus (e.g., Anderson, 1971). Such theories can be formally derived from Bayesian foundations, which also account for factors that change the weighting of information, such as variation in prior beliefs as in our second experiment.

The Bayesian cognitive approach has proven successful in a variety of domains (e.g., Chater, Tenenbaum, & Yuille, 2006; Oaksford & Chater, 2007), and our work builds on this tradition as applied to persuasion (Hahn & Oaksford, 2007). To our knowledge, the only explicit analysis of reversals in Bayesian argumentation relates to the faint praise effect, in which weak positive information impacts beliefs negatively. Harris et al. (2013) use the concept of epistemic closure to account for this effect, arguing that stronger positive information would have been provided were it available, and therefore its absence implies its nonexistence. However, the faint praise effect does not suggest, for instance, why there is any connection between being second place and trying harder. Thus, we add to the catalog of mechanisms that can produce reversals in Bayesian argumentation.

Our approach bears a connection to Bayesian models of belief polarization, in which the same evidence can lead people with opposing priors to strengthen their beliefs (Cook & Lewandowsky, 2016; Jern et al., 2009, 2014). While this literature focuses on the general phenomenon of polarization, belief divergence necessarily implies the possibility of reversals. Indeed, we do observe some participants in our experiments responding to information positively while others respond to the exact same information negatively, and even the same individuals can sometimes be induced to respond in both ways by shifting their prior. Digging into the Bayesian mechanisms that can give rise to polarization may thus shed light on the computational cognitive processes that describe attitude reversals more broadly, and vice versa.

Some research formally studies the interpretation of messages using game theory to capture strategic reasoning (Benz, Jäger, Van Rooij, & Van Rooij, 2005; Parikh, 2010).¹¹ Interaction between agents with various preferences generates statistical regularities that contribute to interpretation. In this way, natural conversational inferences can be derived as the result of equilibria in strategic games between senders and receivers. While valuable, such analysis remains formally predicated on Bayesian reasoning of the sort we consider, since the receiver must interpret the message in light of the sender's position and other context clues. Our work thus clarifies certain mechanisms for reversals that could illuminate corresponding facets of strategic interaction.

We build on the Bayesian framework here to plainly clarify the mechanisms underlying judgment within a unified perspective. Although many core implications of our work do not rest on the assumption of exact Bayesian inference, it would be useful for further research to scrutinize the precise quantitative match between theory and data. Participants in our experiments were not provided with explicit probabilities, which reflects a more naturalistic decision setting but constrains the ability to conduct such an investigation. Other paradigms may be needed to explore a different region in this design space.

Finally, although we describe messages as providing tangible information, it is possible that they also draw attention to certain relationships between attributes. Even if no external evidence is provided, drawing attention to neglected associations can induce a reinterpretation of attributes. In our experiment, some participants stated that the Avis message made them think about the situation in a new way (e.g., "I didn't think at first about the idea that Rightway has to try harder to maintain business, but it makes sense since the other company has almost a monopoly on the market for rental vehicles"). This suggests that the message may not simply provide information, but could actually increase the salience of previously unrecognized implications, causing people to interpret attributes in a different light. Such a mechanism seems psychologically plausible and is compatible with the formal analysis laid out in this article.

Constraints on Generality

We expect our results to generalize to the naturalistic domains described in the vignettes, provided people hold the relevant causal schema and are engaged enough for it to be retrieved. Though the experimental questions were hypothetical, we would expect incentives to sharpen engagement and strengthen the effects. Indeed, customers currently ordering products online seem quite sensitive to the possibility of fake reviews, as evidenced by the development of ways to detect fraud (such as ReviewMeta and Fakespot) and signal trustworthiness (such as Verified Purchaser labels). However, the path from attitude to behavior is not so direct (the "attitude-behavior" gap). The influence of a persuasive slogan on car rentals or of manufacturer identity on product purchases would be partially obscured by other factors such as variation in prices, extra features, or recommendations from friends. Thus, when testing the behavioral implications of attitudinal theories, care must be taken to connect stated intentions to ultimate decisions (Morrison, 1979; Morwitz, Steckel, & Gupta, 2007).

Conclusion

Persuasion is challenging because the same message can be interpreted in different ways depending on one's viewpoint. This variation may lead to counterintuitive preference changes, as seemingly negative information can be interpreted positively, or vice versa. We analyze the multilayered impacts of messages with a Bayesian model of the mental organization of attributes. We view people as reasoning causally about attribute values, and show that this structure can help explain how convincing people find messages including a renowned real-world advertising slogan. In two vignette experiments, we predictably modulate message effectiveness by altering participants' belief networks. Thus, our understanding of high-level attitude change may be enriched by formal descriptions of Bayesian reasoning.

¹¹ This includes Mullainathan, Schwartzstein, and Shleifer (2008) who also study the Avis phenomenon but assume that mental associations are naïve rather than causal. However, their model does not naturally connect persuasiveness with beliefs adjacent to a message's main claim, and so does not obviously account for our treatment effects, in addition to having trouble systematically generating backfires.

References

- Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*, 261–277.
- Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*, 262–270.
- Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review*, *78*, 171–206.
- Benz, A., Jäger, G., Van Rooij, R., & Van Rooij, R. (2005). *Game theory and pragmatics*. New York, NY: Springer.
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287–291.
- Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*, 160–179.
- Crowley, A. E., & Hoyer, W. D. (1994). An integrative framework for understanding two-sided persuasion. *Journal of Consumer Research*, *20*, 561–574.
- Eisend, M. (2006). Two-sided advertising: A meta-analysis. *International Journal of Research in Marketing*, *23*, 187–198.
- Eriksson, P., & Holmgren, H. (1995). *A book about the classic Avis advertising campaign of the 60s*. Dakini Books.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*, 1360–1383.
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.
- Gershman, S. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*, 13–28.
- Gunn, L. J., Chapeau-Blondeau, F., McDonnell, M. D., Davis, B. R., Allison, A., & Abbott, D. (2016). Too good to be true: When overwhelming evidence fails to convince. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *472*, 20150748.
- Hahn, U., Oaksford, M., & Harris, A. J. L. (2013). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 15–38). New York, NY: Springer.
- Hahn, U., & Hornikx, J. (2016). A normative framework for argument quality: Argumentation schemes with a Bayesian foundation. *Synthese*, *193*, 1833–1873.
- Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese*, *152*, 207–236.
- Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*, 704–732.
- Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*, 414–429.
- Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*, *40*, 1496–1533.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer Science & Business Media.
- Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. New York, NY: Harper and Row.
- Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change*. New Haven, CT: Yale University Press.
- Hurlbert, S. H., Levine, R. A., & Utts, J. (2019). Coup de grâce for a tough old bull: “Statistically significant” expires. *The American Statistician*, *73*, 352–357.
- Jern, A., Chang, K.-M., & Kemp, C. (2009). Bayesian belief polarization. In *Advances in neural information processing systems* (pp. 853–861).
- Jern, A., Chang, K.-M., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*, 206–224.
- Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107–128.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*, 270–280.
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433–442.
- McGuire, W. J. (1960a). Cognitive consistency and attitude change. *Journal of Abnormal and Social Psychology*, *60*, 345–353.
- McGuire, W. J. (1960b). Direct and indirect persuasive effects of dissonance-producing messages. *Journal of Abnormal and Social Psychology*, *60*, 354–358.
- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245.
- Morrison, D. G. (1979). Purchase intentions and purchase behavior. *Journal of Marketing*, *43*, 65–74.

- Morwitz, V. G., Steckel, J. H., & Gupta, A. (2007). When do purchase intentions predict sales? *International Journal of Forecasting*, 23, 347–364.
- Mullainathan, S., Schwartzstein, J., & Shleifer, A. (2008). Coarse thinking and persuasion. *Quarterly Journal of Economics*, 123, 577–619.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York, NY: Oxford University Press.
- Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review*, 62, 42–55.
- Parikh, P. (2010). *Language and equilibrium*. Cambridge, MA: MIT Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Columbus, OH: Morgan Kaufmann.
- Pechmann, C. (1990). How do consumer inferences moderate the effectiveness of two-sided messages. *NA - Advances in Consumer Research*, 17, 337–341.
- Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, 37, 1915–1926.
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Cambridge, MA: Academic Press.
- Petty, R. E., & Cacioppo, J. T. (1996). *Attitudes and persuasion: Classic and contemporary approaches*. Boulder, CO: Westview Press.
- Staats, A. W., & Staats, C. K. (1958). Attitudes established by classical conditioning. *Journal of Abnormal and Social Psychology*, 57, 37–40.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ $p < 0.05$.” *The American Statistician*, 73, 1–19.
- Wyer, R. S., Jr., & Goldberg, L. (1970). A probabilistic analysis of the relationships among belief and attitudes. *Psychological Review*, 77, 100–120.
- Zenker, F. (2013). Bayesian argumentation: The practical side of probability. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 1–11). New York, NY: Springer.

Appendix

Simulation of Subgroup Patterns

Here, we split participants up into groups based on the pattern of how they respond (in terms of attitude change) to different star ratings when the retailer is the manufacturer. Four patterns are possible, as the response can (A) be increasing in rating ($3 < 4 < 5$; 710 participants), (B) have an inverted-U pattern ($3 < 4 > 5$; 61 participants), (C) be decreasing in rating ($3 > 4 > 5$; 54 participants), or (D) have a U pattern ($3 > 4 < 5$; 27 participants). In Figure A1, we plot the average responses for participants in each of these four groups. This reveals cases in which higher ratings have both positive and negative effects depending on the condition. Figure A2

displays corresponding hand-tuned simulations (with parameters in Table A1) from an extension of the model in which value is equal to the expected quality minus a penalty based on how likely the reviews are to be fake, $E(q | r) - \delta P(f = 1 | r)$, where δ is a parameter capturing the distaste for fraud. Only the data in Group D is fundamentally at odds with the theory and unable to be captured naturally, as the downward trend going from three stars to four stars is not mirrored by an increase in $P(\text{fake})$, and the possible difference in attitude change between conditions is not recapitulated in beliefs. Encouragingly, this is the smallest group of the four.

(Appendix follows)

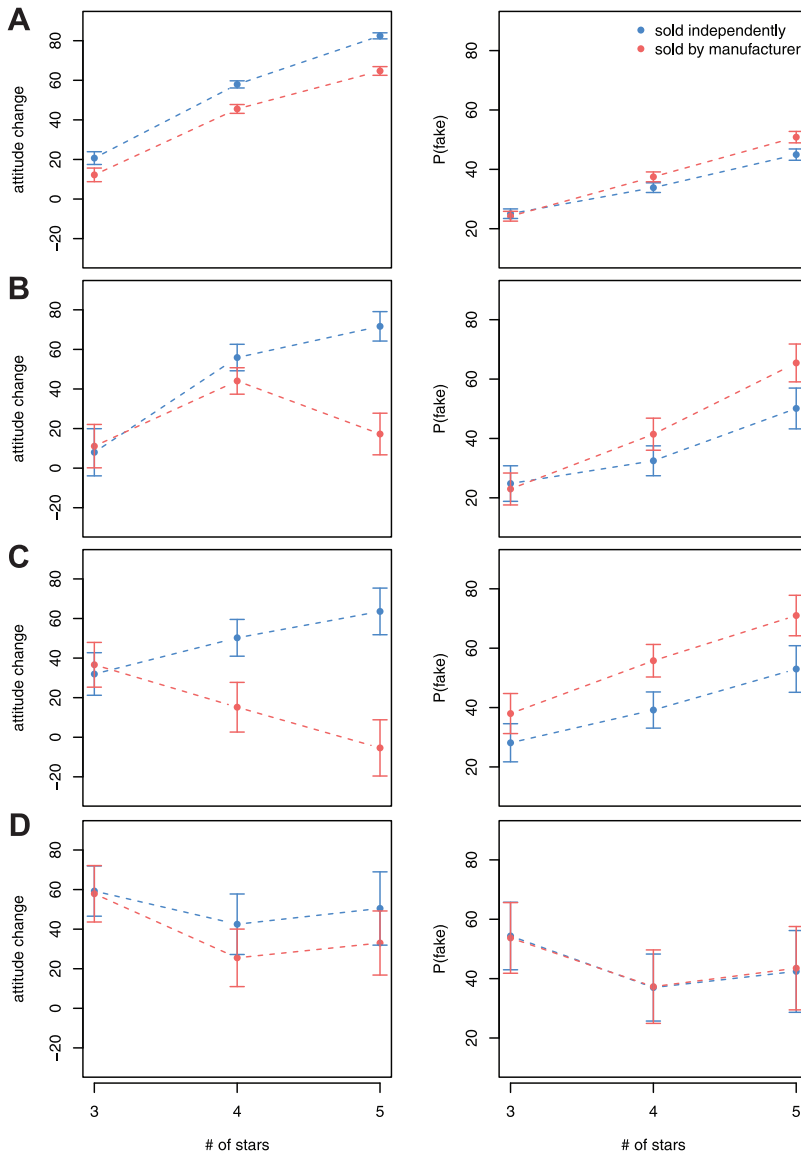


Figure A1. Effect of star rating and manufacturer identity on attitude change and beliefs about fake reviews, with participants split by pattern of attitude change. Bayesian 95% credible intervals shown based on Cauchy (0, 10) prior. See the online article for the color version of this figure.

(Appendix continues)

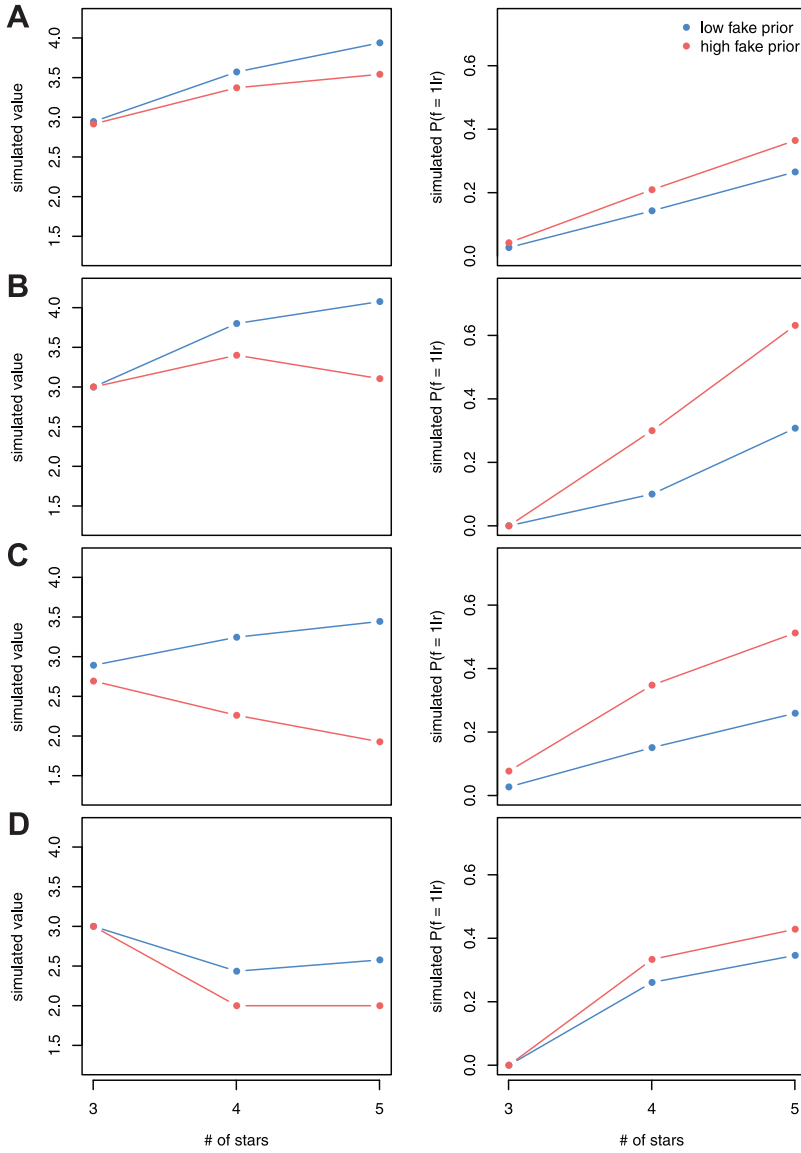


Figure A2. Simulated effect of star rating and prior on attitude change and beliefs about fake reviews, split by pattern of attitude change. Parameters provided in Table A1. See the online article for the color version of this figure.

(Appendix continues)

Table A1
Parameter Values From Simulations in Figure A2

Pattern	Simulation parameters					δ
	$P(r = 3 f = 1)$	$P(r = 4 f = 1)$	$P(r = 5 f = 1)$	Low $P(f = 1)$	High $P(f = 1)$	
A	0.05	0.30	0.65	0.10	0.15	2
B	0.00	0.20	0.80	0.10	0.30	1
C	0.05	0.32	0.63	0.10	0.25	4
D	0.00	0.40	0.60	0.15	0.20	5

Received February 12, 2019
Revision received March 11, 2020
Accepted March 12, 2020 ■