Confirmation Bias: The Role of Messages and Messengers.

By Hongyu Yao[*]

This paper explores how messages from different messengers affect individuals' beliefs and alter their perspectives about the credibility of the messenger. Employing a difference-in-differences strategy with first-hand survey data from a case study about the Covid-19 vaccines, I document that individuals update their belief based primarily on their ex-ante knowledge—while their trust in the messengers has minimal effects—which hints at a great degree of confirmation bias. And individuals use the message as a tool to re-judge their perspectives about source credibility—the credibility score of a reliable messenger can be lowered by as much as 56.3% when people face a disagreeable message. Interestingly, I discover a mechanism of credibility adjustment: a 0.25 point increase for every unit difference between the degree they trust in the messenger and the degree they agree with the message, measured all on a 100-point scale. In addition, there is evidence that the Republicans witness more confirmation bias and relatively more gullible than the Democrats.

JEL: D91, D83
Keywords: Confirmation Bias, Message, Messenger, Covid-19, Political Party

I. Introduction

People's pre-existing knowledge may unintentionally affect their choices, which is a phenomenon called confirmation bias. While most literature focuses on how a piece of message about an incident aligns with an individual's prior knowledge and, as a result, changes the individual's feeling about that incident, I investigate how the messenger, the source where the messages come from, plays a role in the process.

I realize the theory by connecting my experimental design with the relative controversial discussion about the Covid-19 vaccines. Over the year, about half a billion vaccine doses have been administered in the United States, but there are still around 35% of the people not fully vaccinated, and three-fourth of them have not started their first shocks. This fact hints at a great disparity about

[*] Yao: University of California—Berkeley, California, United States, *hongyu_yao@berkeley.edu*.
Acknowledgements: This work would not be possible without the support from Dan Acland: University of California—Berkeley, California, United States, *acland@berkeley.edu*. This project is supported by the Berkeley Experimental Social Science Laboratory (Xlab) Grant #20150-24376-44-NX199 and the Michael and Chris Boskin Scholarship from the Berkeley Department of Economics.

the effectiveness of the vaccines: with all kinds of messages supporting or resisting the vaccines, people's beliefs about vaccines wave. I pick two Fox News hosts: Tucker Carlson and Sean Hannity, who are both supporters of Trump but hold antithetical opinions towards vaccines. Consistent with many of Trump's comments, Carlson initially distrusts the vaccines[1]. On the other hand, Hannity consistently asks people to believe in science and receive vaccinations. Together with a random sample from Amazon Mechanical Turk answering my Qualtrics survey, the experimental setting allows me to explore informative questions like: if an individual receives a piece of disagreeable message from a messenger thought to be trustworthy, will the individual changes opinion to be closer to the message or the credibility of the messenger will be eroded?

I employ a difference-in-differences strategy with two dichotomized dummy variables—people's perspectives about the credibility of the messenger before the treatment and their agreeableness towards the message from that messenger—to explore the effects on people's beliefs about vaccine effectiveness and how the perspectives on messenger credibility are changed. I find strong evidence that people tend to judge a new piece of message and update their belief based on their *ex ante* knowledge: regardless of source credibility, if they agree with the message, their perspectives will change following the direction of the message; if they disagree with the message, their perspectives will change following the opposite direction of the message. In either situations, people tend to reaffirm that their pre-existing beliefs are correct. Also, they may simply disregard a disagreeable message or a message from a unreliable messenger, making the changes in their beliefs insignificant and still hold their origin perspectives. All these results hint at the existence of confirmation bias.

Further, I find that how people update their perspectives about messenger credibility is more impressive. Beyond the combinations of Agree/Disagree and Trust/Distrust, there is an underlying process comparing the degree people agree with the message and the degree people trust in the messenger before reading the message. Including a third dimension—whether the agreeableness score is higher than the pre-treatment credibility score—into the specifications, I record that for each point difference between the agreeableness and the pre-treatment credibility (all measured on a 100-point scale), the subjects' perspectives about the credibility of the messenger will increase by 0.25 point on average. In a similar way, the credibility will become lower if the agreeableness doesn't as high as the pre-treatment credibility. As for the heterogeneity, I find evidence that the Republicans experience a higher degree of confirmation bias by reaffirming their prior beliefs. In addition, the tendency to trust/distrust a messenger is more prominent for the Republicans then for the Democrats, which is robust to demographic and behavioral control variables.

In all, a piece of message coming in becomes mainly a tool for people to judge

[1] The hosts' beliefs about the vaccines may change over time, so the information discussed in this study may not represent the latest beliefs of them.

how reliable the messenger is: thinking and learning from the message to renew their own minds seems to be secondary, or even negligible. People are reluctant to change the minds they have already formed and whatever message, either agreeable or disagreeable, tends to consolidate their original beliefs. This psychological bias has huge implications on public policies: for instance the effectiveness of using "nudges", which is an effective and increasingly-common government policy (DellaVigna and Linos, 2020). And possibly, this is a miniature of how different ideologies—Democratic and Republican, Capitalism and Socialism, etc. —form and growth through self-affirmation over time. This hints at the importance of correct guidance at early stages and the potential necessity for the government to intervene at a certain point.

This work contributes to the understanding of confirmation bias by expanding the two-dimension model with message (information) and messenger (source). When an individual receives a new piece of message, both the content of the message and the source of the message are likely to be the determinants of how strong this message will affect the individual's perspective. Not only that, the individual's perspective also contains two parts: the perspective toward what's discussed in the message and the perspective toward the messenger (source of the message). How these two aspects will be affected is rarely discussed in the current literature. Also, I connect this psychology theory with two specific messages and messengers regarding to the Covid-19 vaccines. Together with an investigation of the role of people's political affiliation, my investigation about the underlying process reveals the novel idea of the Marginal Propensity to Trust. To the best of my knowledge, this is the first research expanding confirmation bias with a new dimension of high relevance, connecting it to a case study of the Covid-19 pandemic, and discovering a mechanism of how people update on source credibility, which will give meaningful implications.

The remainder of this paper proceeds as follows. Section II describes the related literature. In Section III, I describe the experimental design, data, and empirical strategy. Section IV reports the main results and discusses the implication of the results. Section V concludes.

II. Literature Review

The current literature of confirmation bias can be categorized into two groups. One group analyzes the reactions when people are given a new piece of information for which they have no prior experiences. Popper (2005) generalizes a mistake in treating these new information as a tendency to confirm the information received rather than falsify it, which is supported by Wason (1960)'s "2-4-6" experiment and the latter "A, D, 4, 7" selection task (Wason, 1968) that showed people to be illogical and irrational. In these studies, the subjects are asked to guess a rule the experimenter had in mind: surprisingly, the subjects not only formed hypotheses that were more specific than necessary, but also only test positive examples of their hypothesis. Similar phenomenon is shown in other experiments as well: when people are asked to make inquiries about a new hypothesis, they tend to formulate questions that will be answered yes under the situation where the given hypothesis is true (Shaklee and Fischhoff, 1982). Oswald and Grosjean described confirmation bias as “an immunity of the hypothesis”—the possibility to reject the hypothesis is largely reduced—in Pohl (2012)’s handbook.

Another group focuses on the relationship between a piece of new information and people’s current belief, assuming they have one, and their responses conditional on that relationship. Very early philosopher Francis Bacon (1620/1939) proposes that “humans, following the inevitable tendency, draw all things else to agree with the opinion they have adopted”. Subsequent researchers find similar results and conclude that people tend to seek information that is supportive to their existing beliefs and avoid counter-indicative ones, based on their own discretion (Koriat et al., 1980). As a result, confirmation bias makes people focus mainly on one-side of the problem, which precludes new discoveries (Bruner et al., 2017). People’s expectations are constricted inside self-fulling prophecies (Merton, 1957) and our sight is limited to see exclusively what we are looking for (Kelley, 1950). Nickerson (1998) concludes this kind of biased correlation as illusory correlation.

Further research adds more branches to this topic. Jones and Sugden (2001) build on Wason (1968)’s “A, D, 4, 7” selection task and measure positive confirmation bias both in information selection and usage. With cost and benefit, this becomes an anomaly of the standard model of decision-making. Also, Lehner et al. (2008) examines confirmation bias in a more complex setting to test the effectiveness of the Analysis of Competing Hypotheses (ACH) method, a technique developed by the Central Intelligence Agency (CIA), in minimizing the psychological bias. They also propose that current beliefs did not influence the assessment of whether an evidence is confirming or disconfirming. Knobloch-Westerwick et al. (2015) and Westerwick et al. (2017) use a broader term called selective exposure to describe the situation when individuals selectively pay more attention to some of the available messages instead of paying equal attention to all of them. The Elaboration Likelihood Model—which categorizes the information processing into central and peripheral—is employed by Petty et al. (2009). What determines the processing model is the person’s motivation (personal relevance) and ability (pre-existing knowledge). They use a combination of pro/contra message and slanted/unbiased source to measure the selection rate and the time of exposure as the estimates for selection exposure based on Clay et al. (2013).

III. Data and Methodology

A. Experimental Setting

I would like to know how people’s perspectives toward the Covid-19 vaccines and toward the messengers talking about the vaccines change after reading the message. To achieve this goal, I designed an online survey in Qualtrics.

After getting the consent from the subjects, I ask the subjects two preliminary questions: whether they have been vaccinated or not and how effective the Covid-19 vaccines are. The effectiveness is asked on a 100-point slider where 0 is noted with "extremely ineffective"; 50 is noted with "neutral"; and 100 is noted with "extremely effective". Based on the answers to the questions, I try to elicit their willingness to pay (WTP) in two different settings: for those who have been vaccinated or plan to receive vaccination, I ask them the amount they will be willing to pay if they need to pay on their own for another dose 6 months and 12 months after they are fully vaccinated, respectively; and for those who said that they do not plan to receive the vaccine, I ask that if there is a mandate to be vaccinated, how much they will be willing to pay to be exempted from the mandate[2]. For all these questions about willingness to pay, I also ask how confident they are in their answers to see if the message will also change people's confidence about their belief on a similar 100-point scale.

After these pre-message questions, I set the survey flow to randomly show the subjects one of the two messages to read for about 3∼4 minutes: one for Tucker Carlson and another for Sean Hannity. I wrote up the two messages in similar format with direct quotations or paraphrases from the two messengers, trying to report their words as factual and concise as possible and in a comparable style. Before reading the message, the subjects are asked to tell their beliefs about the credibility of the Fox News and the messenger whose message is going to be displayed. Immediately after they finish reading the message, I ask them the degree they agree with the message on the 100-point scale with similar notations.

And then, the same questions asking for the effectiveness of the vaccine, their willingness to pay, and the credibility of the messenger show up again to see if their perspective changes because of the treatment. For all these questions, I set the default choice to be their answers to the same question before reading the message so that they will not randomly enter a number and instead think about how their perspectives change carefully.

Lastly, I collect important demographic information about the subjects, including age, gender, education, political affiliation, household income, time spent on social media every week, and how interested they are in politics. The complete survey is attached in Appendix Section A[3].

Getting the Institutional Review Board (IRB) approval after three cycles of protocol revision through the Committee for Protection of Human Subjects (CPHS) at UC Berkeley, I launched the survey on the Amazon Mechanical Turk. I set a criteria asking for subjects that are labeled "Master": those whose are dedicated survey takers with a high rating given by other researchers. In total, 4 batches were launched one by one so that I could check the quality of the samples continuously during the sample collection process. I got 383 responses and 363 of

[2] I avoid asking them questions like how much are you willing to accept (WTA) to receive the vaccines as people view WTP and WTA differently owing to reference dependency and loss aversion.

[3] The flowing logic of the survey is displayed in the square brackets after each question.

them passed the attention check questions and spent a reasonable time taking the survey, forming my sample of interest.

B. Survey Summary Statistics

Table 1 shows the summary statistics about the 363 subjects in my sample. The first column describes the whole sample, and the second and the third columns describe the two arms of messages respectively. Comparing the two arms is meaningful: I can know not only the demographic characteristics of the subject pool of Mturk, but also ensure the sub-samples in different arms do not vary significantly so that all the post-treatment differences are solely due to the treatment. The similarity of the subjects in the two message arms is formally tested in Table 2.

There are 186 (51.24%) subjects reading the message about Tucker Carlson and 177 (48.76%) in the Sean Hannity arm. Focusing on my primary characteristics of interests, the majority of the subjects, 299 (82.37%) of them, have been vaccinated for 5.4 months on average and 54 (14.88%) of the subjects do not plan to receive vaccination at all, which gives me enough variation to explore how people with different perspectives towards vaccination will respond to the messages holding antithetical standpoints. As for the political affiliations, there are 179 (49.31%) Democrats, 86 (23.69%) Republicans, and 93 (25.62%) Independents. Also, I find that these participants are relatively interested in politics on average—the mean rating is 5.31 on a 7-point Likert scale —and they spend a decent amount of time, about 10 hours, on social media every week.

Throughout this paper, I'm going to compare the results between the two message arms—Tucker Carlson and Sean Hannity—and also try to pool the results together to generalize the idea. Although these two messages and the messengers have apparent differences, to ensure the robustness of the main results, I need to ensure that the subjects in these two arms are not statistically different. In Table 1, we can already have a rough idea of the subjects and I test the equality formally in Table 2. Firstly, I put each of the demographic and behavioral characteristics—including gender, household income, education, degree of interest in politics, time spent on media—of the subjects as a dependent variable while holding other ones constant. The main explanatory variable TC is a dummy variable indicating which message arm this subject belongs to: equaling 1 if this subject read the Tucker Carlson message and 0 otherwise. The point estimates from column (1) to (5) are all close to zero and statistically insignificant, guaranteeing that there is no statistical differences between these characteristics of the two sub-samples. I then test if there is any differences in their answers of the survey questions before reading the messages, holding all these demographic and behavioral characteristics constant. The resulting coefficients for the time used to complete the survey, the willingness to pay, the perspectives about vaccine effectiveness and messenger credibility before the treatment are all statistically insignificant. With these values revealing a random split of the sample into two arms, I can ensure that all the differences witnesses are solely owing to the treatment effects and are

Table 1—Subject Summary Statistics

		By Message Arm	
	Whole Sample	*Tucker Carlson*	*Sean Hannity*
Sample Size	363	51.24%	48.76%
by gender:			
Male	210	54.29%	45.71%
Female	152	47.36%	52.63%
by party affiliation:			
Democratic	179	47.49%	52.51%
Republic	86	56.98%	43.02%
Independent	93	51.61%	48.39%
Lean Democratic	13	76.92%	23.08%
Lean Republic	15	60.00%	40.00%
Neither	65	44.62%	55.38%
Others	5	80.00%	20.00%
by education level:			
High School or Less	49	55.10%	44.90%
Some College	58	53.45%	46.55%
Bachelor's Degree	185	48.11%	51.89%
Graduate Degree	70	54.29%	45.71%
by vaccination status:			
Vaccinated	299	49.16%	50.84%
Time after Vaccination	5.4147	5.4149	5.4145
(Months)	(1.825)	(1.790)	(1.864)
Unvaccinated but Plan to	10	50.00%	50.00%
Unvaccinated and not Plan to	54	62.96%	37.04%
Average Interest in Politics	5.307	5.211	5.407
(7-Point Likert Scale)	(1.357)	(1.385)	(1.324)
Average Household Income	66.20	61.49	71.13
(Thousands Dollars)	(82.82)	(68.67)	(95.34)
Average Time Spent on Media	10.43	10.96	9.90
(Hours per Week)	(9.611)	(9.726)	(9.495)
Survey Duration	300.32	296.94	303.81
(Seconds)	(128.86)	(133.77)	(123.91)

Note: This table displays the summary statistics for the subjects taking the Qualtric survey through Amazon Mechanical Turk (Mturk), separated by two arms of messages. There are 383 subjects originally: 20 of them either fail the attention check questions or go through the whole survey too quickly, and are thus excluded from the analysis. The two attention check questions ask the subjects to enter 100 in the text box or move the slider to 100 and responses completed using less than 150 seconds are thought to be invalid. Standard deviations are in the parentheses below the mean and the units of measurement are in the parentheses below the variable names. The average time spent on media only accounts for values less or equal to 42 hours, which equal to a quarter of the hours in a whole week and thus a reasonable maximum answer for this question. Survey duration is the average time the subjects take to complete the survey, with only values less then 689 (the 95 percentile value) included. See the text for detailed interpretations.

Table 2—Randomized Sample Assignment

	Dependent variable:				
	Gender	*Income*	*Education*	*Poli. Interest*	*Time on Media*
	(1)	(2)	(3)	(4)	(5)
TC	-0.0511	-1270.5	-0.0767	-0.172	0.588
	(0.039)	(6303.885)	(0.089)	(0.108)	(0.732)
Observations	644	644	644	644	644
R^2	0.040	0.064	0.031	0.050	0.062
	Dependent variable:				
	Duration	WTP_6	WTP_{12}	*Vaccine Eff.*	*Messenger Cred.*
	(6)	(7)	(8)	(9)	(10)
TC	-42.24	14.71	0.408	-3.136	1.343
	(45.497)	(13.202)	(14.999)	(2.857)	(3.641)
Observations	644	271	270	321	321
R^2	0.030	0.136	0.072	0.027	0.112

Note: The dependent variables are different demographic characteristics, behavioral characteristics, and survey responses before reading the message. Gender is a dummy variable equaling 1 if the subject is male and 0 if female (no subject in this survey identify themselves as other genders); income and willingness to pay are measured in dollars; degree of interest in politics is measured on a 7-point Likert scale; perspectives about vaccine effectiveness and messenger credibility are rated on a 100-point slider; time spent on media is in hours; and duration is the amount of time the subjects spent in completing the survey (in seconds). Standard errors are in the parentheses. Significant at *** 1%, ** 5%, and * 10%.

comparable between the two arms.

C. Empirical Design

I would like to measure how a message, either consistent or inconsistent with an individual's current belief, from a messenger, either trustworthy or not in the individual's perspective, changes the opinion towards the vaccines and the perspective of source credibility respectively. I employ the following equation:

$$\begin{aligned}\Delta y_i = \beta_0 + \beta_1 Agree_i + \beta_2 Credible_i \\ + \beta_3 (Agree_i \times Credible_i) + \epsilon_i,\end{aligned} \tag{1}$$

where Δy_i is my outcome of interest, which is constructed by deducting each subject's perspective about vaccine effectiveness (or messenger credibility) after reading the message by that before reading the message[4]. $Agree_i$ is dummy variable equal to one for those with agreeableness greater than 50 and 0 for those give a score less than 50. $Credible_i$, representing the credibility of the messenger before the treatment, is dichotomized in the same way. In the survey, these two values are measured on a 100-point scale, where 0 represents extremely disagree/distrust and 100 represents extremely agree/trust and 50 means a neutral perspective, so the dichotomy is intuitively reasonable and also makes the results easier to interpret. A histogram of agreeableness of the two messengers is shown in Figure 1. We can see that most of the answers are polarized—either in the 0∼10 bar or the 90∼100 bar, which justifies the dichotomy.

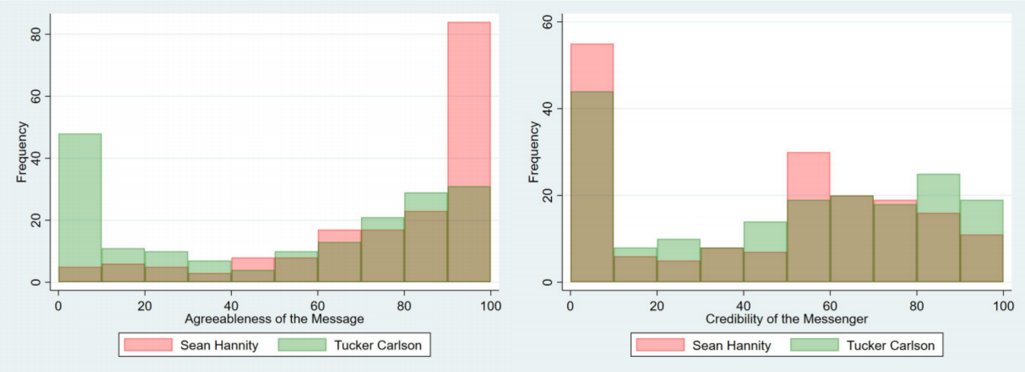

FIGURE 1. AGREEABLENESS AND CREDIBILITY

Note: The histograms depict the agreeableness and credibility (before the treatment) of Carlson and Hannity. Red bars represent Hannity, green bars represent Carlson, and brown bars are the overlaps between the two treatment arms. The minimum is 0, the maximum is 100, and each bin has a width of 10 points.

In this way, the coefficient of the constant term—β_0—tells the change for people who distrust the messenger and disagree with the message; $\beta_0+\beta_1$ tells the change for people who distrust the messenger but agree with the message; $\beta_0+\beta_2$ reveals the change for people who trust the messenger but disagree with the message; and all the coefficients combined tell the effect for people who trust the messenger and also agree with the message. And then a joint test of statistical significance will tell if the change is significant or not.

[4] I don't use percentage changes because there are many subjects answering 0 for the outcome questions before treatment, which will be omitted if using percentage and cause bias.

IV. Empirical Results

A. Treatment Statistics

Firstly, I show some statistics revealing the average change for different survey questions after reading the message in Table 3, separated by political affiliation. Panel A shows the results for the Tucker Carlson arm and Panel B shows that for the Sean Hannity arm. The first two columns show the overall results before and after the treatment; column (3) and (4) shows the average treatment effects for the Democrats; and the last two columns show the effects for the Republicans.

On average, the Republicans agree much more with what Carlson says than the Democrats—the agreeableness is about 28.14 points (69.78%) higher on a 100-point scale—and contrarily, the Democrats' average agreeableness of Hannity's words is 17.47 points (26.75%) higher than that for the Republicans, which give a high variation on one of the explanatory variables to better capture the effect of the messages and also hint at the necessity to sub-sample by political affiliations to detect any differential results.

The beliefs about vaccine effectiveness decrease a little for those reading the Carlson's message and increases for the Hannity counterpart, which makes sense as Carlson questions the vaccines while Hannity supports the vaccines. This may tell that the subjects' opinions, at least on average, are driven by the contents of the messages—a test of significance will be conducted to conclude if there is a real effect. These changes are consistent for both the Democrats and the Republicans, even though the Democrats' average belief about vaccine effectiveness is about 20 points higher.

As for the measurements about the willingness to pay, I can see a positive relationship between the willingness to pay and the belief about vaccine effectiveness, but how the WTP change because of the messages is not the same as expected: if the subjects believe the vaccines to be more effective, they are supposed to be willing to pay more. The results, however, do not follow this intuition mainly because the vast majority of the subjects does not change their willingness to pay before and after reading the message, showing that people consider the money value differently and tend not change that in a short period of time[5].

Looking at the credibility of the sources, I find that the average credibility of Hannity is increased by approximately 10 points (24.14%), leading to a 3.4 point (7.46%) increase in the credibility of the Fox News as well. On the other hand, the change in Carlson's credibility seems to be small, probably because there are highly controversial perspectives towards Carlson's words—as depicted in Figure 1, while most of the agreeableness scores of Hannity's words are greater than 50, that for the Carlson arm is more flatly distributed and is polarized. Also,

[5] The whole survey takes about 5 minutes (300 seconds as shown in Table 1), which is a very short period of time so people may not adjust their perspectives that quickly. This may be a direction of future studies. In this paper, I'll leave out the discussion about willingness to pay because of the limited observations with WTP changed after the treatment.

Table 3—Messages and Average Treatment Effects

	Whole Sample		*Democrats*		*Republicans*	
	Pre	*Post*	*Pre*	*Post*	*Pre*	*Post*
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Tucker Carlson Message						
Number of Subjects	186		85		49	
Message Agreeableness	49.576		40.918		69.061	
(0∼100 Slider)	(37.292)		(38.195)		(29.891)	
Vaccine Effectiveness	71.85	70.76	81.69	80.94	56.37	55.43
(0∼100 Slider)	(26.81)	(28.50)	(17.24)	(19.03)	(33.94)	(36.12)
WTP (6 Months)	115.93	116.85	133.38	134.81	72.50	71.88
(Dollars)	(121.93)	(126.75)	(138.01)	(143.80)	(64.30)	(63.91)
WTP (12 Months)	104.00	108.59	130.05	130.88	57.50	68.44
(Dollars)	(123.99)	(128.99)	(148.29)	(147.00)	(53.28)	(94.43)
Fox News Credibility	48.13	49.39	42.75	42.94	62.22	65.73
(0∼100 Slider)	(33.08)	(36.43)	(37.33)	(40.02)	(24.05)	(28.17)
Carlson Credibility	47.69	46.23	39.49	38.45	65.88	64.10
(0∼100 Slider)	(33.80)	(35.14)	(36.96)	(37.64)	(22.49)	(26.86)
Panel B: Sean Hannity Message						
Number of Subjects	177		94		37	
Message Agreeableness	76.733		82.763		65.297	
(0∼100 Slider)	(27.034)		(21.597)		(32.468)	
Vaccine Effectiveness	75.02	76.86	82.36	83.85	65.89	67.76
(0∼100 Slider)	(23.24)	(24.10)	(16.62)	(17.23)	(25.79)	(27.67)
WTP (6 Months)	107.23	113.95	127.80	135.55	81.07	83.21
(Dollars)	(103.54)	(111.36)	(104.17)	(113.19)	(110.32)	(109.21)
WTP (12 Months)	109.05	111.44	129.39	127.83	81.07	82.86
(Dollars)	(126.81)	(129.81)	(131.81)	(131.97)	(130.04)	(129.22)
Fox News Credibility	45.58	48.93	38.67	41.83	60.30	63.30
(0∼100 Slider)	(34.64)	(35.34)	(36.81)	(37.64)	(28.67)	(27.51)
Hannity Credibility	41.43	51.03	35.52	44.80	56.22	65.19
(0∼100 Slider)	(33.58)	(33.66)	(35.37)	(35.39)	(26.72)	(27.42)

Note: This table displays the average values of survey questions separated by the two arms of messages and before/after reading the message. Outliers are removed as they will affect the average strongly: willingness to pay (6/12 months) includes only values less than 2000. Some of the values, like the willingness to pay to be exempt, cannot be compared between panels because of the huge deviations owing to limited observations. Standard deviations are in the parentheses below the mean and the units of measurement are in the parentheses below the variable names. See the text for detailed interpretations.

it is possible that most of the subjects trust more in Hannity after reading the message, but the number of subjects who trust more in Carlson is very close to the number of subjects who trust less in him, which is proven to be true in Figure 3. More over, the average changes in credibility from both treatment arms do not reveal a clear difference between the Democrats and the Republicans: both change in the same direction with similar magnitude.

I then draw four scatter plots of the effects of the two messages on the subjects' perspectives. On the horizontal axis is the agreeableness of the message and on the vertical axis is the pre-treatment credibility of the messenger. The size of the circles is weighted by the absolute value of the change in the subjects' perspectives about vaccine effectiveness before and after reading the message—namely, the larger the circle, the more this individual changes because of the message. As a result, subjects who do not change their perspectives after the treatment is weighted by zero and therefore not shown on these plots. A red circle implies a decrease and a green circle represents an increase.

The scatter plots about vaccine effectiveness are in Figure 2, where I can see a clear concentration of the points at the upper right corner—for the subjects with agreeableness from 50 to 100 and pre-treatment credibility from 50 to 100—especially in the Hannity arm. And there is another concentration in the Carlson arm when both agreeableness and pre-treatment credibility is lower than 50. This suggests that the dichotomy of the two explanatory variables is reasonable. The scatter plots about the effects on credibility in Figure 3, however, tells a different story: In addition to the concentration of the points noted previously, there is also a clear pattern based on the colors: If I draw a diagonal line where $Agreeableness = Pre\_Credibility$, the majority of the red circles are above the line and most of the green circles are below the line. This finding suggests me to expand equation (1) by another term—ALC_i—which is a dummy variable indicating whether the degree of agreeableness is greater than the degree of trustworthiness before the treatment[6]. Both values are measured on a scale of 100 points so they are comparable. The inclusion of ALC_i can further divide the sub-sample for which $Agree_i = Credible_i = 0$ or $Agree_i = Credible_i = 1$ and better show the differential trends, if any. It turns out that this is one of the most important findings in this paper, which will be discussed in detail in Section IV.

Considering ALC_i, I run the following regressions:

$$
\begin{aligned}
\Delta Messenger\ Credibility_i = \beta_0 &+ \beta_1 Agree_i + \beta_2 Credible_i \\
&+ \beta_3 ALC_i + \beta_4 (Agree_i \times Credible_i) \\
&+ \beta_5 (Agree_i \times ALC_i) + \epsilon_i,
\end{aligned} \tag{2}
$$

$$
\Delta Messenger\ Credibility_i = \beta_0 + \beta_1 ALC_i + \epsilon_i, \tag{3}
$$

where terms like $Credible_i \times ALC_i$ and $Agree_i \times Credible_i \times ALC_i$ are excluded in equation (2) because of perfect multi-collinearity. The coefficient of ALC_i and $Agree_i \times ALC_i$ will tell if there is any difference in the treatment effects if the degree of agreeableness is higher (or less) than the degree of trustworthiness. Imagine a situation where you highly believe in a messenger, and then you read a message from them—even though you agree with the content of the message, but

[6] ALC is the abbreviation of "Agreeableness Larger than pre-treatment Credibility".

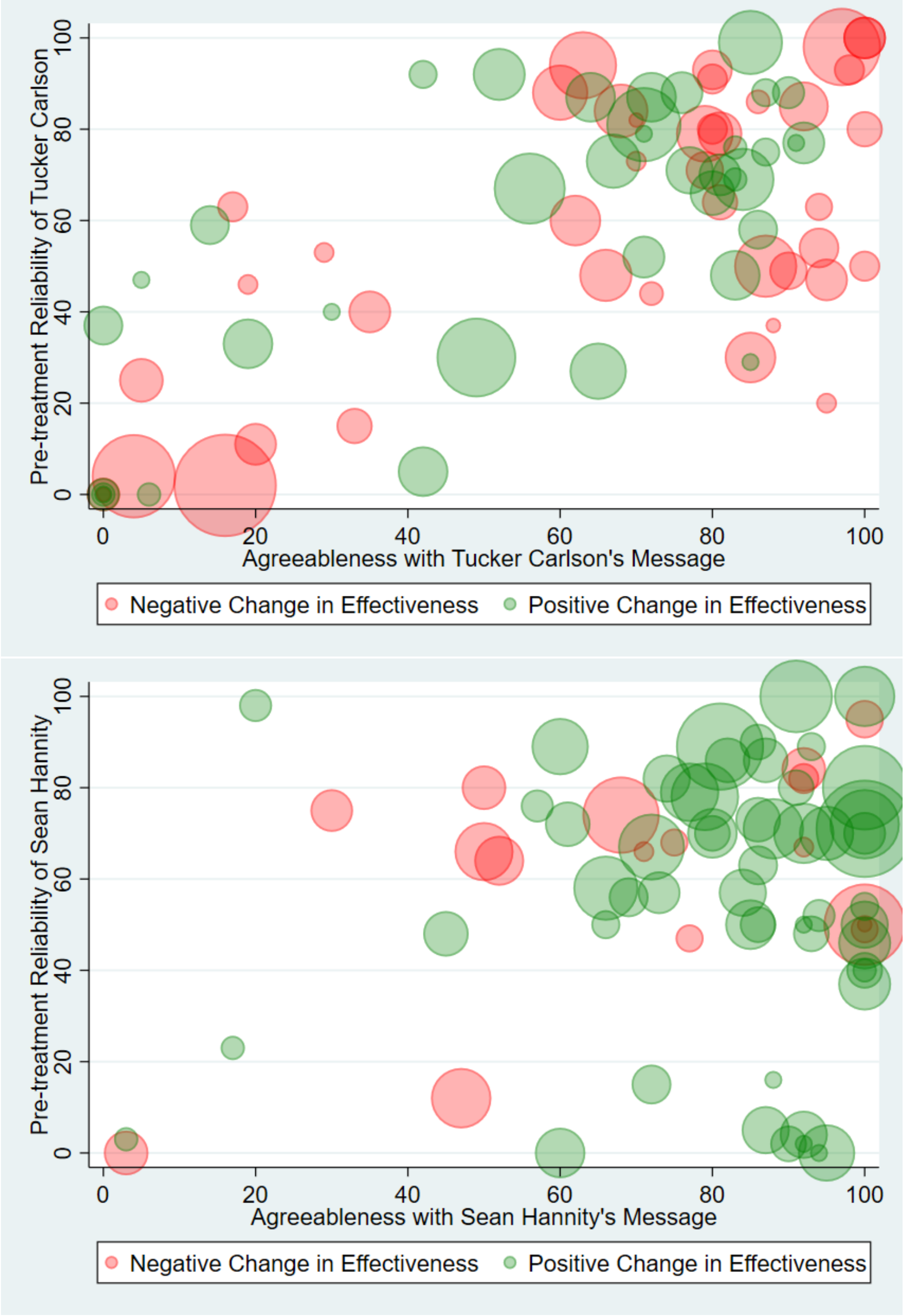

Figure 2. Messages, Messengers, and Perspectives on Vaccine Effectiveness

Note: The scatter plots depict the agreeableness and credibility (before the treatment) of Carlson and Hannity, and the resulting change in the subjects' perspectives on vaccine effectiveness. The agreeableness of the message is on the horizontal axis and the pre-treatment credibility of the messenger is on the vertical axis. The size of the circle represents the size of change. A green circle implies an increase and a red circle represents a decrease. Subjects without a change in their perspectives are not shown on the plot: there are 106 subjects whose perspectives do not change after reading the Carlson's message and 105 subjects do not change their perspectives after reading the Hannity's message.

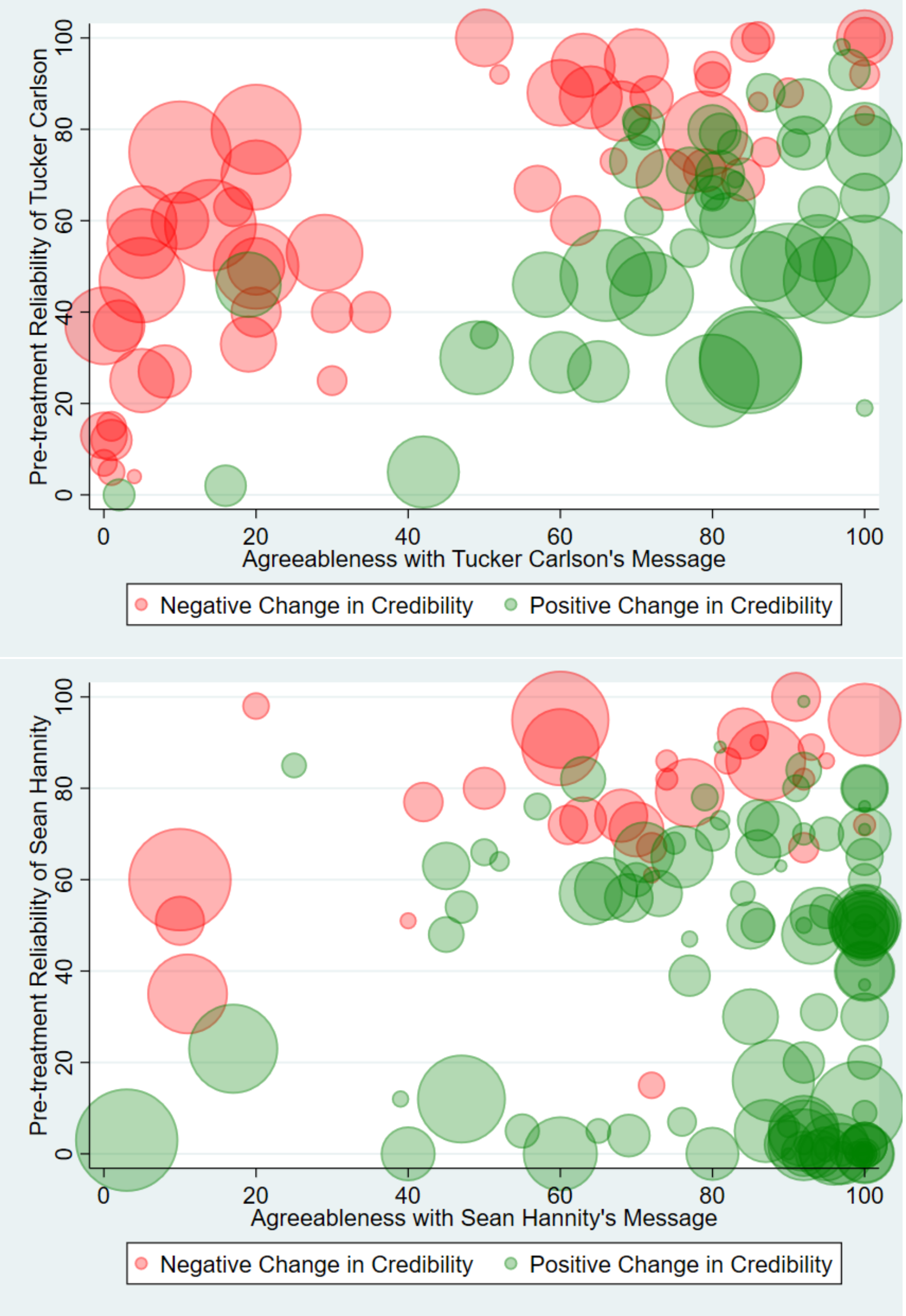

Figure 3. Messages, Messengers, and Perspectives on Messenger Credibility

Note: The scatter plots depict the agreeableness and credibility (before the treatment) of Carlson and Hannity, and the resulting change in the subjects' perspectives on the credibility of the sources (messengers). The agreeableness of the message is on the horizontal axis and the pre-treatment credibility of the messenger is on the vertical axis. The size of the circle represents the size of change. A green circle implies an increase and a red circle represents a decrease. Subjects without a change in their perspectives are not shown on the plot: there are 85 subjects whose perspectives do not change after reading the Carlson's message and 44 subjects do not change their perspectives after reading the Hannity's message.

the agreeableness seems not as high as you expected given the high credibility you gave them. Will you trust less in the messenger? Equation (4) is a simplified version of equation (3) employed in Section IV. C. when talking about the heterogeneity by different political affiliation because the the number of observations in each branch are too small to support a 6-way division after taking sub-samples.

B. *Effect on Vaccine Effectiveness and Source Credibility*

I employ equation (1) and (2) to investigate the effects of the two messages on the effectiveness of vaccines and the credibility of the messengers under different conditions, which are shown in Table 4. Column (1) to (3) display the regression results for the Tucker Carlson arm and column (4) to (6) are for the Sean Hannity arm. The dependent variable in column (1) and (4) is the change in the effectiveness of the vaccines, measured on the same 100-point scale, and the dependent variable in column (2), (3), (5), and (6) is the change in the credibility of the corresponding messenger. Column (2) and (5) uses equation (1) to show the effects of each of the four *Agree*-*Credible* combinations and column (3) and (6) employ equation (2) to further divide into six scenarios with ALC.

In the first two columns, the coefficient of the $Constant$ term represents the effect of the message for people who distrust Carlson and disagree with the message. The coefficient of the *Agree* term shows the additional effect (on top of the $Constant$ coefficient) on those who distrust Carlson but agree with the message. The coefficient on the *Credible* term tells the difference in the effects for people who trust and distrust Carlson, given that they disagree with the message[7]. The interaction term $Agree \times Credible$ tells us the additional effect adding on to the $Constant$ coefficient and both single terms for people who trust Carlson and agree with the message. Column (3) includes the ALC term, which captures the difference when the degree of agreeableness is higher (or less) than the degree of trustworthiness, given that the subjects distrust Carlson and disagree with the message. And $ALC + Agree \times ALC$ captures the same difference when the subjects trust Carlson and agree with the message. Column (4) to (6) have the same interpretations for the Hannity counterpart.

Overall, I can see that the regression coefficients for the change in credibility are larger and more significant than that for the effectiveness of the vaccines, revealing a great degree of confirmation bias—people tend to interpret new message based on their *ex ante* belief. Instead of updating their own belief based on the new piece of information, people are more likely to compare the closeness of the new information with their own belief and then update on the perspective about the credibility of the messenger (source). Looking at column (3) and (6), the coefficient of ALC and $Agree \times ALC$ are statistically significant at 1% level for Carlson arm and Hannity arm, respectively. This shows that the relative degree of message agreeableness and messenger credibility do play an indispensable role,

[7] This is the same to say that it's the additional effect on top of the $Constant$ coefficient.

Table 4—Treatment Effect on Vaccine Effectiveness and Credibility

	Dependent variable By Arm					
	Tucker Carlson			*Sean Hannity*		
	$\Delta Eff.$	$\Delta Cred-4$	$\Delta Cred-6$	$\Delta Eff.$	$\Delta Cred-4$	$\Delta Cred-6$
	(1)	(2)	(3)	(4)	(5)	(6)
Agree	-0.703	23.455***	16.227***	2.459*	-0.953	-18.561***
	(1.877)	(2.525)	(3.147)	(1.485)	(3.823)	(5.875)
Credible	1.530	-24.379***	-22.737***	0.705	-19.080***	-20.232***
	(2.060)	(2.770)	(2.716)	(2.090)	(5.381)	(5.984)
Agree × *Credible*	-0.169	2.097	4.434	2.948	8.885	15.607**
	(2.7699)	(3.724)	(3.721)	(2.288)	(5.889)	(6.525)
ALC			7.739***			-2.634
			(2.553)			(6.709)
Agree × *ALC*			1.130			19.090**
			(3.201)			(7.563)
Constant	-1.364*	-2.455**	-4.096***	-1.250	15.625***	16.778***
	(0.807)	(1.087)	(1.176)	(1.334)	(3.435)	(4.438)
Observations	171	171	171	159	159	159
R^2	0.0033	0.3683	0.4205	0.0818	0.0902	0.1511

Note: The dependent variables are the change in belief about the effectiveness of vaccines (measured on 0∼100 slider) and the change in the credibility of the messenger before or after reading the message, separated by two message arms. Standard errors are in the parentheses. Significant at *** 1%, ** 5%, and * 10%.

in addition to their absolute levels.

In Table 5, I represent the marginal effects for each of the four *Agree-Credible* or six *Agree-Credible-ALC* combinations, calculated based on Table 4. Column (1) and (4) show the marginal effects on the subjects' perspectives about the effectiveness of Covid-19 vaccines, column (2) and (4) show the marginal effects on the credibility of the messenger—either Carlson or Hannity—under four kinds of *Agree-Credible* situations, and column (3) and (6) show that with 6 situations considering *ALC*. Note that some of the results in these two tables, even though significant, may not be implicative because they may be driven by a very few extreme cases—I will focus primarily on any situation where there is a decent number of valid observations (the subjects whose perspectives changes after the treatment).

Looking at the marginal effects on vaccine effectiveness across the messages, I find that many of the changes are very small and not statistically significant. The most implicative results are in column (4) row (2) and row (4) when the subjects

Table 5—Marginal Effect on Vaccine Effectiveness and Credibility

			Marginal Effect By Arm					
			Tucker Carlson			*Sean Hannity*		
Agree	*Credible*	*ALC*	$\Delta Eff.$ (1)	$\Delta Cred-4$ (2)	$\Delta Cred-6$ (3)	$\Delta Eff.$ (4)	$\Delta Cred-4$ (5)	$\Delta Cred-6$ (6)
0	0	0	-1.363* (0.0923)	-2.455** (0.0245)	-4.096*** (0.0006)	-1.250 (0.3496)	15.625*** (0.0000)	16.778*** (0.0002)
		1			3.643 (0.1089)			14.143*** (0.0053)
1	0	1	-2.066 (0.2235)	21.00*** (0.0000)	21.00*** (0.0000)	1.209* (0.0647)	14.672*** (0.0000)	14.672*** (0.0000)
0	1	0	0.167 (0.9300)	-26.833*** (0.0000)	-26.833*** (0.0000)	-0.545 (0.7349)	-3.455 (0.4050)	-3.455 (0.3901)
1	1	0	-0.705 (0.3433)	1.650 (0.2005)	-6.171*** (0.0000)	4.862*** (0.0000)	4.477*** (0.0090)	-6.409** (0.0246)
		1			2.698** (0.0377)			10.047*** (0.0000)

Note: The table displays the marginal effects on the belief about the effectiveness of vaccines (measured on 0∼100 slider) and the credibility of the messengers before or after reading the message, separated by two message arms and 4 (or 6) combinations of treatment branches. P-values are displayed in the parentheses. Jointly significant at *** 1%, ** 5%, * 10%.

agree with what Hannity says: for those who agree with the message and believe in Hannity, their belief about the effectiveness of the vaccine increases by 4.48 points (a 5.97% increase) ; and for those who agree with the message but do not believe in Hannity, their belief increases by 1.21 points (a 1.61% decrease). It intuitively makes sense that a source people trust in will make a greater effect than a source they distrust, given that they agree with the information from this source. While the exact magnitude of the changes and the significance of the effect is hard to be generalized, most point estimates point towards the predicted direction: if people agree with the message, they will update their belief following the direction or the message; otherwise, their beliefs will move against the direction of the message[8]. As Carlson is condemning the vaccines and Hannity is supporting it, it's reasonable that people who distrust them and disagree with the messages will believe the vaccines to be more (and less) effective, respectively. And this reveals the existence of confirmation bias: people tend to judge the new information

[8] The point estimate in col (1) row (1) seems to an anomaly: based on Figure 2, the marginally significant result is driven by two extremely big drops (denoted by two big red circles), which may be a bias because of the limited sample size.

based on their *ex ante* knowledge, and will re-affirm their own beliefs regardless of whether they agree or disagree with the message.

Another main question of my interest is how the credibility of the sources will be affected—if people judge the message based on their own beliefs about the vaccines, they may also use this as a basis to update their perspectives about the trustworthiness of the messengers. Focusing now at the marginal effects on the credibility of the messengers, column (2) and (4) employs equation (1) to conduct a 4-way decomposition and column (3) and (6) employs equation (2) to further decompose the effect into 6 situations.

Between the 4-way and 6-way decomposition, the coefficients in row (2) and (3)—agree but distrust & trust but disagree—will not change as these two conditions necessarily imply that the agreeableness is greater (or less) than the pre-treatment credibility. The 6-way decomposition only divide further the coefficients when the subjects both agree with the message and trust the messenger (both scores are greater than 50) or disagree with message and distrust the messenger (both scores are less than 50). When the subjects disagree with message and distrust the messenger, as presented in row (1), a pooled result in column (2) show a 2.455 point reduction (a 5.15% decrease from the average) in credibility of Carlson. This is misleading as I show in column (3) that there are two opposite effects included: if the degree people trust Carlson is relatively higher than the degree they agree with the message ($ALC = 0$), they will lower Carlson's credibility by 4.1 points (a 8.59% decrease from the average), which is significant at all conventional levels. On the other hand, if $ALC = 1$, they will trust more in Carlson by increasing his credibility by 3.64 points (a 8.59% increase from the average), which is marginally significant with a P-value equaling to 0.1089. Similar situations happen when the subjects agree with the message and trust the messenger, which holds for the Hannity's arm as well[9].

With such an impressive finding, I pooled the two message arms together and regress the change in credibility of the messenger—$\Delta Credibility_i$—on the difference between the agreeableness score and the pre-treatment credibility score—denoted by AMC_i, which is $Agreeablness_i$ minus $Pre\_Credibility_i$. The regression equation is the following:

$$\Delta Credibility_i = \alpha_0 + \alpha_1 AMC_i + \epsilon_i. \tag{4}$$

The regression results are displayed in Figure 7 together with an analysis about the differential effects for subjects with different political affiliations. The constant term (not reported) is insignificant, telling that there is no change in credibility where there is no difference between the agreeableness and the pre-treatment credibility. And for each point difference between the agreeableness

[9] The results in row (1) column (5) and (6), even though statistically significant, should not be taken into consideration as they are driven by very few observations in that category, as depicted in Figure 3.

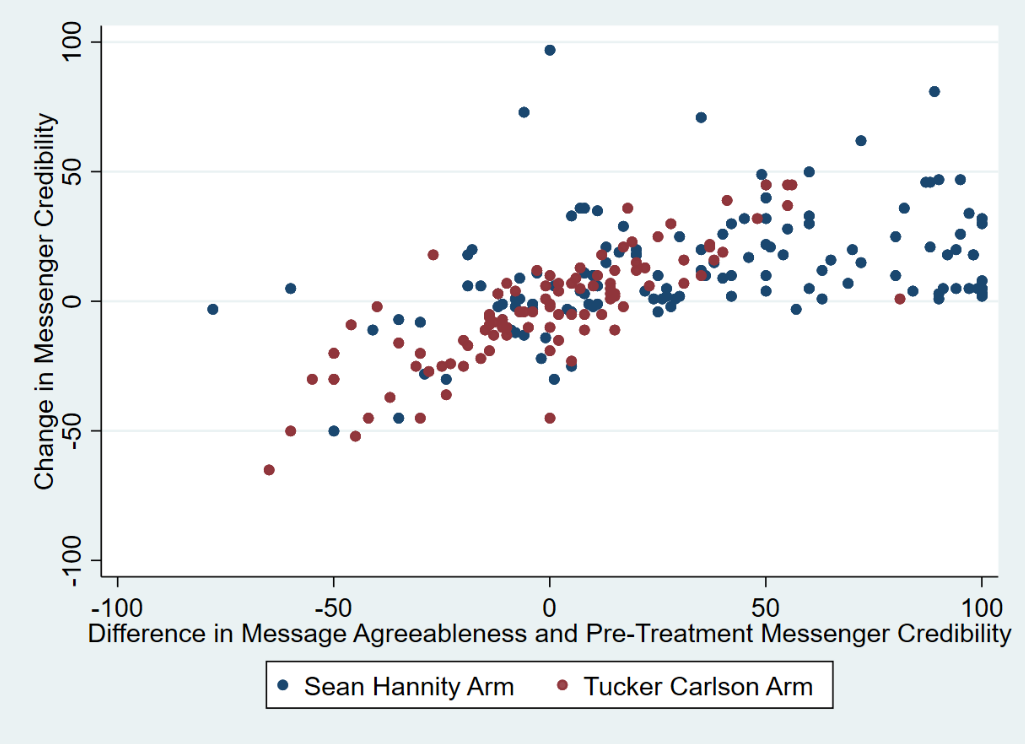

Figure 4. AMC and Change in Credibility

Note: The scatter plot depicts AMC—the difference between the agreeableness fo the message and the credibility (before the treatment) of Carlson and Hannity—and the magnitude of the changes in the subjects' perspectives on the source credibility. AMC is on the horizontal axis and $\Delta Credibility$ is on the vertical axis. Red circles represent subjects in the Tucker Carlson arm and blue circles represent the Hannity counterpart.

and the pre-treatment credibility (all measured on a 100-point scale), the subjects' perspectives about the credibility of the messenger will increase by 0.25 point, which is statistically significant at all conventional levels as shown in the first 3 columns and is robust with additional controls. Figure 4 displays a sactter point of $\Delta Credibility_i$ and AMC_i, which shows a clear positive relationship. This reveals that the process of thinking about credibility is more complicated than a dichotomized one: even if people distrust the messenger and disagree with the message, confirmation bias only let them to reaffirm that what the message says is wrong and have little effect on their perspectives about messenger credibility. Instead of reaffirming that the source is not trustworthy, people tend to compare, on a continuous scale, the message and the credibility of the messenger they used to deem—even though not trustworthy, people may believe the messenger to be relatively more trustworthy as long as what the message says is not "as bad as" they expected[10].

Unlike the perspectives, which deviates from a 2.07 (2.88%) reduction to a 4.86 (6.48%) increase, the credibility is extremely volatile: it can decrease as much as 26.83 points (56.26%) and increase as much as 21 points (44.03%). This result aligns well with intuition: people will update their beliefs following the direction of their pre-existing knowledge, and then adjust the credibility of the messenger accordingly—the messenger, even if a trustworthy one, do not have much power in face of a disagreeable message. Furthermore, it implies that people tend to be "self-centered" when determining who or which source is credible and the so-called credibility is extremely vulnerable: instead of changing their own minds to conform to the source, they distrust a source they thought to be credible easily as soon as the agreeableness of the message from the source does not match its credibility. In conclusion, people interpret the message and update their belief based on *ex ante* knowledge but update their perspectives about source credibility based on the message.

C. Heterogeneity By Political Affiliation

Both messages and the underlying questions about whether the Covid-19 vaccines are effective are highly polarized questions. Therefore, it is necessary to explore if there will be any differences in the effects between the Democrats and the Republicans. In this section, I rerun the regressions using sub-samples to see if there is any heterogeneity in the treatment effects.

Table 6 displays the marginal effects of the messages under different situations, separated by Democrats and Republicans: these effects are calculated based on the regression results attached in Appendix Table A1. Table 6 Panel A shows the marginal effects on the perspectives about vaccine effectiveness using equation (1) and Panel B shows the effects on source credibility using the simplified equation (4)[11].

Looking at Panel A, I find that most of the coefficients are insignificant for both the Democrats and the Republicans, showing that the subjects still stick to their original beliefs after reading the message. There are, however, three statistically significant effects for the Republicans: a 8 point increase in the subjects' perspectives about vaccine effectiveness for those who distrust Carlson and disagree with his words; a 1.4 point decrease for those who trust in Calrson and agree with him; and a 4.4 point increase for those who trust Hannity and agree with his words. All these three estimates are supported by decent number of observations under

[10] Here is a numerical example: if the agreeableness score is 20 and the subject rates the credibility of the source to be 0 (both less than the neutral score of 50), he may give the messenger a score higher than 0 after reading the message because what the messenger says is not as disagreeable as he expected. And as a result the messenger is not as unreliable as he used to believe.

[11] Equation (3) is not used because of limited observations: for instance, there are only 37 Republicans in the Hannity arm as shown in Table 3, so there will be several situations with very unreliable estimates driven by only a few observations if using the 4-way or 6-way decomposition.

Table 6—Marginal Effect By Political Affiliation

		TuckerCarlson			*Sean Hannity*		
Agree	*Credible*	*Whole*	*Dem.*	*Rep.*	*Whole*	*Dem.*	*Rep.*
Panel A: Vaccine Effectiveness							
0	0	-1.363*	-0.476	8.00***	-1.250	-6.333*	-0.143
		(0.0923)	(0.6468)	(0.0080)	(0.3496)	(0.0681)	(0.929)
1	0	-2.067	-2.800	-0.125	1.209*	0.886	1.000
		(0.2235)	(0.3530)	(0.9325)	(0.0647)	(0.3265)	(0.5985)
0	1	0.167	-0.667	0.571	-0.545	1.333	0.000
		(0.9300)	(0.8638)	(0.7174)	(0.7349)	(0.6996)	(1.0000)
1	1	-0.705	-0.824	-1.423*	4.862	4.750	4.429***
		(0.3433)	(0.4760)	(0.0851)	(0.0000)	(0.0000)	(0.0002)
ALC		*Whole*	*Dem.*	*Rep.*	*Whole*	*Dem.*	*Rep.*
Panel B: Messenger Credibility							
0		-8.038***	-5.982***	-14.783***	0.667	2.059	0.188
		(0.0000)	(0.0000)	(0.0000)	(0.7402)	(0.5204)	(0.9571)
1		7.075***	8.000***	9.731***	13.198***	11.013***	15.667***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Note: The table displays the marginal effects on the belief about the effectiveness of vaccines (measured on 0$\sim$100 slider) and the credibility of the messengers before or after reading the message, separated by two message arms, political affiliations, and different treatment branches. P-values are displayed in the parentheses. Jointly significant at *** 1%, ** 5%, * 10%.

those situations. The significant 6.33 point reduction for those who distrust Hannity and disagree with his message, however, is driven by very few observations as depicted in Figure 2, which might be unreliable. As a result, I can conclude with some confidence that both the Democrats and the Republicans have a clear tendency to interpret new messages based on their *ex-ante* beliefs. Their beliefs are hardly affected by these messages, but the Republicans are more likely to witness self-affirmation, believing significantly more in their original beliefs.

The results for the effects on the credibility in Panel B are more statistically significant and resonate well with the previous interpretations. For all the significant marginal effects, the Republicans experience higher changes in messenger credibility than the Democrats. For example, Democrats who trust Carlson more than they agree with the message decrease the credibility by 5.98 points while Republicans decrease the credibility by 14.783 points under the same situation. However, this effect is not conclusive as the higher changes may be driven by the fact that the Republicans in general trust in these two Fox News host more than the Democrats do, and therefore simply have more room to change. To

Table 7—Credibility and Agreeableness-Credibility Difference

	Dependent Variable: Δ *Credibility*					
	(1)	(2)	(3)	(4)	(5)	(6)
AMC	0.246*** (0.016)	0.251*** (0.016)	0.250*** (0.017)	0.359*** (0.037)	0.362*** (0.037)	0.400*** (0.041)
AMC × *Demo*				-0.178*** (0.043)	-0.180*** (0.043)	-0.224*** (0.047)
Demographic Control	No	Yes	Yes	No	Yes	Yes
Behavioral Control	No	No	Yes	No	No	Yes
Observations	720	714	670	528	524	494
R^2	0.252	0.263	0.258	0.241	0.255	0.263

Note: The dependent variable is the change in the belief about the credibility of the messenger (measured on 0∼100 slider) before or after reading the message. *AMC* is the difference between the agreeableness and the pre-treatment messenger credibility. Standard errors are in the parentheses. Significant at *** 1%, ** 5%, and * 10%.

circumvent this problem, I employ equation (3) again, interacted with *Demo*, an indicator of whether the subject is a Democrat or a Republican[12]:

$$\Delta Credibility_i = \alpha_0 + \alpha_1 AMC_i \times Demo_i + \epsilon_i. \tag{5}$$

The regression results are shown is Table 7. Column (1) to (3) is a pooled regression as discussed in the previous section and column (4) to (6) are interacted with *Demo*. Column (2) and (5) include demographic control variables like age, gender, education, and household income. Column (3) and (6) further include behavioral control variables like the amount of time spent on social media each week.

The pooled regression tells a 0.25 point increase in messenger credibility for every unit difference between the agreeableness of the message and the pre-treatment credibility. For the Republicans, this effect becomes larger—a 0.36 to 0.4 point increase for each unit different—and for the Democrats, this effect is about 0.18 to 0.22 points smaller then that for the Republicans. Across different specifications, the resulting marginal of one-point difference is about 0.18 point for the Democrats. All these point estimates are statistically significant at 1 % level and are robust with the demographic and behavioral control variables.

These results can be generalized into a simple mechanism of credibility adjust-

[12] Note that $Demo_i = 0$ does not mean that the individual is not a Democrat because there is a considerable subjects who are Independents or with other political affiliations. These subjects are assigned a missing value to the variable *Demo* and not considered in the discussion in this section

ment when receiving a piece of new information:

$$Post\_Credibility = Pre\_Credibility + \delta(Agreeableness - Pre\_Credibility), \tag{6}$$

where δ belongs to 0 and 1. If $\delta = 0$, then this individual is completely loyal to the source and is apathetic about the message: whatever the source says, the individual will trust (or distrust) the source steadfastly as he used to be. On the other hand, if $\delta = 1$, then this individual has no royalty toward this messenger at all: the credibility of the source will be exactly the same as the agreeableness of the latest message from him this individual hears. This δ measures how much percent of the difference is deemed by people as their "mistake"—the degree they trust in this source too much (or too less) previously—and will be rectified. Therefore, I call this δ the Marginal Propensity to Trust. In this study, given the same messages, the Marginal Propensity to Trust for the Republicans is significantly higher than that for the Democrats, which gives the first evidence that the Republicans tend to be more gullible.

V. Conclusion

This paper reveals an underlying psychological phenomenon in our judgement process. People have a tendency to interpret new things based on their preexisting knowledge, I confirm and expand this concept with a case study on Covid-19 vaccines.

Employing a Qualtrics survey with a high-quality sample from the Amazon Mechanical Turk, I gain first-hand data on people's beliefs about vaccine effectiveness and credibility of the sources—either Tucker Carlson or Sean Hannity. Across different specifications, I document a considerable degree of confirmation bias in that people's perspectives towards vaccine effectiveness are either unaffected by the message or, instead, their *ex-ante* beliefs are re-affirmed. Moreover, people utilize the message as a basis to adjust the reliability of the messenger: they can easily distrust a messenger they used to believe in if they receive a disagreeable message from the messenger. Beyond the dichotomy, the relative degree of agreeableness and pre-treatment credibility matters in a continuous sense: people will distrust the messenger as long as the degree they agree with the message is not as high as the degree they trust the messenger and *vice versa*. Furthermore, I find evidence that this tendency to reaffirm prior beliefs and to readily trust/distrust a source is much more prominent for the Republicans than for the Democrats.

These findings have implications on public policies like the use of "nudges" to promote welfare programs and the necessity for the governments to make some paternalistic interventions. What other characteristics of a piece of information may influence the thoughts of the readers differently and how the mechanism can be further consolidated across diverse messages are left for future studies.

REFERENCES

F. Bacon. Novum organum. *The English Philosophers from Bacon to Mill*, pages 24–123, 1620/1939.

J. S. Bruner, J. J. Goodnow, and G. A. Austin. *A study of thinking.* Routledge, 2017.

R. Clay, J. M. Barber, and N. J. Shook. Techniques for measuring selective exposure: A critical review. *Communication Methods and Measures*, 7(3-4): 147–171, 2013.

S. DellaVigna and E. Linos. Rcts to scale: Comprehensive evidence from two nudge units. Technical report, 2020.

M. Jones and R. Sugden. Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50(1):59–99, 2001.

H. H. Kelley. *The warm-cold variable in first impressions of persons.* 1950.

S. Knobloch-Westerwick, C. Mothes, B. K. Johnson, A. Westerwick, and W. Donsbach. Political online information searching in germany and the united states: Confirmation bias, source credibility, and attitude impacts. *Journal of Communication*, 65(3):489–511, 2015.

A. Koriat, S. Lichtenstein, and B. Fischhoff. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2):107, 1980.

P. E. Lehner, L. Adelman, B. A. Cheikes, and M. J. Brown. Confirmation bias in complex analyses. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(3):584–592, 2008.

R. K. Merton. Social structure and anomie. *Social Theory and*, 1957.

R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

R. E. Petty, P. Briñol, and J. R. Priester. Mass media attitude change: Implications of the elaboration likelihood model of persuasion. In *Media effects*, pages 141–180. Routledge, 2009.

R. F. Pohl. Introduction: cognitive illusions. In *Cognitive illusions*, pages 13–32. Psychology Press, 2012.

K. Popper. *The logic of scientific discovery.* Routledge, 2005.

H. Shaklee and B. Fischhoff. Strategies of information search in causal analysis. *Memory & Cognition*, 10(6):520–530, 1982.

P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140, 1960.

P. C. Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

A. Westerwick, B. K. Johnson, and S. Knobloch-Westerwick. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs*, 84(3):343–364, 2017.

APPENDIX

A1. Qualtrics Survey

1. INTRODUCTION.

University of California at Berkeley
Consent to Participate in Research
Confirmation Bias: The Role of Message and Messenger
CPHS Protocol ID: 2021-10-14711 (Daniel Acland)

Introduction and Purpose

My name is Dan Acland. I am a faculty member at the University of California, Berkeley in the School of Public Policy. I would like to invite you to take part in my research study with Hongyu (Randol) Yao from the University of California, Berkeley. This study concerns how messages and messengers will affect your perspective.

Procedures

If you agree to participate in our research, We will ask you to complete the attached online survey. The survey will involve multiple-choice, slider, and Likert-scale questions about Covid-19 vaccines, and should take less than 10 minutes to complete.

Benefits

There is no direct benefit to you from taking part in this study. It is hoped that the research will help better understand how confirmation bias works in reality under different situations.

Risks/Discomforts

Some of the research questions may make you uncomfortable or upset. You are free to decline to answer any questions you don't wish to, or to stop participating at any time. As with all research, there is a chance that confidentiality could be compromised; however, we are taking precautions to minimize this risk. You can protect your privacy by clearing your browser's history, cache, cookies, and other browsing data. (Warning: This will log you out of online services.)

Confidentiality

Your study data will be handled as confidentially as possible. If the results of this study are published or presented, individual names and other personally identifiable information will not be used. No identifiable information will be collected. To minimize the risks to confidentiality, we will store the data in password-protected computers and destruct the survey results on Qualtrics. When the research is completed, de-identified data will be retained indefinitely for possible use in future research done by ourselves or others.

Compensation
Amazon Mechanical Turk will compensate in accordance with the suppliers' incentive guidelines.

Rights
Participation in research is completely voluntary. You are free to decline to take part in the project. You can decline to answer any questions and are free to stop taking part in the project at any time. Whether or not you choose to participate, to answer any particular question, or continue participating in the project, there will be no penalty to you or loss of benefits to which you are otherwise entitled.

Questions
If you have any questions about this research, please feel free to contact us. You can reach us at acland@berkeley.edu or at hongyu_yao@berkeley.edu. If you have any questions about your rights or treatment as a research participant in this study, please contact the University of California at Berkeley's Committee for Protection of Human Subjects at 510-642-7461, or e-mail subjects@berkeley.edu.

If you have any questions about your rights or treatment as a research participant in this study, please contact the University of California at Berkeley's Committee for Protection of Human Subjects at 510-642-7461, or e-mail subjects@berkeley.edu.

If you agree to take part in the research, please print or download a copy of this consent form to keep for your records and then choose the "Accept" button below. Click "Next" to start the survey.

○ ACCEPT

2. INSTRUCTION.
Firstly, you'll be presented with a set of questions—please answer based on your previous impression and experiences. And then, you'll be presented with a short passage (3∼4 minutes) to read—you may agree or disagree with the passage, but please read through carefully and understand all the points the passage is making. Finally, you'll be presented another set of questions—please answer based on your experiences and also feelings after reading the short passage.

3. COVID AND VACCINES. (Pre-Message)
Q0. Have you ever received any Covid-19 vaccinations?
○ Yes!
○ No, but plan to
○ No and don't plan to

Q1.1. Let's think about Covid vaccines. [display if the answer of Q0 == Yes!]
How long it has been since you finished your two doses? (7-point Likert Scale)

Q1.2. Let's think about the effectiveness of Covid vaccines...
How effective do you think the vaccines are? (0∼100 Slider)

Q1.3. If you have to pay for your own, what's your willingness to pay for another dose of Covid vaccine 6 months after you have been fully vaccinated? (please enter a whole number in USD; e.g. put 200 if you're willing to pay $200) [display if the answer of Q0 == Yes!]

Q1.3.5. Our willingness to pay may vary based on how certain we are, please indicate on the slider how confident are you in your answer to the previous question? [display if the answer of Q0 == Yes!] (0∼100 Slider)

Q1.3. If the government issues a vaccine mandate, how much do you willing to pay to be exempted from the mandate? (please enter a whole number in USD; e.g. put 200 if you're willing to get vaccination when you're rewarded $200) [display if the answer of Q0 != Yes!]

Q1.4. Our willingness to pay may vary based on how certain we are, please indicate on the slider how confident are you in your answer to the previous question? [display if the answer of Q0 != Yes!] (0∼100 Slider)

Q1.4. If you have to pay for your own, what's your willingness to pay for another dose of Covid vaccine 1 year after you have been fully vaccinated? (please enter a whole number in USD) [display if the answer of Q0 == Yes!]

Q1.4.5. Our willingness to pay may vary based on how certain we are, please indicate on the slider how confident are you in your answer to the previous question? [display if the answer of Q0 == Yes!] (0∼100 Slider)

[Version 1] Q1.5. Let's think about Tucker Carlson, who currently serves as a host of FOX News Channel's flagship primetime cable news program, Tucker Carlson Tonight (weekdays 8PM/ET)
How reliable do you think FOX News is? (0∼100 Slider)
How reliable do you think Tucker Carlson is? (0∼100 Slider)

[Version 2] Q1.5. Let's think about Sean Hannity, the host of The Sean Hannity Show, a nationally syndicated talk radio show, and also the host of a commentary program on Fox News.
How reliable do you think FOX News is? (0∼100 Slider)
How reliable do you think Sean Hannity is? (0∼100 Slider)

Tucker Carlson Believes Vaccines Unreliable, Telling Viewers to Ignore 'Medical Advice on Television'

Fox News host on the July 19 edition of his show continued to encourage viewers to question the efficacy of COVID-19 vaccines Fox News

- **Tucker Carlson on Monday urges viewers to question COVID-19 vaccines.**
- **Fox News has been distrusting the vaccine in parts of the United States.**
- **Joe Biden's plan to promote booster shots is condemned by Tucker Carlson.**

Over the course of the COVID pandemic, Tucker Carlson, host of the Tucker Carlson Tonight show on Fox News, has shown that COVID vaccines are not very effective. On Monday, Tucker Carlson encourage viewers to question the effectiveness of Covid-19 vaccines because the experts who recommend getting vaccinated are not trustworthy: "There are a lot of those people giving you medical advice on television, and you should ignore them. The advice they're giving you isn't designed to help, it's designed to make you comply. And you shouldn't comply mindlessly."

Carlson showed evidence that vaccines are not effective: UK public health official Sir Patrick Vallance stated that 60% of hospital admissions in the country were among the vaccinated. Carlson suggested that the vaccines don't prevent death from COVID: "According to the CDC, thousands of vaccinated Americans have died so far of COVID-19. And about 13,000 more the vaccinated have been hospitalized with life-threatening COVID symptoms."

Currently, more than 3,000 people have died from the COVID-19 vaccines in the U.S. because of side effects like thrombus and cardiac infarction. "Perhaps the actual number is vastly higher than that," Carlson said, pointing to a report submitted to the Department of Health and Human Services in 2010 concluding that "fewer than 1% of vaccine adverse events are reported by the VAERS system".

Moreover, Carlson stated that the trustworthiness of the Covid-19 booster shot is unfounded. Commenting on the Biden administration's announcement of a plan to give additional shots of COVID vaccine - so-called boosters - to millions of Americans, Carlson pointed out that in fact, scientists at the CDC disagreed about the value of giving booster shots.

“I Believe in the Science of Vaccination”: Hannity Urges Viewers to “Please Take COVID Seriously”

Fox News host on the July 19 edition of his show continued to encourage viewers to believe in the efficacy of COVID-19 vaccines Fox News

- **Sean Hannity on Monday urged viewers to take the pandemic seriously.**
- **With hundreds of thousands of deaths due to the pandemic, Hannity believes COVID-19 vaccines to be trustworthy.**
- **Hannity states that fatal side effect of the vaccines are quite rare, only affecting those with severe chronic disease.**

Over the course of the COVID pandemic, pro-Trump Fox News star Sean Hannity has consistently stated that COVID vaccines are trustworthy. Hannity took some time out of his broadcast Monday night to deliver a direct message to Fox News viewers, telling them to take the coronavirus pandemic “seriously” and declaring that he believes in the “science of vaccination”.

Nearly 609,000 people have died in the Covid-19 pandemic in the US but, as Hannity has pointed out, vaccination has slowed this rate. “I can't say it enough. Enough people have died. We don't need any more death,” Hannity encouraged his viewers to “do your own research”, suggesting that they should consult their doctors and medical professionals they trust what’s the right thing to do.

After highlighting the importance of medical privacy and doctor-patient confidentiality, Hannity added: "And it absolutely makes sense for many Americans to get vaccinated. I believe in science, I believe in the science of vaccination."

As for the side effects of the vaccines, Hannity said there are only very "rare cases where people have serious underlying health conditions that could be aggravated by the vaccine." Hannity told the viewers that he has been fully vaccinated and proposed people make decisions based on research and science and on their unique medical condition.

4. After Reading. So you've finished reading this passage...
Do you agree with what's stated in the article in general? (0∼100 Slider)

5. COVID AND VACCINES. (Post-Message)
Q2.1. Based on what you read, let's rethink the effectiveness of Covid vaccines. The default answer shown below was your answer to the same question before reading the message.
How effective do you think the vaccines are? (0∼100 Slider)

Q2.2. Based on what you read, if you have to pay for your own, what's your willingness to pay for another dose of Covid vaccine 6 months after you have been fully vaccinated? The default answer shown below was your answer to the same question before reading the message. (please enter a whole number in USD; e.g. put 200 if you're willing to pay $200) [display if the answer of Q0 == Yes!]

Q2.2.5. Our willingness to pay may vary based on how certain we are, please indicate on the slider how confident are you in your answer to the previous question? [display if the answer of Q0 == Yes!] (0∼100 Slider)

Q2.2. Based on what you read, If the government issues a vaccine mandate, how much do you willing to pay to be exempted from the mandate? The default answer shown below was your answer to the same question before reading the message. (please enter a whole number in USD; e.g. put 200 if you're willing to get vaccination when you're rewarded $200) [display if the answer of Q0 != Yes!]

Q2.3. Our willingness to accept may vary based on how certain we are, please indicate on the slider how confident are you in your answer to the previous question? [display if the answer of Q0 != Yes!]

Q2.3. Based on what you read, if you have to pay for your own, what's your willingness to pay for another dose of Covid vaccine 1 year after you have been fully vaccinated? The default answer shown below was your answer to the same question before reading the message. (please enter a whole number in USD) [display if the answer of Q0 == Yes!]

Q2.3.5. Our willingness to pay may vary based on how certain we are, please indicate on the slider how confident are you in your answer to the previous question? [display if the answer of Q0 == Yes!] (0∼100 Slider)

[Version 1] Q2.4. Let's think about Tucker Carlson, who currently serves as a host of FOX News Channel's flagship primetime cable news program, Tucker Carlson Tonight (weekdays 8PM/ET). The default answers shown below were your an-

swers to the same questions before reading the message.
How reliable do you think FOX News is? (0∼100 Slider)
How reliable do you think Tucker Carlson is? (0∼100 Slider)

[Version 2] Q2.4. Let's think about Sean Hannity, the host of The Sean Hannity Show, a nationally syndicated talk radio show, and also the host of a commentary program on Fox News. The default answers shown below were your answers to the same questions before reading the message.
How reliable do you think FOX News is? (0∼100 Slider)
How reliable do you think Sean Hannity is? (0∼100 Slider)

6. DEMOGRAPHIC.

Q3.1. What's your age? (please enter the nearest whole number)

Q3.2. What's your gender identity?
○ Male
○ Female
○ Others

Q3.3. What's your education level?
○ Less than High School
○ High School Diploma
○ Some College
○ Bachelor's Degree
○ Graduate Degree

Q3.4. What's your political affiliation?
○ Democratic
○ Republic
○ Independent
○ Others

Q3.4.5 Which party do you typically lean towards? [display if the answer of Q3.4 == Independent]
○ Democratic
○ Republic
○ Neither

Q3.5. On a 7-point scale, how interested in politics are you? (7-point Likert Scale)

Q3.6. To your best knowledge, what's your household income per year? (please enter the nearest whole number)

Q3.7. Approximately how many hours are you spending on social media every week? (please enter the nearest whole number)

End of Survey

Thank you very much for taking this survey! Your response has been recorded.

Please note that the news reports you read were written by the researchers, but the quotes from Carlson/Hannity are direct quotes from their shows.

Your validation code is:
$\$\{e : //Field/mTurkCode\}$

To receive payment for participating, click “Accept HIT” in the Mechanical Turk window, enter this validation code, then click “Submit”.

A2. Additional Tables and Figures

Table A1—Treatment Effect By Political Affiliation

	Dependent Variable by Panel					
	Tucker Carlson			*Sean Hannity*		
	Whole	*Dem.*	*Rep.*	*Whole*	*Dem.*	*Rep.*
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Vaccine Effectiveness						
Agree	-0.703	-2.324	-8.125**	2.459*	7.220**	1.143
	(1.877)	(3.180)	(3.291)	(1.485)	(3.565)	(2.473)
Credible	1.530	-0.190	-7.429**	0.705	7.667	0.143
	(2.060)	(4.017)	(3.338)	(2.090)	(4.878)	(2.473)
Agree × *Credible*	-0.169	2.167	6.130	2.948	-3.803	3.286
	(2.7699)	(5.148)	(3.738)	(2.288)	(5.059)	(3.310)
Constant	-1.364*	-0.476	8.000***	-1.250	-6.333*	-0.143
	(0.807)	(1.037)	(2.944)	(1.334)	(3.449)	(1.596)
Observations	171	84	43	159	86	31
R^2	0.0033	0.0033	0.1099	0.0818	0.0813	0.1186
Panel B: Messenger Credibility						
ALC	15.114***	13.981***	24.513***	13.865***	8.954***	15.479***
	(1.494)	(2.346)	(3.096)	(2.328)	(3.536)	(4.610)
Constant	-8.038***	-5.981***	-14.783***	-0.667	2.059	0.1875
	(1.176)	(1.394)	(2.256)	(2.009)	(3.197)	(3.473)
Observations	184	85	49	176	93	37
R^2	0.2186	0.1746	0.3950	0.0920	0.0337	0.1354

Note: The table displays the regression results on the belief about the effectiveness of vaccines (measured in 0∼100 Slider) and the credibility of the messenger after reading the message, separated by two message arms and political affiliations. Panel A displays the results for vaccine effectiveness and Panel B displays the results for messenger credibility with the simplified regression equation to ensure a decent number of observations.