Does More Speech Correct Falsehoods?

ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

DOES MORE SPEECH CORRECT FALSEHOODS?

Edward L. Glaeser
Cass R. Sunstein

Published in the *Journal of Legal Studies*, Vol. 43, No. 1 (2014)

Discussion Paper No. 777

06/2014

Harvard Law School
Cambridge, MA 02138

Final 12/3/13
Forthcoming Journal of Legal Studies

Does More Speech Correct Falsehoods?

Edward Glaeser* and Cass R. Sunstein**

Abstract

According to a standard principle in free speech law, the remedy for falsehoods is "more speech," not enforced silence. But empirical research demonstrates that corrections of falsehoods can actually backfire, by increasing people's commitment to their inaccurate beliefs, and that presentation of balanced information can promote polarization, thus increasing preexisting social divisions. We attempt to explain these apparently puzzling phenomena by reference to what we call Asymmetric Bayesianism: purported corrections may be taken to establish the truth of the proposition that is being denied, and the same information can have diametrically opposite effects if those who receive it have opposing antecedent convictions. In our primary model, recipients whose beliefs are buttressed by the message, or a relevant part, rationally believe that it is true, while recipients whose beliefs are at odds with that message, or a relevant part, rationally believe that the message is false (and may reflect desperation). We also show that the same information can activate radically different memories and associated convictions, thus producing polarized responses to that information, or what we call a memory boomerang. These explanations help account for the potential influence of "surprising validators." Because such validators are credible to the relevant audience, they can reduce the likelihood of Asymmetric Bayesianism, thus ensuring that corrections are persuasive and also promoting agreement.

I. Introduction

One of the underlying principles of a system of freedom of expression, and indeed of democracy itself, is that if "there be time to expose through discourse the falsehood and fallacies, to avert the evil by the process of education, the remedy to be applied is more speech, not enforced silence." (Whitney v. California, 1927). The principle seems both important and unexceptionable, but it rests on empirical assumptions that might not always hold true. Under what circumstances is "more speech" actually a remedy? Might more speech be ineffective and even counterproductive? If so, can we identify the reasons

* Fred and Eleanor Glimp Professor of Economics, Harvard University.
** Robert Walmsley University Professor, Harvard University and Harvard Law School.
For valuable comments and discussions, the authors are grateful to two anonymous referees, Eric Johnson, Sendhil Mullainathan, and participants in a workshop at the National Bureau of Economic Research. Cassie Chambers and Yueran Ma provided valuable research assistance.

and the circumstances, and perhaps specify remedies that do not simply involve "more speech"?

It is well-known that when like-minded groups deliberate, they tend to polarize, in the sense that they generally end up in a more extreme position in line with their predeliberation tendencies (Sunstein, 2008). For example, people who are inclined to believe that climate change is not occurring, and that it is some kind of hoax, are likely to become more unified, more confident, and more extreme in that belief after discussion with one another (Schkade et al., 2007). The phenomenon of *group polarization* can be explained in part by rational updating as information is exchanged among group members (Brown, 1984). We have suggested, however, that this explanation is inadequate and that group polarization occurs in part as a result of Credulous Bayesianism: Group members insufficiently adjust for idiosyncratic features of particular environments and put excessive weight on the statements of others in the face of common sources of information and unrepresentative group membership (Glaeser and Sunstein, 2009).

In light of group polarization and its underlying mechanisms, it might be thought that consistent with widespread faith in the potential power of "more speech," corrections of falsehoods would lead people to truth, and that the provision of balanced, objective information, would help to produce a consensus, and perhaps even a rational one, where it did not exist before. But a great deal of evidence suggests otherwise, at least in certain circumstances. If people begin with a strong prior conviction, a purported correction might actually backfire, leading them to an even stronger belief in that prior conviction. And if people begin with highly disparate beliefs about climate change, the provision of balanced information might increase polarization (Lord et al., 1979). No less than discussion by like-minded people, balanced information can led people to have greater confidence and conviction about their antecedent beliefs – and thus to make antecedently divided opinion even more divided than it was before. In short, balanced news can unbalance views.

These findings are puzzling as well as disturbing, and they raise empirical doubts about the beneficial effects of more speech in counteracting falsehoods and fallacies. They raise immediate questions: Why and when do corrections backfire? Why and when does a presentation of competing views increase polarization? Our principal goal in this Article is to answer these questions. Our central explanation involves the relationship between the informational signal and people's antecedent beliefs, which can lead to what we call *Asymmetric Bayesianism*. We show that if antecedent beliefs are sharply divided, the same signal (whether it is balanced information about a familiar topic, balanced information about an unfamiliar topic, or some kind of purported factual correction) may produce highly disparate responses, leading to even sharper divisions (Suen, 2004).

The mechanisms in the model involve signals that are produced in two different situations. Consider corrections that backfire. We assume that truthful messages are usually easy to send, which leads to a group of truth-tellers, genuinely seeking to correct misunderstandings; but false messages often have the biggest expected impact, which leads to a group of deceivers, trying to persuade people to believe falsehoods.

Responses are disparate because people begin with different prior beliefs and hence different degrees of skepticism about the motives of the messenger. Individuals who believe that the messenger is a truth-teller largely have their beliefs buttressed, while individuals who are skeptical think that the message is deceitful, which reinforces and even increases their skepticism.

Asymmetric Bayesianism sheds light on a pervasive difficulty. Influential intermediaries, such as general interest newspapers and magazines, often attempt to correct falsehoods and to present both sides of an issue, with the honorable goal of informing, and not merely reinforcing, people's opinions. If the consequence of such efforts is merely to increase people's commitment to what they thought before, and thus to increase polarization, there is a natural question: What is the point?

If Asymmetric Bayesianism is at work, then there is a natural solution to the problem. Our principal suggestion, briefly noted in the conclusion, involves *surprising validators*. Messages need to come from sources that are seen as credible to the relevant audience and not as likely to be lying (and especially not doing so out of desperation). When balanced information produces polarization, it is in part because people credit information that is consistent with their preexisting convictions while dismissing information that is inconsistent with those convictions. But when information that is unwelcome (in the sense that it casts doubt on one's prior beliefs) comes from someone who is highly credible and difficult to dismiss, a change in view is more likely. In this respect, surprising validators can overcome Asymmetric Bayesianism.

II. Backfiring Corrections and Increased Polarization

A diverse collection of studies, from multiple areas, attests to the possibility that corrections of falsehoods may backfire, and that balanced informational signals sometimes intensify polarization.

A. Backfiring Corrections

Suppose that a society is divided on some proposition. The first group believes A and the second group believes not-A. Suppose that the first group is correct. Suppose finally that truthful information is provided, not from members of the first group but from some independent source, in support of A. It would be reasonable to suppose that the second group would come to believe A. But in important settings, the opposite happens. The second group continues to believe not-A, and even more firmly than before. The result of the correction is to increase polarization.

In a relevant experiment (Nyhan and Reifler, 2010), people were exposed to a mock news article in which President George W. Bush defended the Iraq war, in part by suggesting (as President Bush in fact did) that there "was a risk, a real risk, that Saddam Hussein would pass weapons or materials or information to terrorist networks." After reading this article, they read about the Duelfer Report, which documented the lack of weapons of mass destruction in Iraq. Subjects were then asked to state their agreement,

on a five-point scale (from "strongly agree" to "strongly disagree") with the statement that Iraq "had an active weapons of mass destruction program, the ability to produce these weapons, and large stockpiles of WMD."

The effect of the correction greatly varied by political ideology. For very liberal subjects, there was a modest shift in favor of disagreement with this statement; the shift was not significant, because very liberal subjects already tended to disagree with it. But for those who characterized themselves as conservative, there was a statistically significant shift in the direction of *agreeing* with the statement. "In other words, the correction backfired – conservatives who received a correction telling them that Iraq did not have WMD were more likely to believe that Iraq had WMD than those in the control condition." (Nyhan and Reifler, 2010) It follows that the correction had a polarizing effect; it divided people more sharply, on the issue at hand, than they had been divided before.

An independent study confirmed the more general effect. People were asked to evaluate the proposition that cutting taxes is so effective in stimulating economic growth that it actually increases government revenue. They were then asked to read a correction. The correction actually increased people's commitments to the proposition in question. "Conservatives presented with evidence that tax cuts do not increase government revenues ended up believing this claim more fervently than those who did not receive a correction" (Nyhan and Reifler, 2010).

Or consider a test of whether apparently credible media corrections alter the belief, supported and pressed by former Alaska Governor Sarah Palin, that the Affordable Care Act would create "death panels" (Nyhan et al., 2013). Among those who viewed Palin favorably but had limited political knowledge, the correction succeeded; it also succeeded among those who views Palin unfavorably. But the correction actually backfired among Palin supporters with a high degree of political knowledge. After receiving the correction, they became *more* likely to believe that the Affordable Care Act contained death panels.

Liberals are hardly immune to this effect (Nyhan et al., 2013). In 2005, many liberals wrongly believed that President George W. Bush had imposed a ban on stem cell research. Presented with a correction from the New York Times or FoxNews.com, liberals generally continued to believe what they did before. By contrast, conservatives accepted the correction. Hence the correction produced an increase in polarization. Importantly but not surprisingly, it mattered, in terms of the basic effect, whether the correction came from the New York Times or Fox News: Conservatives distrusted the former more, and liberals distrusted the latter more. Source credibility is important – a point to which we will return.

B. The Effect of Balanced Presentations

The findings just described are relatively new, but for over three decades, it has been well-known that information might not produce consensus, even if it is balanced and

appears directly to address the concerns that led to divided views in the first place. The underlying phenomenon is usually described as *biased assimilation* (Lord et al., 1979 Monro et al., 2002; McHoskey 2002; Sharot et al., 2012). The basic idea is that people assimilate information in a way that is skewed in the direction of support for their antecedent beliefs (Munro and Ditto, 1997).

The initial studies involved capital punishment (Lord et al., 1979). People were asked to read several studies arguing both in favor of and against the view that capital punishment has deterrent effects. A key finding was that both supporters and opponents of the death penalty were far more convinced by the studies supporting their own beliefs than by those challenging them. After reading the opposing studies, both sides reported that their beliefs had shifted toward a stronger commitment to what they originally thought. One consequence is that the two sides were more polarized than they were before they began to read.

Similar findings have been made in many contexts (Lord et al., 1979). In one experiment, both confirming and disconfirming information was provided on the questions whether sexual orientation has a genetic component and whether same-sex couples are likely to be good parents. After receiving that information, people's preexisting beliefs were strengthened, and there was greater, not less, polarization on those questions (Lord et al., 1979). In studies of this kind, people are provided with "pro" and "con" arguments, and at least under certain conditions, provision of such arguments leads to an increase in polarization, even on questions of fact (Hardistey et al., 2010).

These findings raise an obvious question: What if the underlying issue is not familiar? In that event, will balanced information produce polarization or instead consensus? A measure of agreement might well be expected, if only because people do not begin with strong antecedent convictions. A study of nanotechnology attempted to answer these questions (Kahan et al., 2007).

A large set of Americans was divided into two groups. In the "no information" condition, people were simply told that nanotechnology is a process for producing and manipulating small particles. In that condition, people did not divide about nanotechnology. Apparently the issue seemed highly technical, and the mere name and description did not split people along any relevant lines (Kahan et al., 2007).

In "information exposed" condition, people were given factual material on the potential risks and benefits of nanotechnology. Exposure to such information sharply divided people in accordance with their preexisting political orientations. Those who tended to like free markets, and to distrust government interference, ended up far more favorably disposed toward nanotechnology. Those who tended to favor social equality, and to trust government to promote social goals, ended up far less favorably disposed. In the no-information condition, there was essentially no division between the two groups in their belief that the benefits of nanotechnology outweighed the risks. By a small majority (61 percent), both groups tended to accept that belief. But after exposure to balanced information, the split grew quite dramatically, from 0 to 68 percent, with 86 percent of

free marketeers believing that the benefits outweighed the costs, and only 23 percent of egalitarians so believing (Kahan et al., 2007).

C. Existing Explanations

Within the social science literature, the prevailing explanation of these findings points to biased assimilation. But that phenomenon itself requires explanation. Psychologists have emphasized the importance of *motivated reasoning*, which suggests that people will credit information that they like and refuse to believe information that they dislike (Molden and Higgens, 2013). This explanation points to the significant role of the emotions in producing biased assimilation. There is a close relationship between motivated reasoning and reduction of cognitive dissonance. Confronted with information that produces dissonance, people tend to ignore it (Eil and Rao, 2011; Sharot et al., 2012), a finding that might be explained in affective terms.

A competing or supplementary explanation, also found within the psychological literature, is purely cognitive (Vallone et al., 1985). The central idea is that people have different prior knowledge, and they process new information in light of that knowledge. For example, those who have reason to believe that capital punishment has no deterrent value might well disregard arguments to the opposite effect, simply because those arguments are not credible in light of what they currently know. Similarly, corrections will fall on deaf ears, because those with a great deal of prior information "know" that those corrections cannot be correct. But within the existing literature, the cognitive account has not been much elaborated as an explanation for biased assimilation.

Our central goal in this Article is to elaborate, specify, and and formalize that argument. We show that without resort to motivated reasoning, it is possible to give a plausible and unified account of why corrections backfire, and of why balanced information can produce polarization. Our account is detailed and somewhat technical, but in brief, our primary model explores the interaction between informational signals and people's antecedent beliefs. The intuition here is straightforward. Suppose that there is a report of corrupt behavior on the part of a high-level public official. If the official denies the report in the strongest and most detailed terms, his supporters, given their antecedent convictions, might well be convinced and hence dismiss the report, whereas his critics, given their different antecedent convictions, might believe that the denial is further indication that the report is true, and perhaps even a form of proof. The disparate reactions represent a form of rational updating on the basis of prior convictions. We believe that a mechanism of this kind, which we formalize here, helps to explain the findings that we have sketched.

Of course we do not deny the possibility that emotions may play a significant role. Existing work does not offer a clean test of the respective roles of motivated reasoning and the more cognitive account that we offer here. Our claim is only that the latter account fits the data and offers a parsimonious explanation of some longstanding puzzles. Empirical work would be necessary to establish the precise role of that explanation in different contexts.

III. Asymmetric Bayesianism

Our analysis considers reasons why different individuals can respond differently to the same signal, so that a common signal can induce a widening, rather than a narrowing, of belief differences. We recognize that this widening is not universal and indeed that in most cases, common signals do induce beliefs to converge. Yet as our previous discussion suggests, there are important settings where common signals appear to push views further apart. Our examples are not meant to be merely existence proofs, showing that such divergence can exist under reasonable assumptions, but rather guides to further empirical work. The assumptions and added implications of the models provide implications that might guide such work.

The common structure of our analysis is that individuals with different prior beliefs get exposed to a common message. The message is uniformly simple, but to trigger divergence, it must generate some added knowledge that differs across the population of listeners. We first focus on inference about the communicator's private information—the possibility that more desperate communicators might be more prone to provide fake information. In the second case, the message triggers memories that differ across listeners. We will refer to the first case as Asymmetric Bayesianism and the second case as Memory Boomerang.

The two structures are necessary to deal with the two types of evidence that we have already discussed. The Asymmetric Bayesianism theory is suited to cases where a leader presents a piece of information that appears to favor one conclusion (there were WMDs in Iraq). The Memory Boomerang theory can explain that phenomenon, but also casts light on cases where balanced information is put forward by a neutral party. One primary difference between the two theories is that the identity of the communicator will have a major impact on inference in the Asymmetric Bayesianism model, but not in the case of "Memory Boomerangs." Tests that examine the impact of communicator identity provide a possible tool for testing between the two theories.

In both cases, listeners are imperfectly rational and carry different information sets into the game. The core assumption of Asymmetric Bayesianism is that listeners have different views about the motives of a decision-maker, which may be shared by the communicator. The core assumption of Memory Boomerang is that individuals have different stock of pre-game signals, some of which they have forgotten, and that some of these are recalled as a result of the new signal (Mullainathan, 2002).

In all cases, there is some underlying state of the world that is either 1 (good) or 0 (bad). We will focus on the pre and post signal beliefs about the state of the world. The structure of the game always includes a pre-game, and this differs between the two settings. We begin with Asymmetric Bayesianism.

A. The Basic Phenomenon

We will focus on the response of listeners to a message, but as we have emphasized, that response depends on prior knowledge and beliefs. In this formal discussion, we describe this prior knowledge as reflecting beliefs about a pre-game, during which a leader first acquires information that convinces the leader whether or not to undertake an action (going to war, introducing a product). The leader's information may reflect myriad, diverse facts but it can be distilled down to a single probability that the world is good, which is denoted π.

The outside public has less information than the decision-maker, and we go so far as to assume that the only information about the state of the world comes from beliefs about the unconditional distribution of π and inference based on the decision-makers action and signal.[1]

The leader's welfare from undertaking the action is assumed to be $A + B \cdot \pi$, where A is the general preference for the action and B represents an extra benefit that occurs if the state of the world is good. In the Iraq context, A would represent the potential benefits or costs to the leader of going to war whether or not there were weapons of mass destruction, and B represents the benefits of going to war if there were weapons of mass destruction. In the product context, A would represent the costs of developing the product, and B represents the benefit if it is a success. The leader therefore only undertakes the action if $\pi > -A/B$, and we will only consider settings in which the action has taken place.

Individuals only differ in their assessment of A/B, and as such "optimists" believe that $-A/B$ is large, so that the leader would have only chosen the action if the state of world is good, i.e. the President would only have invaded Iraq if he believed that there was a high probability that the country had weapons of mass-destruction. We assume that individuals assume that they know A/B with perfect certainty and do not update their views about the motives of the leader. This assumption is the primary form of limited rationality in the model. While agreeing to disagree is incompatible with standard Bayesian assumptions, it is a fairly universal phenomenon, especially in political contexts (Aumann, 1976). While we have micro-founded the disagreements based on differing views about the policymaker's motives, any differences of opinion that led to different views about the possible range of states of the world can generate similar results.

Our approach in this paper is to hew as closely to standard Bayesianism as possible, and still deliver a model explaining the increased divergence of beliefs discussed in experiments above. Allowing individuals to agree to disagree is a bare

[1] The core results of the model would weaken if the leaders' informational monopoly became weaker, but the basic phenomena could still persist as long as the leader has some information that was only revealed through actions and signaling.

minimum, because without it, the model couldn't even allow individuals to have different opinions before the external message.

The listeners may be exposed to a message from a communicator, who may or may not be a representative of the leader. This message is assumed to be simple but at least somewhat costly, especially if the state of the world is not good. We have in mind a report that combines hard facts with an implied interpretation, such as a Defense Department report allegedly confirming the presence of weapons of mass destruction in Iraq.

With probability γ, the communicator is a representative of the "leader" or "connected communicator" and with probability $1-\gamma$, the communicator's only objective is to send a message. If $\gamma = 1$ or $\gamma = 0$, then the communicators' identity is obvious, and we will be able to address these cases with relatively general assumptions about the costs of communications and the distribution of beliefs. In both cases, the costs of creating and sending messages are the same. With probability ϑ, an independent communicator receives no information and with probability $1-\vartheta$, the communicator has acquired information identical to that possessed by the leader.

Communicators of all types have the possibility of sending a simple binary message to listener who represents the subjects in a game. The message is meant to convey the idea that the state of the world is good, and might take the form of military intelligence suggesting that they have found weapons of mass destruction in Iraq. The difference between the communicators lies in their objective functions. Connected communicators are interested in persuading the larger public; independent communicators simply speak the truth.

If the communicator's preferences are tied to the leader, then the benefit is assumed to be tied to the change in views associated with the message. In particular, we assume that communicators tied to the leader want a wider share of the population to believe that the state of the world is good, and hence that the action was broadly desirable. The message will create a change in aggregative opinion that generates Δ worth of value to the communicator relative to not hearing a message, which we will evaluate later, after describing the beliefs of the two groups conditional upon hearing the message.

But with probability $\varphi\pi$, the state of the world will be revealed to be good before listeners take any actions, and as such, the message will have no impact on any actions taken by the listeners. The state of the world cannot be revealed to be bad.[2] We have assumed an asymmetry which seems natural in settings like the presence of weapons of mass destruction. If there were such weapons, then with some probability it is reasonable

[2] This assumption is made for simplicity and it is not strictly necessary. Revelation that the state of the world is bad is compatible with belief backlash, as long as the returns to sending the signal decrease, at least over some range, with the leaders' belief that the state of the world is good.

to expect that the public will learn that fact. If there are no weapons then it is hard to imagine compelling evidence that will force everyone to that conclusion.

With probability $1 - \varphi\pi$, there will be no revelation, and the population will base their opinions on two pieces of information (1) that the communicator undertook the action to begin with and (2) that the communicator chose to send a message. Our assumptions now ensure that the returns of revelation to the leader are higher when the state of the world is bad, because it is less likely that the truth will come out anywhere. This assumption is crucial to belief backlash, for if there was both a lower cost and a higher benefit from sending a positive if the state of the world was good, then it is impossible for listeners to infer from the message that the state of the world is bad. We will later prove that divergent beliefs are impossible if $\varphi = 0$ which formalizes the point. For the purposes of generating belief divergence, it would be enough to assume anything that generates a higher return from messaging is that probability of the good state is lower, and this revelation structure ensures that will happen.

When the identity of the communicator is itself uncertain, then the message will also provide information about that identity and in that case, we will work with a particularly simple structure for information and signals. In all cases, the communicator has the same information as the leader, i.e. shares the leaders knowledge of π, but the only possible signal is to convey the simple message that the state of the world is good.

To recap, the timing of the model is that (1) the leaders acquires information about the state of the world, (2) the leader decides whether or not to take an action, (3) an individual has the opportunity of communicating a message to the wider public, and that individual may either share the leaders' objectives or be disinterested (in both cases, the communicator either shares the leader's information or has no information), (4) the communicator then chooses whether or not to reveal a signal to the public, (5) public revelation of the state of the world may occur, and (6) the public responds to the signal. We don't specifically model what actions the public takes, which can include buying the product or re-electing the leader, but we assume that the leader will benefit if a wider fraction of the public believes that the state of the world is good.

A more general treatment allows a full distribution of π and in that case, we assume that the cost of sending the signal is $c(1 - \pi)$, where $c'(1 - \pi) > 0$. This assumption is crucial, for the model's core result, and it is worth discussing it at length. The most natural justification is that the communicator's belief is itself a reflection of a collection of private observations—some of these which are easy to transmit and some of which are hard to transmit. If π is high, then presumably the communicator has a robust stock of real data showing that the state of the world is good. In that case, communicating simply involves releasing a bit of that information to the public, which is presumably pretty cheap. If π is low, then the communicator will have no such stock of information, and as a real, signaling will require the manufacture of false information, which is presumably somewhat more costly.

A complementary view is that higher values of π reflect a clearer signal about the real world. That clarity makes it easier to fashion a plausible public signal than if the communicator only has access to murky private signals. Even if the signal is basically false, then more accurate supportive information will presumably make the signal more plausible.

Our more specific treatment assumes that the leader and communicator have each potentially received a signal that may be good or bad. If the state of the world is good then with probability $\frac{\lambda}{1+\lambda+\sigma}$, there is a good signal, with probability $\frac{1}{1+\lambda+\sigma}$ there is a bad signal and with probability $\frac{\sigma}{1+\lambda+\sigma}$, there is no signal. If the state of the world is bad then with probability $\frac{\lambda+.5\psi}{1+\lambda+\sigma}$, there is a bad signal, with probability $\frac{1+.5\psi}{1+\lambda+\sigma}$ there is a good signal and with probability $\frac{\sigma-\psi}{1+\lambda+\sigma}$, there is no signal. We assume that $\lambda > \frac{\sigma(1+.5\psi)}{\sigma-\psi}$, and that $\sigma > \psi$.

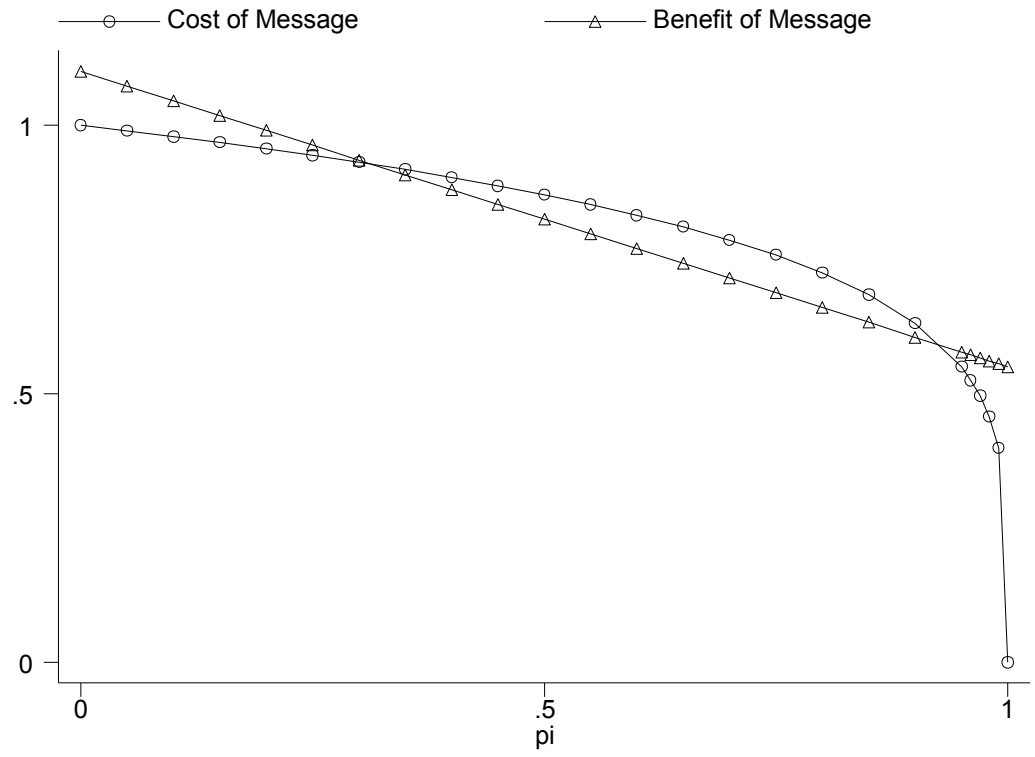
In the case of these binary signals, we assume that the cost of transmitting information is $\underline{c}$ if the communicator has received a good signal and $\overline{c} > \underline{c}$, if the communicator has not received a signal or received a bad signal. Those costs are the same because in either case, the communicator is essentially creating a false signal. One extreme possibility is that $\underline{c} = 0$—if a signal exists it is free to communicate—and that $\overline{c}$ reflects the costs of forging a signal.

We assume that independent communicators receives a fixed benefit from sending the message and will only broadcast a message if he has good information about the state of the world. In the binary case, this implies that a positive signal from an independent communicator also implies that the leader has received a positive signal. In the continuous case, this implies that that the value of π, the probability known to the leader and the communicator is above some threshold level, which we denote π_{Ind}.[3]

The behavior rules for the non-independent communicators are somewhat more complicated, because they will depend on the change in beliefs generated by their communication. We will first treat this change, and the associated benefit to the communicator, as being exogenous. We will then endogenize those benefits in the next subsection. The signal will create a change in aggregative opinion that generates Δ worth of value to the non-independent communicator relative to not hearing a signal, which we will evaluate later, after describing the beliefs of the two groups conditional upon hearing the signal. As such, an interested communicator will send a signal if and only if the benefits of sending the signal ($(1-\varphi\pi)\Delta$) are greater than the costs ($c(1-\pi)$). We are particularly interested in the case where the benefits of sending a message when

[3] This behavior can easily be justified as the result of maximizing behavior on the part of the independent communicator. For example, if that communicator has some fixed benefit of sending a broadcast, then in the binary structure, we are assuming that $\overline{c} > benefit > \underline{c}$. In the continuous case, the threshold value of π will satisfy $c(1-\pi_{Ind}) = benefit$.

$\pi = 1$ (i.e. the world is known to be good) are greater than the costs, when that also holds when $\pi = 0$ and when there is also at least some value of π at which the costs of sending a message exceed the benefits. The next figure plots functions for which this might hold.[4]



In this case, two types of interested communicators would send messages—those who have very good signals and those who have very bad signals. The communicators with very good signals will communicate because the cost is so low—since they have a lot of positive information about the state of the world. The communicators with very bad signals will communicate even though the cost is higher because the benefit of signaling is higher for lower levels of π, since they know that it is very unlikely that a good state will be revealed before decision-making needs to occur.

To prove a more general theorem, still treating the benefits of persuasion (Δ) as a fixed parameter rather than as the product of people's beliefs, we assume that costs are twice continuously differentiable c$(1-\pi)$ is twice continuously differentiable increasing and concave function which goes to zero when the world is known to be good ($\pi = 1$) and some fixed constant ($\overline{c}$) when the world is known to be bad ($\pi = 0$). Moreover, c'(1)=0.

Proposition 1: If the benefits of persuasion (Δ) is less than the maximum cost of communicating ($\overline{c}$), then messaging by an "connected communicator" will be optimal if

[4] In this case, the cost function is $(1-\pi)^{.2}$ and $\Delta = 1.1$ and $\varphi = .5$.

and only if the state of the world is sufficiently good. If the benefits of persuasion (Δ) is greater than the maximum cost of communicating ($\overline{c}$), but close to that amount (in a way formalized in the appendix), then sending a message will be optimal when the state of the world is very good or very bad, but when the state of the world is intermediate. The upper limit on the range of beliefs for which sending a message is optimal will fall with the benefits of persuasion (Δ) and rise with the connection between the state of the world and the probability that the truth is automatically revealed (φ). The impact of those same two variables on the lower limit on the range of beliefs for which sending a message is optimal is reversed. If the benefits of persuasion are sufficiently high, then it is always optimal to send a message.

This proposition illustrates that there are three regions of outcomes based on the benefits of persuasion (Δ), given our assumptions about the shape of the cost function. If the benefits of persuasion are sufficiently low, then the message will only be sent when the costs of the message are extremely low, which is when the probability that the state of the world is good is quite high. This low benefits region essentially implies something akin to truth-telling, since messages are only sent when it is quite likely that they will prove to be accurate.

When the benefits of persuasion (Δ) are greater than the maximal costs of persuasion ($\overline{c}$), then it is also optimal to send the message when $\pi = 0$, because the rewards from sending the message have become high enough to justify the costs. The assumption that as π goes to zero, the derivative of costs with respect to π also goes to zero implies that the benefits of sending the message will initially fall with π more sharply than the costs. This assumption essentially means that a complete lie is about as costly as an almost complete lie. As long as the benefits of persuasion (Δ) are not too much higher than the maximum costs of sending a message ($\overline{c}$), then it will not be optimal to send the message for some intermediate values of π. However, when π is sufficiently high, it will again be optimal to send the message.

As Δ rises, these two regions, one essentially truth-telling and the other essentially deceiving, when it is optimal to send a message also expand, and at some value of Δ (which happens must be less than $\overline{c}/(1-\varphi)$) it becomes optimal for all communicators. When the benefits of persuasion are low, then only truthful signals are sent. When the benefits take on an intermediate level, then there is mixture of truth-telling and highly misleading messaging. When the benefits are extremely high, then any interested communicator messages.

Belief reversals can never occur when $0 = \varphi$, because in that case, messages are always sent only by leaders with better information about the state of the world. In reality, the benefits of sending information may not decline with the probability that the state of the world is good, but in that case, the model predicts that there will be no belief backlashes.

Before making the specific assumptions that enable us to endogenize the benefits of persuasion, we consider the impact of a signal on respondents' beliefs, assuming that

that interested communicators, with either very good or very bad information both decide to send messages. We also assume that respondents start with a prior distribution over communicators beliefs with density $f(\pi)$ and cumulative distribution $F(\pi)$.

For the next proposition, we assume that we are in a setting where connected communicators only withhold messages if their beliefs are in the range $[\pi^{-}_{cross}, \pi^{+}_{cross}]$. Sufficiently optimistic or pessimistic communicators both send messages, as discussed in Proposition 1. Disinterested communicators only send messages if π is above some minimum threshold π_{Ind}.

Proposition 2: The message will have no impact on individuals for whom $-A/B \geq Max[\pi^{+}_{cross}, \pi_{Ind}]$. If $\pi^{+}_{cross} > -\frac{A}{B} > \pi^{-}_{cross}$, then the belief that the world is good will increase after seeing the message, and this belief will increase more as γ rises if and only if $\pi^{+}_{cross} > \pi_{Ind}$. If $\pi^{-}_{cross} > -\frac{A}{B}$, then if γ =1, then ex post beliefs are lower than ex ante beliefs if and only if ex ante beliefs are lower than the expectation of π if π lies between π^{-}_{cross} and π^{+}_{cross}. If that is the case, then ex post beliefs are always falling with γ. A message known to be from a disinterested communicator cannot decrease ex post optimism, and will always increase ex post optimism if $-\frac{A}{B}$ is less than π_{Ind}.

Proposition 2 indicates the ways in which beliefs can differentially respond to a message. Those listeners who have a great deal of faith in the motives of the leader are essentially immune to the effects of the message. They were confident that the world was good before the signal and are neither more nor less confident after the message, because they were sure that anyone given the ability to signal would do so.

Individuals who are slightly less confident, but are still generally optimistic will see their beliefs reinforced by the message. In this case, it is unclear whether a more independent message will have more or less effect on their posterior beliefs. If π^{+}_{cross} is greater than the threshold level for disinterested communicators, then interested communicators will be somewhat more selective in sending their messages, so an increased belief that the message came from an interested communicator will actually increase posterior optimism more. If π^{+}_{cross} is less than the disinterest communicators threshold, then it is the disinterested communicators who are choosier, and hence a belief that the news is independent will have a more positive effect on posterior beliefs.

Individuals with a sufficiently pessimistic starting point may see their pessimism grow as a result of the message, at least if they know that the message comes from an interested party. The key condition for this divergence to occur is that ex ante beliefs are lower than the expectation of πif π lies between π^{-}_{cross} and π^{+}_{cross}, which would be the posterior belief for a listener who knows that an interested party chose not to send a message. In other words, the message from an interested party will reduce posterior views if the decision to send a message itself is seen as a sign of desperation rather than good knowledge.

If this condition holds, so that a message from an interested party depresses beliefs, then posterior beliefs will be rising with the probability that the message comes from an outsider. Therefore, a higher probability that the signal comes from an outsider will reduce the probability that the signal will lead to divergence. This fact motivates our later discussion of surprising validators. While it is possible that a surprising validator may have less positive impact than an interested communicator, a surprising validator cannot reduce optimism, and will typically increase optimism if the listener is initially highly skeptical.

B. Endogenous Benefits of Sending a Message

We now endogenize Δ by assuming that there are two groups in the population that differ in their assessment of $-A/B$. One group, which we refer to as optimists believe that $-\frac{A}{B} = \pi_o$, while the second group (pessimists) believe that $-\frac{A}{B} = \pi_P$, where $\pi_p < \pi_o$. In a war-related context, the optimists believe that the President had little interest in starting a war unless weapons of mass-destruction was real which means that $-\frac{A}{B}$ is high. Pessimists think that the President is interested in fighting whether or not there are weapons of mass destruction and hence $-\frac{A}{B}$ is low.

We assume that the benefit of communicating for connected communicators is $v_O(Change\ in\ optimists'\ beliefs) + v_P(Change\ in\ pessimists'\ beliefs)$, where the change in beliefs reflect the difference between sending and not sending a message. The term v captures both the value of changing one person's beliefs and the number of people impacted. We first consider the case where communicators are exposed only to a binary signal, again assuming that independent communicators will send a message if and only if they have received a positive signal. We use the term message to refer to the transmission between communicator and listener and the term signal to refer to information initially received by the communicator. We will also assume that optimists and pessimists have material differences in beliefs. One possibility is that optimists believe that the leader would not have taken the action unless he received a positive signal and that pessimists believe that the action would have been undertaken at least in the case where there was no signal. A second possibility is that optimists believe that the leader also would have taken the action if he had received no signal and that pessimists believe that the leader would have taken the action in all situations.

We also assume that $\underline{c}$ is sufficiently low so that interested communicators who have positive signals will always send messages as long as these messages have a positive impact on beliefs. In our first appendix, we characterize the equilibria of the game. The behavior of independent communicators is always the same, but the actions of connected communicators depends on the value of $\overline{c}$, the costs of sending what is essentially a false message. When those costs are sufficiently high, then it only connected communicators who have received a positive signal send a message. In that case, there is essentially truth-telling among both independent and connected communicators.

For intermediate levels of $\overline{c}$, then connected communicators who have been exposed to negative signals also chose to send message. We assumed—critically—that the benefits of sending a rosy message are higher if the state of the world is worse. In our model, this came from the probability of independent revelation (which eliminates the advantage of sending a message) rising with the probability that the state of the world is good. This feature of the model could be duplicated by any force that pushes the desperate into deception. Finally, when the costs of $\overline{c}$ are extremely low, then all connected communicators send singals.

For some ranges of parameter values, there is the possibility of multiple equilibria, which occur because the benefits of sending a message depend on who is sending a message. So it is possible that for the same parameter values either (a) all connected communicators send messages or (b) only connected communicators with good or bad messages. In case (a), the benefits from sending a message are higher and that induces even those connected communicators who have not received a signal to send a message. Since our focus lies on belief divergence, not the characterization of the equilibrium itself, we leave these details to the appendix.

We are now ready to describe the parameter ranges for which a belief reversal occurs.

Proposition 3: A belief reversal where messages make pessimists more pessimistic can only occur if optimists are uncertain about the state of the world, and if those optimists are made more optimistic by the signal. A belief reversal can occur if and only if $\frac{\gamma}{(1-\gamma)(1-\vartheta)} > \frac{\lambda-1-.5\psi}{\psi} > \frac{\lambda}{\sigma}$ and if $\overline{c}$ lies in the segment $\left[\overline{c}^{-}, \overline{c}^{+}\right]$, where $\overline{c}^{-}$ and $\overline{c}^{+}$ are defined in the appendix as functions of the parameters, $\varphi, \lambda, \sigma, \vartheta, \psi, v_0$ and v_P. If $\overline{c}^{-} \leq \overline{c} \leq \overline{c}^{+}$, then the negative effect of the message on beliefs will be decreasing with $\overline{c}$, φ and $v_P$and rising with v_0.

Proposition 3 explains that a belief reversal cannot happen whenever optimists are sufficiently sure about the state of the world. Connected communicators don't want to make people more pessimistic, so if a belief reversal occurs, this will push them not to send a signal in the first place. Unless there is the positive effect of the message on optimists that outweighs the negative effect of the message on pessimists, the message will not be sent. The only way that a belief reversal is compatible with sensible strategic communication behavior is that there are enough people who react positively to the signal to offset those who become more pessimistic. . Whenever optimists are sure about the state of the world, because they believe that the leader has to have seen a good signal, then there is no chance of a belief reversal among pessimists. As such, divergence is not so much a paradox as a necessary condition for messages to be sent in equilibrium, when one group reacts negatively to them.

The proposition shows that if optimists and pessimists are both uncertain, there are two crucial conditions for a preference reversal that causes the identical message to pull optimists and pessimists further apart. The first condition is that $\frac{\gamma}{(1-\gamma)(1-\vartheta)} >$

$\frac{\lambda-1-.5\psi}{\psi} > \frac{\lambda}{\sigma}$, which essentially means that ψ, which determines whether the extent to which no news is good news, must be low, but not too low. The second half of the inequality can be written as $\lambda > \frac{1+.5\psi}{\sigma-\psi}$, and we have assumed that this always holds. This condition implies merely that the state of the world is more likely to be good conditional upon hearing a good signal than conditional upon hearing no signal.

The more important condition is the first part of the inequality which will be easier to meet when γ or ϑ or both are close to one. As such, belief reversals cannot occur when information is sufficiently likely to come from an independent source—high probabilities of connection are critical. That first inequality is also more likely to hold when λ is lower, so the initial signal can't be too informative, or when ψ is higher, which means that no news is more likely to mean bad news. When signals are more likely to be independent, or sent by a surprising validator, then $\frac{\gamma}{(1-\gamma)(1-\vartheta)} < \frac{\lambda}{\sigma}$ and a reversal is impossible.

An added requirement for a reversal is that the cost of transmitting a somewhat inaccurate message is neither too low nor too high. Intermediate costs of transmission are necessary for belief reversals, since they require messages to be sent by connected communicators with bad signals, but not connected communicators with no signals. As long as costs are intermediate then we have the possibility for belief reversal.

In principle, these comparative statics provide a series of implications that can be tested. If messages are thought to be very expensive or very dear, then belief reversals shouldn't occur. If messages are more likely to come from independent sources, and if those independent sources are particularly likely to have real information, then belief reversals will also be impossible.

Proposition 3 delivers the main result of this section: *divergence is possible as long as different groups start with different assessments about the character of the decision-maker*. Moreover, the pessimists must believe that the signal is unlikely to have come from a disinterested communicator. Optimists, who trust the decision-maker, are always likely to increase their optimism, because they assume that the signal reflects a low cost of signaling, and a high degree of accuracy. Our technical assumption essentially forced that result.

.

IV. Memory Boomerang

In this section, we return to the binary signal structure but change the model in several critical ways. For we assume that –A/B<0 for everyone, so that all types belief start believing that the state of the world is good with probability .5 We furthermore assume that $\gamma = 0$, and $\sigma = 0$, so that there is no biased information, and that all

individuals are presented only with independent communicators who have actually received a signal. The critical difference here is that we assume that optimists and pessimists have been exposed to many past independent signals, and they may not recall all of them.

As before if the state of the world is one, then with probability $\frac{\lambda}{1+\lambda}$, independent communicators will receive information to use as a signal. With probability $\frac{1}{1+\lambda}$, they receive a bad signal if the state of the world is one and send no message. If the state of the world is zero, then independent communicators will receive a good signal with probability $\frac{1}{1+\lambda}$ and a bad signal with probability $\frac{\lambda}{1+\lambda}$. Individuals understand when a communicator has been given an opportunity to signal, but declines, and these moments are interpreted as negative signals about the event. Individuals have been given many periods of possible exposure to signals, but individuals may forget some share of their signals that they have received.

Given these assumptions, this section presents an alternative explanation for opinion divergence following new information revelation, which we call memory boomerang. In this model, we follow Mullainathan (2002). and assume that individuals have forgotten many of their past experiences, but that an intervention may cause forgotten facts to be remembered. Even if an experimental intervention favors one view of the world (i.e. there were weapons of mass destruction in Iraq), the intervention may create a memory boomerang by causing the subject to recall forgotten evidence against that worldview that is far more compelling than the experimental intervention (Hardistey et al., 2010).

An extreme example of this phenomenon would occur if individuals take in evidence and keep only a brief summary judgment of that evidence at the top of their mind. Over time, that summary judgment may weaken, perhaps because people have forgotten why they held that opinion in the first place. But in an experiment, they are exposed to information that suddenly brings back all the evidence that had been dormant in their longer-term memory and the effect of recalling that lost information overwhelms the direct impact of the experiment. If the recalled information contradicts the experimental intervention, then we say that the experiment created a memory boomerang.

To formalize these ideas, we assume individuals enter an experiment and are exposed to a new piece of data about the state of the world. We assume that before the experiment, each subject has been exposed to a history of signals about an aspect of the state of the world, such as whether there were weapons of mass destruction in Iraq. We also assume that the new signal may jolt the memory of past facts or stories, and cause a past signal to be remembered. In Mullainathan's terminology (Mullainathan, 2002), the new information is "associated" with the past information and that makes it more likely for the past information to be remembered. Humans appear more likely to remember past events if they are similar to current events.

Data takes the form of signals—a stock of events during which the signal equaled either zero or one. For simplicity of exposition, we will use the term optimism to refer to the belief that the state of the world is one and pessimism to refer to the belief that the state of the world is zero, but that is an entirely arbitrary terminology.

We assume that individuals believe that if the real state of the world is one, then the probability that any signal will equal one is $\frac{\lambda}{1+\lambda} > .5$. If the state of the world is zero, then a fraction $\frac{1}{1+\lambda} < .5$ of signals have a value of one. Individuals enter the experiment with a stock of N_1^R remembered signals with a value of one, and N_0^R remembered signals.

The experimenter provides the subject with a new signal. This experimenter-provided signal also has a value of one, but subjects don't believe that experimenter has any ulterior motives for providing this signal. They do think that the experimenter provided signal is provided by its own stochastic process, that would have generated a positive signal with probability $\frac{\lambda_E}{1+\lambda_E}$ if the state of the world is one, where $1 \leq \lambda_E$. If the state of the world is zero, then the experimenter provided signal would have been one with probability $\frac{1}{1+\lambda_E}$. The experimenter-provided signal may be more or less accurate than previous signals, but in order to get a memory boomerang, the experimenter-provided signal cannot be too accurate.

The term λ_E reflects the accuracy or precision of the signal. If $\lambda_E = 1$, then the signal is perfectly balanced, so this includes the case, discussed in the data section, where listeners are presented with balanced information. As λ_E rises to infinity, then the signal itself essentially perfectly discloses the state of the world. The signal will always serve to activate memories, regardless of its precision. However, if this new signal is itself so accurate that it reveals the world, then the memory activation, and indeed all prior information, will have little impact on final beliefs.

If there was no other effect of the signal, then exposure to the experimenter-provided signals changes the posterior positively as long as $\lambda_E > 1$. The direct effect of the signal is to increase optimism or the posterior belief that the state of the world is one.

But we allow the signal to have a secondary effect—bringing back similar lost memory of a past signal. In a previous draft (Glaeser and Sunstein, 2013), we consider two possible structures: random recollection and endogenous recollection. Here we only consider random recollection. With probably m, the presence of related information cues some other fact, once buried deep in the brain, but now brought forward by this similar event or piece of information. The probability m determines whether a forgotten fact returns but not whether that fact favors optimism or pessimism.

With random recollection we assume that the remembered signal is particularly accessible, and that with probability $(1-\delta)a + \delta\frac{N_1^R}{N_0^R+N_1^R}$, this remembered signal is positive. As long as accessibility of memories is correlated with the sign of the signal,

the bias towards accessibility will do little to impact the posteriors. The δ parameter determines the extent to which recalled signals resemble the initial remembered stock of signals. If $\delta = 0$, then the remembered signal is positive with probability $a \geq \frac{\lambda}{1+\lambda} > .5$, to reflect the fact that the new positive signal may be more likely to bring back a forgotten signal that is also positive. If $\delta = 1$, then the forgotten signal remembered after the intervention has the same probability of being recalled as the share of one-signals in the pre-intervention stock of signals.

The process of recall makes it possible that the new information will end up reinforcing past beliefs and increase the divergence of beliefs across individuals, especially if the experimenter is providing information that is balanced or not too important:

Proposition 5: As long as the new information is sufficiently uninformative (i.e. λ_E is low), then the information will on average cause the belief that the state of the world is one to increase for individuals for whom $\frac{N_1^R}{N_0^R+N_1^R} > a + \frac{1}{\delta}\left(\frac{\lambda}{\lambda+1} - a\right)$ and to decrease for individuals for whom $\frac{N_1^R}{N_0^R+N_1^R} < a - \frac{1}{\delta}\left(a - \frac{1}{\lambda+1}\right)$.

When the new information revealed is weak, individuals who begin with a strong belief that the state of the world is one ($\frac{N_1^R}{N_0^R+N_1^R}$ is high), will increase that belief. There is both the direct effect of the new information, but even when the new signal has little information value it will trigger memories that will typically increase optimism for the initially optimistic, because initially optimistic people also have a relatively optimistic stock of forgotten memories. Recalling one of those memories will only reinforce the person's belief that the state of the world is one.

But for pessimists, who begin with a low value of $\frac{N_1^R}{N_0^R+N_1^R}$, the new information can actually create a memory boomerang that moves their beliefs in the opposite direction. As long as $\frac{1}{\lambda+1} > (1-\delta)a$, which requires a high value of δ—the probability that the newly remembered signal will look like the pre-experiment stock of remembered signals—then sufficiently pessimistic people who only get more pessimistic because of the new information. The correlation between remembered signals and the stock of pre-experiment signals creates the force driving divergence. As long as pessimists are more likely to remember a forgotten pessimistic signal, then any new information, even optimistic information that jogs past memories will create the possibility of generating more pessimism.

We see this proposition as directly speaking to the experiments where balanced information has been provided. Truly balanced information presumably implies that $\lambda_E = 1$, so this will typically create divergence if the memory boomerang model is relevant. In a sense, this effect seems far more general than the asymmetric

Bayesianism discussed above, but it does provide at least one set of testable implications if the memories of the individual can either be measured or manipulated. Measurement might involve surveys of past exposure to content, such as reading practices. Manipulation might involve delivering information in a lab setting at some prior data. The model predicts that beliefs differ across people with different memories, and that the impact of those past memories will widen with exposure to balanced information.

The model also suggests other comparative statics on when we should expect to see settings where the new information creates a memory-related backlash that pushes beliefs in the opposite direction. In the online appendix, we state and prove a proposition that if $\frac{\left(\lambda+\lambda^{N_0^R-N_1^R}\right)}{(1+\lambda)\left(1+\lambda^{N_0^R-N_1^R}\right)}$ is sufficiently large, then for higher values of λ_E, the signal will increase average optimism while for lower values of λ_E, the signal will increase pessimism. Optimism ends on the perceived precision of the new information from the experimenter. When the experimenter is giving out really good data, then this is likely to increase optimism and if the data is relatively meaningless, this will increase pessimism. There is a cutoff value of precision that will determine whether the experiment generates added optimism.

The cutoff value of λ_E that determines whether the signal will increase optimism is decreasing with a and $N_0^R + N_1^R$, holding $N_0^R - N_1^R$ constant and increasing with the total stock of constant, and increasing with the exogenous probability of recalling a memory. Parameters that reduce the cutoff value should be understood as suggesting that they will also increase optimism, while parameters that increase the cutoff values should be understood as increasing the likelihood that the new information will produce a memory boomerang. Higher values of λ_E don't change the memories that are awoken, but it does reduce their importance, since the new signal itself has a stronger impact on posterior beliefs.

Increases in a and the stock of signals both decrease the cutoff value and make a memory boomerang less likely. Higher values of "a" make it more likely that the new signal will jog a positive memory and unsurprisingly that makes an increase in optimism more likely. Increases in the total number of signals, holding $N_0^R - N_1^R$ constant, pulls the share of past signals that are positive or negative closer to zero, and that makes it harder to have a high probability of pulling a negative signal out of one's past memory. This suggests that individuals with less experience are more likely to experience a memory boomerang, since they are most likely to have a really skewed set of past forgotten signals.

In the relevant parameter space, higher values of δ also typically make memory boomerang more common. The stronger the correlation between the stock of remembered signals and forgotten signals jogged by the experiment, the more likely the experiment is to create a memory boomerang, because it is more likely that people who begin as pessimists will pull a lost negative signal from their memory.

The results are relatively similar with an endogenous memory model, which we explore in a previous version of the paper (Glaeser and Sunstein, 2013) where individuals have a cost of accessing past memories. In our setting, we assume that humans like to feel as if they know the state of the world and that there are cognitive costs to being uncertain. It follows that if any new information decreases the current level of uncertainty by supporting people's current views, then that new information will in turn *decrease* the incentive to remember forgotten data. If new information increases uncertainty, then it will increase the incentive to recall other information, and potentially that recalled information can overwhelm the new data.

As a result, the impact of any information depends on the initial signals remembered by the participants, and their costs of recalling memories divided by the benefit of certainty. An optimistic signal will reassure those individuals who begin as optimists, and create no incentive to recall other information. People who are mildly pessimistic will have their beliefs reversed, then they may try to recall information that supports their changed opinion. In this case, memory will reinforce the new information and cause an overreaction. Memory boomerangs are most likely among individuals who begin as diehard pessimists. They will remain pessimists after the new information is provided to them, but the new information will make them less certain and increase the incentives to remember other pessimistic facts. If their costs of recall are low and the benefits of certainty are high, then they may recall a pessimistic memory that may make them even more pessimistic than they were before receiving the optimistic signal.

Conclusion

Corrections may backfire if people have strong antecedent convictions and understand purported corrections of those convictions as evidence that those convictions must be right. Presentation of balanced information may turn out to intensify polarization, simply because different people will believe different parts of the presentation. Assurance that all is well – that a politician is not corrupt, that a war was fought for the right reasons, that a product is safe, that an alarming rumor is baseless – will be credible to some but not to others. As a result, such assurance might well increase polarization.

Our goal here is explanatory, not prescriptive, but both the empirical findings and our explanations raise an obvious question: If the goal is to avoid polarization and to produce some kind of reasonable consensus, might anything be done?

To see a potential answer, consider a striking finding (Cohen, 2003). When liberals and conservatives are asked for their private views about a generous welfare policy and a more stringent one, they react in the predictable ways, with liberals favoring the former and conservatives the latter. But things change dramatically when they are informed of the distribution of views within the House of Representatives. More specifically, conservatives end up disapproving of the more stringent policy, and favor the generous one, when they are told that 90 percent of House Republicans favor the generous policy. Liberals show the same willingness to abandon their private opinions,

and thus end up favoring the stringent policy, when told that this is the position of 90 percent of House Democrats. Notably, the effect of learning about party views is as strong among those who are knowledgeable about welfare policy as it is among people who were not (Cohen, 2003). Also notably, both conservatives and liberals believe that their judgments are driven largely by the merits, and not by what they learn about the views of their preferred party – but in that belief, they are wrong (Cohen, 2003).

The general lesson is that surprising validators can help to counteract Asymmetric Bayesianism (Kahan et al., 2007). Consider the influence of the *convert communicator*, who once believed the opposite of his current position – say, a former member of the National Rifle Association turned gun control advocate, or a former pacifist turned strong defender of a particular war. In one study, a reformed alcoholic was found to be substantially more persuasive than a teetotaler when extolling the importance of abstaining from alcohol (Levine and Valle, 1975; Lord et al., 1984). A similar effect has been found in political context -- where, for instance, a politician taking a pro-environmental stand turned out to be more persuasive when he was perceived by audience members as generally pro-business (Eagly et al., 1978).

Surprising validators have special credibility to precisely the people who would otherwise be inclined to dismiss them. If a longtime critic of an allegedly corrupt politician rises to his defense, contending that the corruption charges are baseless, then there is little reason for the kind of polarization that we have explored. And if a pro-business speaker, typically critical of environmentalists, asserts that climate change is indeed a serious problem, the polarization that we have sketched should be less likely to take place. Surprising validators are credible, and reduce rather than create polarization, because they counteract Asymmetric Bayesianism.

References

Aumann, R. J. (1976). Agreeing to disagree. *The annals of statistics*, 1236-1239.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of personality and social psychology*, *85*(5), 808.

Eagly, A. H., Wood, W., & Chaiken, S. (1978). Causal inferences about communicators and their effect on opinion change. *Journal of Personality and Social Psychology*, *36*(4), 424.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114-138.

Glaeser, E. L., & Sunstein, C. R. (2009). Extremism and social learning. *Journal of Legal Analysis*, *1*(1), 263-324.

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? Attribute framing, political affiliation, and query theory. *Psychological Science*, *21*(1), 86-92.

Kahan, D., Slovic, P., Braman, D., Gastil, J., & Cohen, G. (2007). Affect, values, and nanotechnology risk perceptions: an experimental investigation. *GWU Legal Studies Research Paper*, (261). Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=968652 (2007)

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, *47*(6), 1231.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

McHoskey, J. W. (1995). Case closed? On the John F. Kennedy assassination: biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology*, *17*(3), 395-409.

Molden, D. C., & Higgins, E. T. (2005). Motivated thinking. *The Cambridge handbook of thinking and reasoning*, 295-317.

Mullainathan, S. (2002). A memory-based model of bounded rationality. *The Quarterly Journal of Economics*, *117*(3), 735-774.

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*(6), 636-653.

Munro, G. D., Ditto, P. H., Lockhart, L. K., Fagerlin, A., Gready, M., & Peterson, E. (2002). Biased assimilation of sociopolitical arguments: Evaluating the 1996 US presidential debate. *Basic and Applied Social Psychology*, *24*(1), 15-26.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303-330.

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical care*, *51*(2), 127-132.

Sunstein, C. R. (2009). *Going to extremes: how like minds unite and divide*. Oxford University Press.

Schkade, D., Sunstein, C. R., & Hastie, R. (2007). What happened on deliberation day?. *California Law Review*, 915-940.

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences*, *109*(42), 17058-17062.

Suen, W. (2004). The Self-Perpetuation of biased beliefs. *The Economic Journal*, *114*(495), 377-396.

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of personality and social psychology*, *49*(3), 577.

Whitney v. California, 247 U.S. 357, 377 (1927) (Brandeis, J., concurring).