October 4, 2019

Dear Harris School Political Economy Workshop Participants!

Thank you all for the chance to come present some work to you this coming Thursday. I want to discuss with you probably the shakiest portion of my book manuscript "Persuasion in Parallel." I'm including in this document the opening chapter that lays out the main empirical claim of the book (everyone responds to persuasive information by updating their views in the same direction by about the same amount). I'm *not* including here Chapters 2 - 6 that lay out all the definitions, important theoretical distinctions, and empirical evidence in favor of the claim, though if anyone is interested in those chapters, I'm ready to share them.

I am also including Chapter 7, which engages what the main finding (taking it as granted, which you are not required to do) means for models of information processing.

I'm really looking forward to discussing this with you all!

Alex

Persuasion in Parallel

Alexander Coppock

Draft, October 2019

Contents

CHAPTER 1

Persuasion in post-fact America

I married into a large and loving family whose political views differ from my own. I can't help talking about politics at holiday gatherings and I also have many nieces and nephews. It has now dawned on me: I'm a crazy uncle who ruins Thanksgiving. Our disagreements are always polite, but I'm nevertheless quite sure that everyone would have preferred we didn't disagree at all.

In our cultural mythology, it's usually a medium-close (often male) relative who derails dinner conversation into shouting matches and recriminations. What should be a celebration of love and togetherness devolves into a political argument in which one or more family members (often uncles) reveal themselves to be racist, sexist, or otherwise bigoted. Learning that members of *your own family* hold political attitudes with which you vehemently disagree is painful and talking about it makes it feel worse.

When your uncle tries to persuade you that we need to build Trump's border wall because criminal immigrants are taking over America, you are enraged at the ethno-nationalism, the blatant disregard for America's history as a nation of immigrants (and conquerors), the ridiculous notion that even if the wall were built that it would be at all effective. He supplies half-remembered statistics from Fox News about crime rates among immigrants and the supposed ease with which terrorists can infiltrate the US – your blood boils. You are totally unconvinced by the arguments and your opinion of your uncle couldn't be lower. It feels like you are further apart than when you started. Not only did he not persuade you, you're now more convinced than ever that the border wall is the racist fantasy of a lunatic president.

How does this conversation affect your political attitudes and beliefs? As I am not able (or allowed) to conduct a real experiment in which I randomly assign which of your family members visit for the holidays, let's conduct a brief thought experiment instead.

Imagine two counterfactual states of the world, one in which your uncle comes to Thanksgiving and one in which he misses it. As a "treatment," the dinner conversation is a bundle of partisan cues, possibly false or fallacious arguments, and a healthy dose of negative affect. When we do experiments that randomly assign treatments, we try to vary one feature of a treatment at a time while holding others constant; in this case everything about the conversation moves together.

In both worlds, I ask you three survey questions over dessert:

1. There has been a lot of discussion lately about immigration. On a scale from 1 to 100, do you oppose or support building a wall along the United States border with Mexico, with 1 meaning completely oppose and 100 meaning completely support?

2. Remaining on the topic of immigration, on a scale from 1 to 100, do you oppose or support increasing spending on border security measures *other than* building a wall along the US border with Mexico, with 1 meaning completely oppose and 100 meaning completely support?

3. Finally, on a scale from 1 to 100, how much do you like your uncle, with 1 meaning completely dislike and 100 meaning completely like him?

Suppose you are a typical NPR-listening progressive. In the counterfactual world in which your uncle missed Thanksgiving, your answer to the first question might be 2 out of 100. Your opposition to the border wall is very solid, but there's room to move. On the second question, you recognize the need for some smart protections, so your response is 20 out of 100. On the last question, for all his faults, he's your uncle, so you say 90 out of 100.

In the world where dinner was a disaster, your answers are a bit different. You're a solid 1 out of 100 on the border wall, that's for sure. The debate did, however, highlight some security threats that probably should be addressed, so you're a 25 out of 100 on non-wall border measures. Because you're extremely angry with your uncle right now, you say you only like him 8 out of 100, though at least some portion of that is just the rage talking.

Counterfactually speaking, the conversation causes a 1 - 2 = -1 point decrease in your support for the wall. This is a "backlash" in the sense that your uncle was advocating for the wall

but you ended up liking it less than you would have if he had said nothing. In my view, the best explanation for the backlash is a *group cue*. As a result of the treatment, the wall is associated with xenophobia, bigotry, and hatred, now more than ever. You like the wall less because you infer it must be bad if all these bad people like it. Group cues serve as a powerful heuristic wherein people figure out their policy views by learning what people who are like them (or who are unlike them) think. This book is mostly *not about* group cues.

The conversation also causes a 25 - 20 = 5 point increase in your support for border spending. The affirmative arguments about the need for increased border security (regardless of their accuracy) constitute *persuasive information*. Throughout the book, I will focus almost exclusively on the extent to which persuasive information affects policy attitudes and opinions.

Finally, the thought experiment highlights the distinction between policy attitudes and *affective evaluations* of sources or evidence. The hypothetical conversation causes a 8 - 90 = -82 decrease how much you like your uncle. Hearing people make arguments with which you disagree makes you like them less, at least a little. This book is mostly *not about* affective evaluations of the persuasion information itself.

Very crucially, this vignette underlines that persuasive information could have a positive effect on policy support but a negative effect on affective evaluations. Making someone angry doesn't mean you didn't persuade them. The difference between the persuasive effect of information on the one hand and people's affective evaluations of information on the other is a source of deep confusion, both in the popular press and the academic literature on attitude change. Sometimes, authors will say that people "reject" evidence when what they mean is that people don't like the evidence. The word reject makes it seem like the evidence didn't have any effect on a person's attitudes, or worse, actually strengthened their prior attitude. As we'll see, people are persuaded by evidence they don't like every day on issues great and small.

1. *The persuasion in parallel hypothesis*

This book makes a single argument, over and over: Persuasion occurs in parallel. Persuasion itself is the causal effect of information on attitudes. Saying that persuasion occurs in parallel means that everyone responds to persuasive information by updating their views in the same

direction and by about the same amount. While baseline political views are very different from person to person, responses to information are quite similar.

The claim that persuasion always occurs in parallel for everyone, regardless of the content or provenance of the persuasive information is obviously far too broad. I promise that important caveats and scope conditions are coming for readers who forge on. In the meantime, I want to emphasize that although there some interesting exceptions, persuasion in parallel is the norm. It happens many millions of times a day as people scroll through social media feeds and talk with friends and coworkers about the news. Politicians, journalists, pundits, academics, and advertisers are constantly barraging people with persuasive attempts on issues great and small. These attempts are probably a little bit effective for most everyone who hears them.

The amount of persuasion is usually small. Small means something like 5 percentage points or a tenth of a standard deviation in response to a treatment like an op-ed, a video advertisement, or a précis of a scientific finding. Small changes make sense. If persuasive effects were any bigger, wild swings in attitudes would be commonplace and people would be constantly changing their minds depending on the latest advertisement they saw.

Persuasive effects decay. Ten days after people encounter persuasive information, average effects are about one-third to one-half their original magnitude. After ten days, we have only limited evidence about whether they persist or fade. Whether this pattern of decay means that information has long-lasting or fleeting effects depends on your point of view about how long is long. The fact that we can detect persuasive effects days or weeks after exposure to persuasive information indicates to me that we're talking about a real phenomenon and not just some experimental artifact.

The strongest evidence for the claim that people are persuaded in parallel derives from randomized experiments in which some people are exposed to information (the treatment group) while others are not (the control group). These experiments show over and over that the average treatment effects of information are positive. By positive, I mean that they are in the direction of the information. These average effects also hold for subgroups: young and old, better and less-well educated, Republican and Democrat, black and white, male and female, they all respond in the direction of information by about the same amount.

The strong form of the persuasion in parallel thesis is that information has the exact same effect for everyone. Falsifying this thesis is trivially easy. All we would need is one statistical test that shows that effects are stronger for one group than another. The experiments described in this book offer many examples of such tests. When the average effect for Democrats is 3 percentage points and the average effect for Republicans is 5 percentage points, a sufficiently large experiment would declare these two responses "different." But these sometimes statistically significant differences are rarely politically significant.[1] In the main, responses to treatment are qualitatively speaking quite similar even across wildly diverse groups of people.

The weak form of the thesis is that *backlash* doesn't occur. Backlash (or backfire – I don't draw any distinction between the two terms) happens if information has positive effects for some but negative effects for others. Falsifying this weaker thesis would also be easy. All we would need is one statistical test that finds evidence of a positive effect for one group but a negative effect for a different group. None of the many experiments reported in this book measure any instances of backlash, but other authors have claimed to find them (Lazarsfeld, Berelson and Gaudet, 1944; Nyhan and Reifler, 2010; Zhou, 2016; e.g.,). Backlash has probably occurred, but it is definitely not the norm (Wood and Porter, 2018; Nyhan et al., 2019; Porter and Wood, 2020).

Many theories accommodate and predict backlash, including the Receive-Accept-Sample model (Zaller, 1992), the "John Q. Public" model (Lodge and Taber, 2013), and the Cultural Cognition model (Kahan, 2012). These models may be useful for explaining all sorts of phenomena, but at a minimum, their predictions that backlash is common is (in my view) incorrect.

2. *Example: Flat tax op-ed experiment*

Emily Ekins, David Kirby, and I ran an experiment to measure how much, if at all, an opinion piece in the Wall Street Journal could have changed minds on tax policy, a topic over which

[1] A difference is "statistically significant" if, supposing the true difference is equal to zero, we would be unlikely observe a difference as large or larger than the one we did observe in the revealed data. Those chances depend on a number of features of how the study is designed, but they depend most critically on the number of subjects in the experiment. As long as the true difference is not *exactly* equal to zero, the probability gets smaller and smaller as the sample size increases. Annoyingly, with very large experiments, even subtantively meaningless differences can be declared "significant."

Americans are solidly divided. Opinons about tax policy may be crystalized, which is to say that they aren't likely to move in response to new information. At worst, resistance could be so strong that any persuasive attempt would result in people becoming further entrenched in their previously-held opinions.

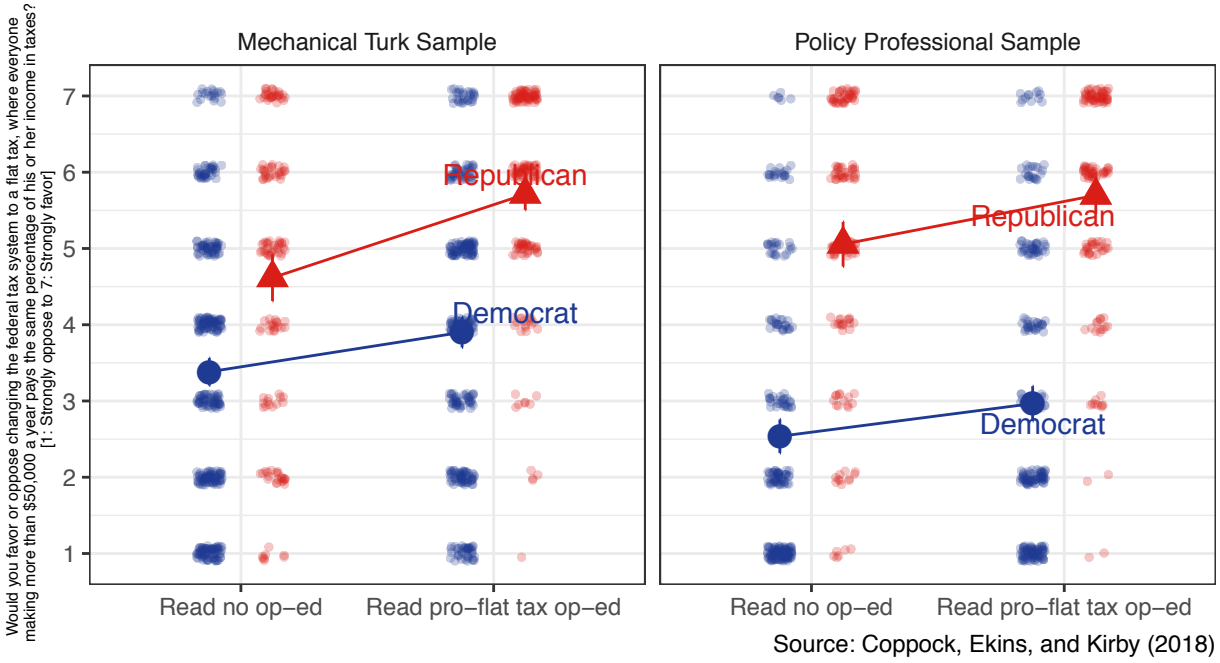
We randomly assigned subjects to treatment and control groups. The treatment group read an op-ed by Senator and then-(at the time of the experiment)-presidential hopeful Rand Paul of Kentucky. In "Blow Up the Tax Code and Start Over" (Paul, 2015), Senator Paul proposed a 14.5% flat tax that he predicted would cause the economy to "roar." Paul primarily argues for his flat tax proposal on fairness grounds. He anticipated the objections that the proposal is a giveaway to the rich (he'd close loopholes) and that the proposal would induce massive deficits (he'd balance the budget). He called the IRS a "rogue agency" and blamed Washington corruption on the convolutions of the tax code. The op-ed is 1,000 words long, makes a complicated argument, and demonizes relatively obscure bureaucrats most Americans wouldn't have heard of. It's insidery, a little punchy, and lacks strong evidence for its claims. It's the sort of thing we might think is unlikely to change many minds. Post-treatment, we asked both the treatment and control groups "Would you favor or oppose changing the federal tax system to a flat tax, where everyone making more than $50,000 a year pays the same percentage of his or her income in taxes?"

We ran this experiment twice, once with a convenience sample of approximately 1,000 Americans obtained online via Amazon's Mechanical Turk (MTurk) service. The people on Mechanical Turk aren't representative of all Americans, but they are nevertheless Americans.[2] We also ran the experiment on a sample of policy professionals. These people are also not representative of all Americans – they are DC staffers, journalists, lawyers, and other professionals with some degree of connection to policymaking. For the moment, please don't let the sample's lack of representativeness bother you. We'll return to the questions of generalizability and external validity in Chapter 4.

Figure 1.1 shows the results of both versions of the experiment. The MTurk experiment

[2] "Convenience sample" is a term of art that means the sample is made up of easy-to-interview people rather than randomly selected people from a well-defined population. MTurk is an online labor market where people get paid to perform small tasks, like tagging photos, transcribing videos – or answering academic surveys. MTurk makes it easy to interview large, diverse, and indeed, convenient samples.

Figure 1.1: Paul Op-Ed



is on the left and the policy professionals experiment is on the right. I've overlaid the group averages by sample partisanship, and treatment condition on top of the raw data. Democrats and Republicans clearly differ with respect to the flax tax. On MTurk, partisans in the control group differ on average by over a full point on the 1 to 7 scale – among the policy policy professionals, the gap is closer to 2.5 points. Despite these baseline differences, both Republicans and Democrats change their minds in response to the op-ed by similar amounts, something between half a point and a full point. This pattern holds for both the MTurk respondents and the policy professionals, even though these two groups of people are quite different from each other. We see very clear evidence of persuasion in parallel. If there were backlash along partisan lines, the Democrat and Republican slopes would have oppositely signed slopes. Instead of parallel motion, we would have contrary motion, but that's not what we find here – nor is it what we find in any of the persuasive information experiments in this book.

The following pages contain many figures that look just like Figure 1.1 with various elements swapped in and out. I will report experiments that I've conducted myself with collaborators, experiments by others that I have replicated on new samples, and experiments by others that I have reanalyzed using my preferred set of tools. While the specifics of the persuasive

information, the survey items used to measure policy opinions, and the subgroup divisions will vary, the pictures tell very similar stories: small effects in the direction of information.

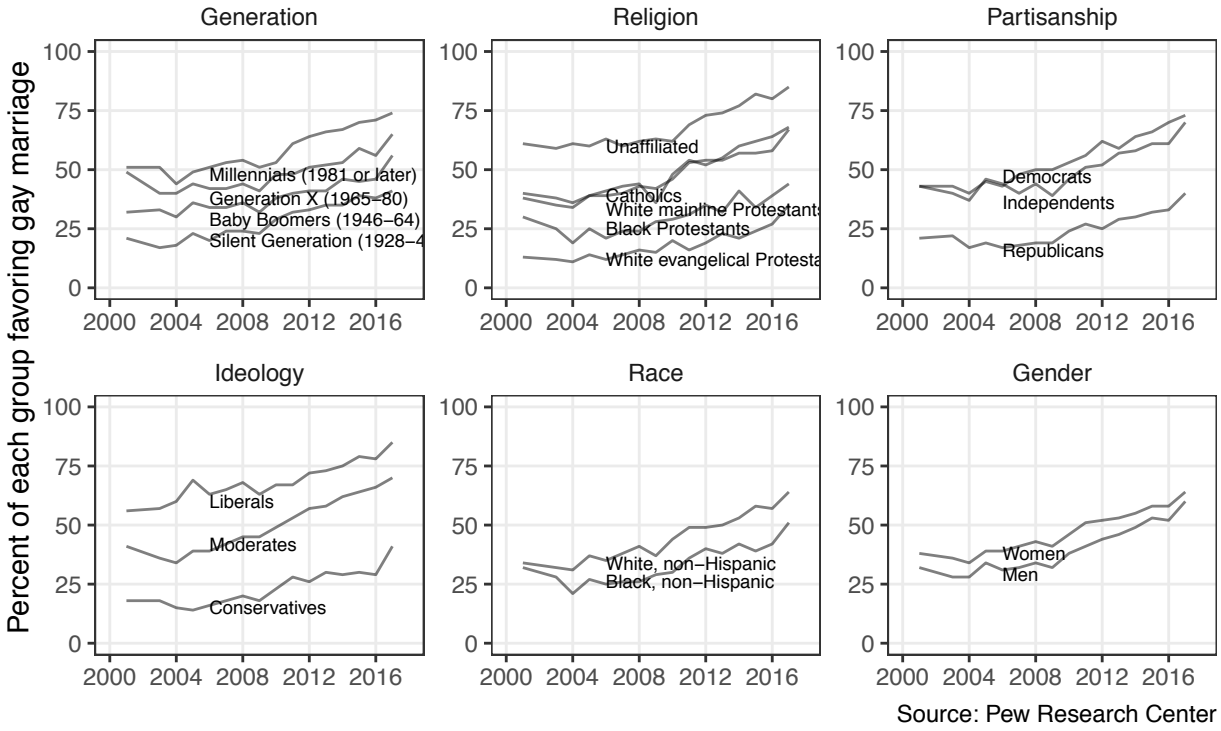
3. *Parallel publics*

In *The Rational Public*, Page and Shapiro (1992) used surveys that ask the same questions over many years to repeated cross-sections of the U.S. population to argue for the existence of "Parallel Publics." They found that for most issues, opinion changes trend in the same direction for many segments of society. According to Page and Shapiro, opinion doesn't tend to polarize in the sense that as Democrats become more supportive of an issue, Republicans become less supportive of it. On the contrary, if Democrats warm to an issue, so too do Republicans. Since *The Rational Public*, political scientists have amassed evidence in favor of the parallel publics thesis in a huge number of domains: defense spending, redistribution, presidential approval, crime, even healthcare. A brief dip into Google Scholar turns up dozens of articles that echo this find. To name a few, see Huxster, Carmichael and Brulle (2015) on climate change, Eichenberg and Stoll (2012) on defense spending, Enns (2007) on welfare spending, or Kellstedt (2003) on busing.

To see some evidence of persuasion in parallel from observational data, consider Figure 1.2, which shows how attitudes toward same-sex marriage have evolved over time. The data come from repeated cross-sectional polls of Americans conducted by Pew Research Center between 2001 and 2017. Pew estimated support for same-sex marriage in 19 separate demographic subgroups based on generation, religion, partisanship, ideology, race, and gender. In all 19, the proportion favoring gay marriage was higher in 2017 than in it was 2001, the beginning of data collection. Fitting straight lines to each series, we can estimate the average amount each group changed its position over time. Overall, the average change (or slope with respect to time) is about 1.6 percentage points every year. The slopes for some groups are slightly larger (Democrats: 2.1 points per year, White mainline Protestants: 2.1 points per year) than for others (Republicans: 1.2 points per year, Black Protestants: 1.2 points per year) but the overall pattern across subgroups is very similar from one to the next.[3]

[3] As we'll delve into in Chapter 4, comparison of magnitudes is not so straightforward. Whereas Democrats saw a greater percentage point change (30 points) than Republicans (19 points); Republicans experienced a larger *percent* change (90% increase) than Democrats (70% increase). It is not at all obvious which increase is "bigger." Suffice it to

Figure 1.2: American attitudes toward gay marriage 2001-2017

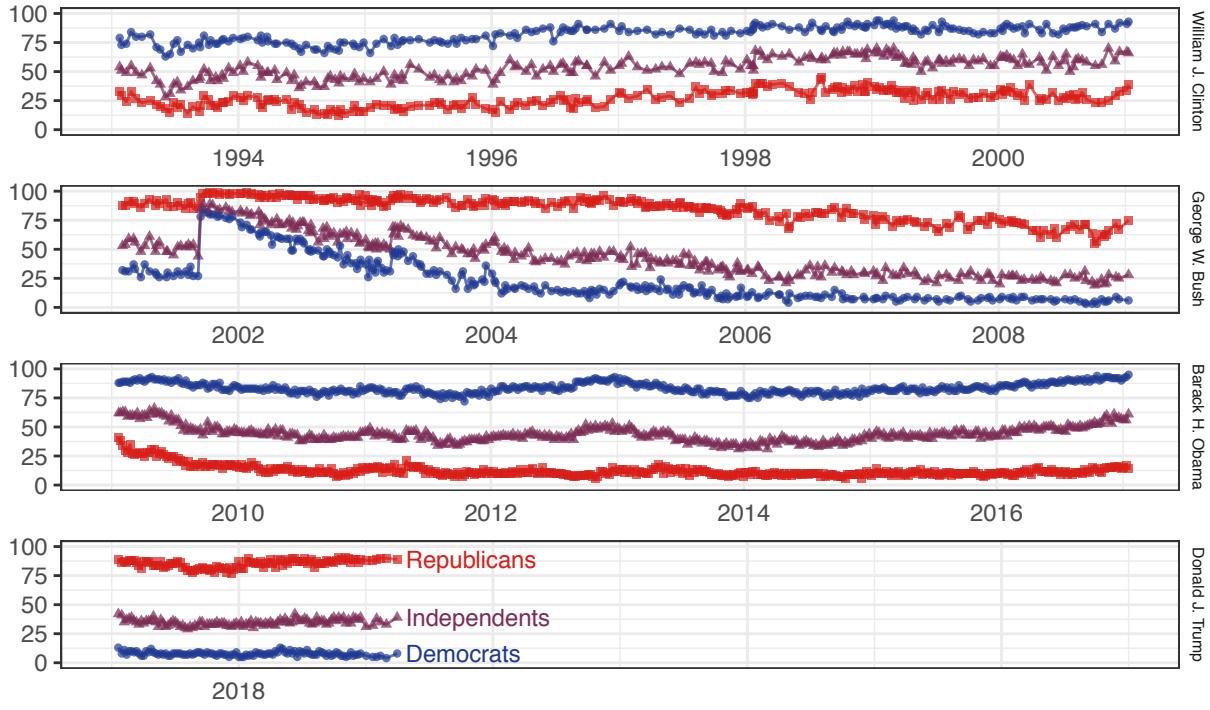


The parallel publics thesis extends to other, more obviously partisan opinions. Green, Palmquist and Schickler (2002) showed that while Democrats, Independents, and Republicans express vastly different baseline levels of presidential approval, corresponding quite plainly with the party of the current presidential office-holder, *changes* in presidential approval take place in parallel. Green, Palmquist and Schickler (2002) conducted their analysis for the presidencies of Harry Truman through Ronald Reagan using the presidential approval series from the Gallup Organization. Figure 1.3 updates their analysis from Bill Clinton through Donald Trump. These graphs are remarkable for their within-president consistency. Support for Bill Clinton rose for all three groups over the course of his presidency; support declined over the course of George W. Bush's term, and has stayed essentially stable over the course of the Obama and Trump presidencies. Some of the changes have obvious causes: the spikes in approval after 9/11 and the invasion of Iraq; the gentle rise in Obama's approval among Democrats and Independents during the 2012 campaign. Other changes, such as the long decline in Bush's approval, likely have many causes. Clearly, these partisan groupings change

say that the magnitudes of change are similar but not the same and they are difficult to rank.

their approval of the president in the same direction as one another, but not obviously by the same amount. For example, one might conclude that, following 9/11, Republicans' approval of the President did not move much, at least not as much as that of Independents or Democrats. A critic of this reading might point out that Republicans were constrained by the scale, and that a change from 88% support to near universal support (98%) is at least as impressive as Democrats' jump from 30% to 81%.

Figure 1.3: Presidential approval by partisanship

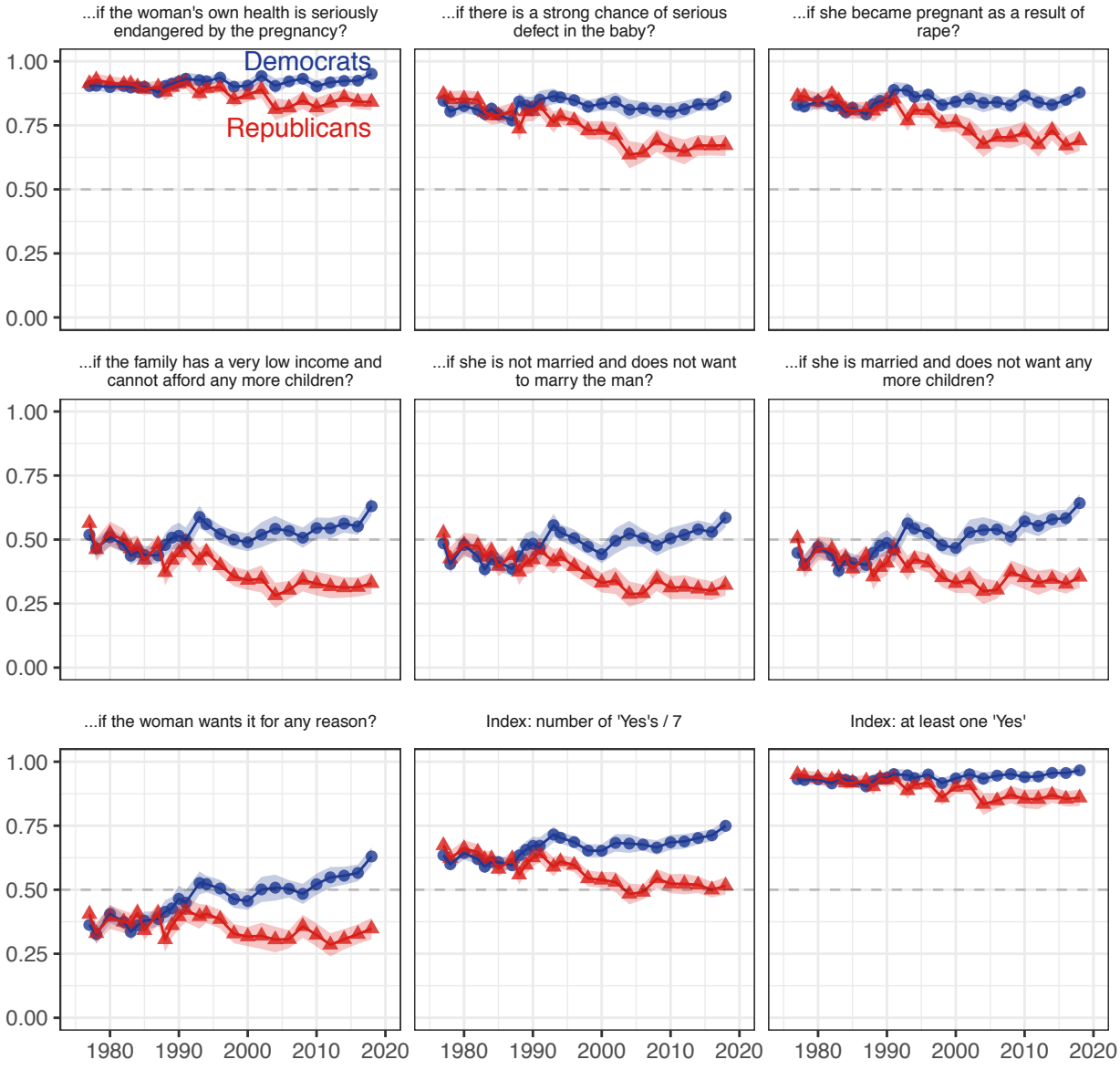


Source: The Gallup Organization

The parallel publics pattern is quite widespread but there are, however, some clear-cut exceptions. For example, Figure 1.4 shows how the last 30 years have seen a dramatic partisan divergence in abortion attitudes. Republicans and Democrats in the '70s and '80s held almost identical average opinions about the circumstances under which abortion should be allowed but have moved in opposite directions on this issue ever since. In passing, I will note that the parties appear to mostly agree on the ranking of the "reasons," and very large majorities of both parties support legal abortion in at least some cases.

The evidence from repeated cross-sectional polls like those shown in Figures 1.2, 1.3, and 1.4 can only take us so far. Observing that opinions by various subgroups of society do or do

Figure 1.4: Fraction of partisans supporting legal abortion by circumstance



Source: General Social Survey 1977 – 2018

not move together is interesting, but important methodological issues arise when we want to use these descriptive facts to draw causal inferences. These issues fall into three main categories.

First, we *want* to think of overtime change in opinion as the result of a causal process – but what treatment are these changes in response to? The variable on the horizontal axis of all these graphs is time, not some particular set of persuasive messages. Second, even if we could reasonably claim that the treatment is something like the balance of mass media messages, we can't be sure that different segments of the population are exposed to the same set of messages. Third, these repeated cross-sectional polls aren't panel studies. In a panel study, the same people are reinterviewed at multiple points in time but in a repeated cross-sectional design, the people who respond to the survey are different each year. Crucially, that means that the composition of the groups could be different each year. The kind of people who would call themselves Republican in 1980 may be at least somewhat different from the kind who call themselves Republican today. Stated differently, the partisan divergence in abortion attitudes might not be the result of Republicans and Democrats being persuaded in opposite directions, but rather the result of pro- and anti-choice people sorting themselves into the parties differently over time.

Because of these inferential difficulties, the main source of evidence in this book will come from randomized experiments. These studies have their own weaknesses and infirmities too, and I'll try to be as clear and forthcoming about those as possible throughout. But the main reason to turn to experiments to study persuasion is that we can be in control of the main causal agent we want to study. We're studying persuasive information, so that's what we'll randomize.

4. What's at stake

This single finding – parallel changes in attitudes in response to persuasive information – has major implications for our national politics.

First and foremost: *the other side is not lost*. It is worth our time to make arguments in favor of our preferred policies because we end up changing minds, even if just a little. That said, having conversations about politics is often painful. It's painful with family at Thanksgiving

dinner, it's painful on social media, it's painful among friends and coworkers. We pay a social cost when we disagree with others, but that doesn't mean the attempt has no effect at all on others' attitudes.

Second, political misinformation is dangerous, because people are persuaded by false information just like any other kind of information. We have to hold those who control media platforms of every stripe – print, broadcast, or social – accountable for the spread of lies, conspiracy theories, and propaganda. Corrections to misinformation are effective since they too are a kind of persuasive information, but we would obviously be far better off if misinformation were not spread in the first place.

Finally, we must recognize that we are ourselves persuadable as well. Being open to arguments from our opponents doesn't make us hypocrites, it just means we are like everyone else: a little bit persuadable.

5. *Where we're headed*

This book is aimed at Chapter 5, which will lay out the evidence from a large number of survey experiments that persuasion occurs in parallel. To get there, we're first going to correct the record on one of the most influential studies that claimed the opposite. The main claim of Lord, Ross and Lepper (1979) is that information causes "attitude polarization," which is equivalent to backlash as we've defined it. Chapter 2 will show that that claim is not correct, and unpacking the study's research design explains why. In Chapter 3, I will provide the definitions and scope conditions for the persuasion in parallel hypothesis and in Chapter 4, I will explain how my research design (the panel survey experiment) allows us to evaluate that hypothesis. Chapter 5 presents the evidence from those experiments. Chapter 6 is something of a stand-alone chapter that demonstrates the over-time durability of these persuasion effects. The first six chapters of this book will be light on theory, but Chapter 7 will show what the evidence from these persuasion experiments does or doesn't mean for two of the most important theories of information processing, Bayesian learning and motivated reasoning. Chapter 8 will sum up.

CHAPTER 7

Models of information processing

Up to now, I have consciously sidestepped *information processing*, or what may or may not be going on inside people's minds when they encounter new information. The main reason I have avoided the topic is that, unfortunately, the empirical evidence of persuasion in parallel is altogether silent on the question of what happens in between exposure to information and attitude change. I can tell from these experiments *that* information causes attitude change, but I can't tell *why*. I know *whether*, but I don't know *how*.

Nevertheless, this chapter will elaborate two theories of information processing: Bayesian reasoning and motivated reasoning. Bayesian reasoning posits that people evaluate information by considering the likelihood of the information in alternative states of the world. Information is interpreted as evidence in favor of whichever alternative state of the world is most likely to have generated the information. Bayesian reasoning has garnered an undeserved reputation as being "rational" or "reasonable," because it imagines that individuals coolly and calmly update their views in line with a mathematical formula. But depending on the inputs to that formula, we might entirely disagree that a person is updating their views reasonable. A perfectly Bayesian conspiracy theorist could interpret video footage of Saturn V rockets blasting off as further evidence that we faked the moon landing because that just what NASA *would* broadcast if they wanted to sell the lie. This line of reasoning may seem kooky, but with the right likelihood function, such a person is just as Bayesian as Spock.

Motivated reasoning, by contrast, posits that people reason in line with goals or motivations. Their goal is to arrive at a conclusion of a particular type. Within political science, a distinction is often drawn between accuracy motivations and directional motivations, and which set of motivations dominates for a particular person in a particular setting will color how they interpret evidence. If their goal is to come to a "correct" conclusion, then they will try to incor-

porate the information in ways that are most likely to yield the most accurate answer. If their goal is to come to a "congenial" conclusion, then they will incorporate information in ways that are mostly likely to yield that congenial answer.

1. *The trouble with mediation*

Information processing theories are about how and why information changes attitudes. In other words, information processing is about the set of mediators along the causal path from exposure to persuasive information to political attitudes and beliefs. Some theorists also consider information search, or the process by which people self-select into exposure to information, as a part of information processing. I will briefly touch on selective exposure below, but this section is mainly focused on how processing mediates the causal effect of persuasive information on attitudes and beliefs, conditional on exposure.

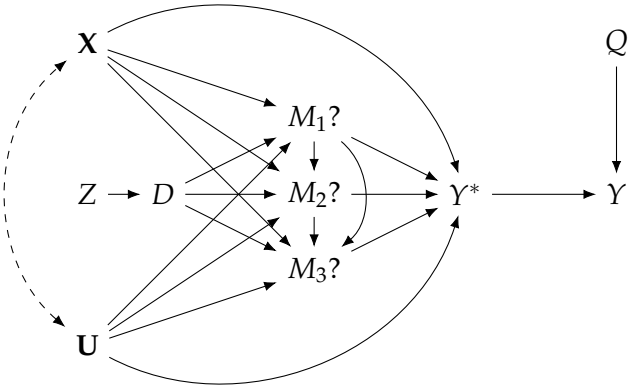
Figure 7.1 is an elaboration of the causal model described in Chapter 4. As before, on the left-hand side of the graph, we have the variable Z, which is the random assignment to treatment and D is the actual exposure to persuasive information. No variables lead into Z because in our experiments, Z will be randomly assigned, so it will be uncorrelated with all pre-treatment variables regardless of whether we measure them. The assignment perfectly manipulates exposure, thereby snipping any other paths in to D.

Exposure to information D leads into three intermediate variables M_1, M_2, and M_3. These are the mediators, or the variables that represent information processing. I include multiple mediators on the graph because multiple mediators are of course possible. I allow each mediator to affect all subsequent mediators: M_1 could affect M_2 and M_3, and M_2 could affect M_3. Later mediators can't affect earlier mediators; otherwise, the graph would contain a cycle, and so would not meet the definition of an acyclic graph. I include question marks on the graph on the mediators to represent my ignorance of what they are or how many of them there are. All mediators can affect the outcome Y.

At the top of the graph, we have X, which is the set of *observed* pre-treatment variables that might affect both the mediators and the outcome. X includes things we can measure like demographic characteristics like race, age, or gender, political variables like party identification, turnout or interest in politics, socioeconomic measures like income or education, and

psychological variables like personality or need for cognition. All of these may exert influence on the latent level of support for some policy (Y^*) and they might affect the mechanisms by which exposure to information might change that level of support. X could include everything social scientists have thought to measure. At the bottom of the graph, we have U, the set of *unobserved* variables. These are things we either don't know about yet, we haven't figured out how to measure, or failed to measure in a particular study. Just because they don't appear in our dataset doesn't mean they don't affect both the mediators and the outcome. I also include a double-headed path between U and X to indicate that I don't know anything about the causal structure of which variables in X affect which variables in U and vice-versa. The double-headed arrow indicates general confounding in unknown ways. Finally, we use a survey question Q that translates the latent outcome Y^* into the measured outcome Y.

Figure 7.1: Schematic DAG of information processing



Using data from experiments to understand the causal role of mediators is extraordinarily difficult. Within psychology, the more or less casual use of mediation models is common, owing perhaps to the wide influence of Baron and Kenny's 1986 article describing an easy-to-implement regression approach to studying mediation. Unfortunately, the Baron-Kenny approach is prone to bias (Bullock, Green and Ha, 2010; Bullock and Ha, 2011; Gerber and Green, 2012). One of the reasons Baron-Kenny fails is that it relies on very stringent modeling assumptions, in particular that the effect of Z on each M is exactly linear and exactly the same for all units. The Baron-Kenny approach to mediation analysis requires strict homogeneity of effects.

Alternative approaches to mediation relax some of the restrictive modeling assumptions of the Baron-Kenny model but nevertheless require the very strong assumption of sequential ignorability (Imai et al., 2011). Sequential ignorability is so named because it requires that a *sequence* of variables be as-if randomly assigned. First, the treatment variable (in our case, exposure to persuasive information D) must be as-if randomly assigned. This requirement is easily satisfied by actually randomly assigning exposure to treatment, as in an experiment. This first assumption is encoded in the DAG because there is no arrow from any of the variables (including U and X) to Z. Second, within each treatment condition, each mediator itself must be as-if randomly assigned (possibly after statistical adjustment for the observed variables in X). This second ignorability assumption is not justified in any way by the random assignment of the treatment. To justify making this second assumption, we have to imagine that there are not variables in U that affect any of the mediators. This isn't a matter of controlling for more variables, it's a matter of asserting that there are no more variables to control for. The quality of the inferences about the mechanisms by which treatments influence outcomes depends crucially on whether this second requirement of sequential ignorability assumption is correct. In my experience, I have rarely been convinced of a sequential ignorability assumption. If the researchers were worried enough about ignorability of the treatment variable to go to the trouble of assigning it at random, it seems inconsistent to brush away those worries when it comes to the mediators. All of this is to say that empirically studying the pathways by which information affects attitudes is difficult and I am unwilling to make the statistical assumptions that would be required to do so.

Because of the difficulty of "unpacking the black box," many analysts (myself included) elide the direct study of mechanisms and instead focus on treatment effect heterogeneity. Will the effect of exposure be higher or lower for different types of people? This is sometimes referred to as treatment effect "moderation," though that term is confusing because it connotes a causal role for the covariates that are merely correlated with the size of the treatment effect. To say that partisanship "moderates" the treatment effect of information suggests that it is *because* of the difference in partisanship that we observe differences in treatment response. A causal role for partisanship is of course possible, but partisanship could equally simply be a marker for the true unobserved causes of treatment effect heterogeneity. This is indicated in the DAG

by the double-headed arrow between X and U.

Why might a moderator be associated with treatment effect heterogeneity? Within the domain of information processing, the treatment (information) is supposed to operate via different channels for different kinds of people. That is to say, moderators predict treatment effect heterogeneity because they are associated with or are causally related to mediators. Confusingly, many of our theories about why treatment effects might be different for different kinds of people – theories about moderation – come down to predictions of how the moderators correlate with mediators. Stated differently, they come down to beliefs about how observed variables in X correlate with unobserved variables in M.

2. *Bayesian reasoning*

Bayesian reasoning is a model of cognition that is sometimes described as the model according to which "rational" people ought to process new information. Under this model, individuals are endowed with two characteristics, prior beliefs and a likelihood function. When individuals encounter new information, they update their prior beliefs to form posterior beliefs. The magnitude of the update depends on the new information and the individual's likelihood function. These updates are computed via application of Bayes' Rule. As we'll see, Bayesian reasoning's reputation as a "rational" is mostly undeserved, since people are free to have totally bonkers prior beliefs and likelihood functions.

Bayes' Rule itself is a formula that describes how to calculate a conditional probability that follows from the foundational premises of probability. [1] Bayes rule says that the conditional probability $P(A|B)$ is a function of the unconditional probability $P(A)$, and two more conditional probabilities $P(B|A)$ and $P(B|\neg A)$. Here is one way to write Bayes's Rule using this quantities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

$P(A|B)$ is called the posterior belief, $P(A)$ is called the prior belief, and together, $P(B|A)$

[1] For a clear description of how to derive Bayes Rule from the Kogomorov Axioms, see Aronow and Miller (2019; p. 11)).

and $P(B|\neg A)$ form the likelihood function. Although the terms used to describe Bayesian reasoning have a temporal flavor (prior beliefs followed by posterior beliefs), we can draw the connection between Bayesian reasoning and counterfactuals by saying that $P(A) = Y_i(0)$ and $P(A|B) = Y_i(1)$. Under the Bayesian model, the treatment effect of information[2] can therefore be written as $P(A|B) - P(A)$.

As an example of Bayesian reasoning in action, consider the treatment effect of a report that the earth is warming on belief that climate change is real. To fix ideas, let's imagine we expose a group of climate skeptics to this report.

The skeptics' prior probability that climate change is real is very low, say $P(A) = 0.01$. In order to predict the effect of information on their beliefs, we need to know the skeptics' likelihood function. That is, we need to know what they think the probability the report would conclude the world is warming if climate change is real and if climate change is not real. Because the climate skeptics think scientists can almost write whatever they want regardless of the truth, they estimate both probabilities to be very high: $P(B|A) = 0.99$ and $P(B|\neg A) = 0.98$. I'm imagining that they think $P(B|A)$ is ever so slightly higher than $P(B|\neg A)$, because after all, maintaining a global scientific conspiracy is tough. Plugging these numbers into Bayes' rule yields a posterior of $P(A|B) = 0.0101$. If these really are the prior belief and likelihood functions of the climate skeptics and if they really do follow Bayes' rule when forming posterior beliefs, then we would predict they update their views in the direction of information by one hundredth of a percentage point. Their posterior beliefs aren't equal to 1.0 – they still disagree vehemently with the report and its conclusions – but a tiny bit of incremental progress has been made.

Using language of mediators and moderators described above, we can say that in the Bayesian model, the treatment effects of information are moderated by individuals' likelihood *functions* and their priors; the effects are mediated by the *value* the likelihood takes on in response to treatment. This distinction is very subtle. A subject's likelihood function can't be affected by treatment – it's a fixed attribute. The actual number that the likelihood takes on *is* affected by treatment, precisely because the likelihood function takes the value of the treatment

[2]Depending on what the relevant counterfactual is, the treatment effect could also be written as $P(A|B) - P(A|\neg B)$

as an argument.

2.1 *Empirical assessments of Bayesian reasoning*

The Bayesian reasoning hypothesis may be traced at least as far back as de Finetti (1937), who appears to have been the first to explicitly invoke Bayes' rule as a model of human cognition.[3] The main thrust of that work is that although individuals make subjective probability judgments, they nevertheless respond to new information in the manner suggested by "objective" laws of probability:

> Observation cannot confirm or refute an opinion, which is and cannot be other than an opinion and thus neither true nor false; observation can only give us information which is capable of *influencing* our opinion. The meaning of this statement is very precise: it means that to the probability of a fact conditioned on this information – a probability very distinct from that of the same fact not conditioned on anything else – we can indeed attribute a different value. (Reproduced in Kyburg and Smokler 1964; p. 154, emphasis in original.)

The Bayesian learning hypothesis developed alongside the introduction of Bayesian statistical methods to the social sciences (Edwards, Lindman and Savage, 1963). Bayesian statisticians were arguing that Bayes' rule offered a sensible method for the integration of research findings, leading to the normative claim that scientists and humans generally *should* incorporate new evidence according to the cool calculus of conditional probability. It was then a short step from advocating this normative position to testing the positive claim that Bayes' rule provides a fair description of human cognition.

A long series of demonstrations (Edwards and Phillips, 1964; Peterson and Miller, 1965; Phillips and Edwards, 1966) showed that humans perform poorly compared to (a specific version of) the Bayesian ideal. In the prototypical experiment, subjects are shown two urns, one filled with a majority of blue balls and the other with a majority of red balls. They are told that they will be given a series of draws from a randomly selected urn and will have to estimate the probability that the draws come from the majority blue urn. Bayes' rule (with a binomial likelihood function) dictates how much subjects "should" update their priors; subjects are consistently too conservative and fail to update their beliefs far enough.

A more optimistic take comes from the field of linguistics. Frank and Goodman (2012)

[3] See, however, Ramsey (1931) for a formulation that comes close.

tackle the problem that listeners reason under uncertainty about speakers' intended messages when the signal isn't sufficient to discriminate among potential interpretations. Listeners need to infer what speakers mean on the basis of beliefs about speakers, most notably that they will prioritize signaling relevant information. Bayesian listeners would have priors over what speakers mean, likelihood functions that describe what speakers *would say* depending on what they mean, then would calculate posteriors about what speakers did mean on the basis of what they did say. In a language game experiment, Frank and Goodman (2012) randomly divide subjects into three groups: those who give their priors, those who give their likelihoods, and those who give their posteriors. In a crazy bit of luck and magic, the posteriors come extremely close to what would be obtained from multiplying the priors and likelihoods together. This doesn't prove that listeners are actually Bayesian, but it does show that they behave very similarly to how Bayesians would.

Within political science, the Bayesian models of cognition have been most frequently applied to models of partisan evaluation and choice. Zechman (1979) and Achen (1992) propose statistical models of party identification based on a Bayesian model in which individuals update their impressions of the political parties based on new evidence. Gerber and Green (1999) argue that so-called "perceptual bias" can be explained in Bayesian terms: partisans accord negative evidence for their candidates less weight not because they irrationally disregard discordant information, but because it conflicts with their priors. Partisans are nevertheless moved by "bad news," as evidenced by the parallel public opinion movements by partisans on both sides in response to changing political and economic conditions (see Chapter 1). Bartels (2002) disagrees, arguing that a lack of convergence (i.e., parallel motion) indicates partisan bias, because after enough evidence, Bayesians with different priors (but the same likelihood functions) ought to agree. Bullock (2009) shows that the prediction of convergence requires an assumption that the underlying parameter (in this case, the "true" value of the parties) is not changing.

The disagreement between Gerber, Green, and Bullock on the one hand and Bartels on the other about whether voters update their beliefs as Bayesians is premised on a specific model of Bayesianism; the dispute is about what that model predicts. The model in question is the "normal-normal" model, in which subjects have a prior belief that is normally distributed

around a "best guess," but includes some uncertainty. The evidence arrives as a point estimate, also with uncertainty. The likelihood function embedded within the normal model requires that a Bayesian take a weighted average of the prior and the evidence, where the weights are equal to the precision of the evidence. If the evidence is unequivocal (i.e., more precise), subjects will update further in the direction of evidence. Subject to the condition that the true value of the parameter is unchanging (which it might not be, as shown by Bullock (2009)), normal-normal Bayesians should converge on the true value. But what about people who are Bayesian in their own right, but simply do not employ the normal-normal model? Bayesians of every stripe are entitled to their own likelihood functions, and each likelihood function predicts a different pattern of updating. Unless one is willing to make strong assumptions about precisely what kind of Bayesian everyone is, neither evidence of parallel trends nor converging trends is sufficient to confirm or disconfirm Bayesian reasoning.

Formal theorists have proposed many models of Bayesian reasoning in which everyone learns from information, but nevertheless persists in disagreement. Acemoglu, Chernozhukov and Yildiz (2016) show that even small amounts of heterogeneity in how uncertain individuals are about $P(B|A)$ and $P(B|\neg A)$ can lead to a long-run failure to converge. Benoît and Dubra (2014) offer a theory of "rational" attitude polarization in which some (but not all) individuals have access to auxiliary information that colors how they interpret evidence: in other words, the auxiliary information allows different individuals to have different likelihood functions. Bohren (2016) describes a process of "informational herding" in which Bayesian beliefs can fail to converge on the truth because the likelihood functions misinterpret multiple signals as being independent rather than correlated. Cheng and Hsiaw (2018) allow individuals' likelihood functions to themselves vary as a function of their beliefs about source credibility, which can generate long-run disagreement. The model proposed in Fryer, Harms and Jackson (2018) accomplishes something similar by allowing individual likelihood functions to be conditional on priors. This is far from a complete accounting of the variety of ingenious ways formal theorists have invented to allow Bayesians to interpret the same evidence differently. For more, see Fudenberg and Levine (2006); Shmaya and Yariv (2016); Koçak (2019); Little (2018); Lockwood et al. (2017); Stone (2018).

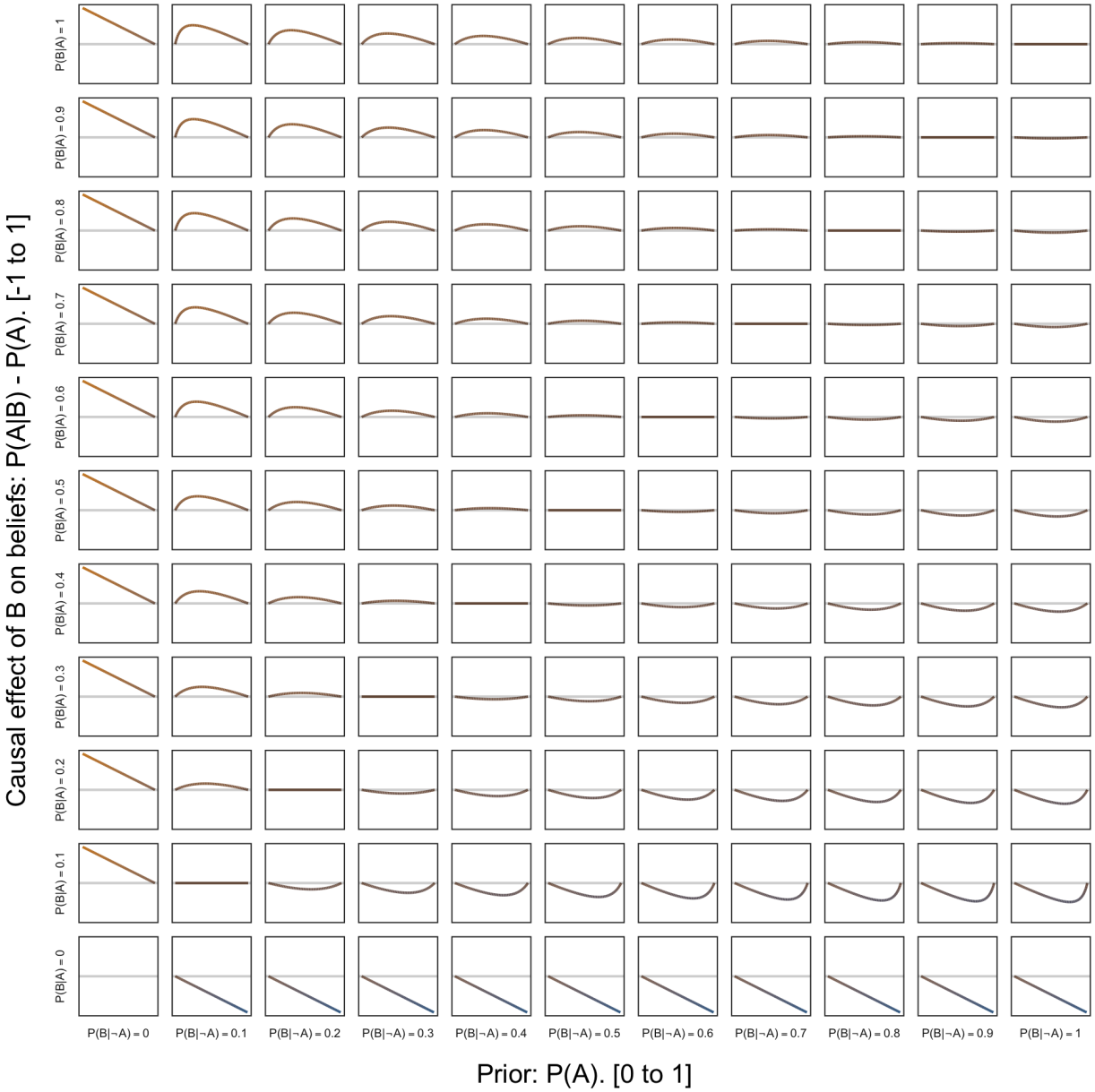
In light of the many ways Bayesians can fail to update "correctly," it makes little sense to

hold up Bayesian reasoning as a normative ideal against which frail, biased, and imperfect human information processing should be measured. The components of Bayesian reasoning – prior beliefs and likelihood functions – are unconstrained and subjective, so treatment effects of any sign and magnitude are consistent with Bayesian reasoning. If our experiments showed (contra the main claim of this book) that the treatment effect of persuasive information was positive for some people, but negative for others, we would not have evidence against Bayesian reasoning. Those "other" people might just have a likelihood function in which $P(B|A)$ is *lower* than $P(B|\neg A)$. Why people might have different likelihood functions is a difficult thing to know. The reasons could be observable (somewhere in X) or they might be unobservable (somewhere in U) and we just don't know.

Figure 7.2 shows how deeply frustrating the problem that Bayesianism doesn't mean "reasonable" truly is. Each plotted point represents how a different Bayesian agent would respond to information. The treatment effects are plotted on the vertical axis and priors on the horizontal axis. The facets group together Bayesians according to their likelihood function, with $P(B|A)$ in the rows and $P(B|\neg A)$ in the columns. Facets along the 45-degree line show no updating at all, because the likelihood functions accord equal probability to $P(B|A)$ and $P(B|\neg A)$. People with these beliefs think that the evidence is equally likely to occur regardless of whether A is true or false. Facets above the 45-degree line show positive treatments, which is to say that seeing evidence B increases the posterior probability of A. To them, the evidence is no more informative than a random number generator. Facets below the 45-degree line show negative updating. These people think that the evidence is more likely if A is not true than if it is! This is perverse if we think of B as being evidence in favor of the proposition that A is true. Every treatment effect from -100 percentage points to 100 percentage points (exclusive) is represented somewhere on the plot.

The fact that every possible pattern of updating could be accommodated within a Bayesian theory of information processing is very troublesome from a philosophy of science perspective because it means that the theory cannot be falsified by measuring the causal effect of treatments on beliefs. If the theory cannot be falsified, then we are in the position of not being able to demonstrate that it is false even if it is. Any treatment effect estimate that we obtain from a persuasion experiment can be accommodated unless further structure is put on the problem.

Figure 7.2: Bayesian updates in response to evidence can have any sign or magnitude. Each facet represents a different likelihood function.



One way to put more structure on the problem is to assume that not all likelihood functions are possible. For example, we might restrict likelihood functions to those that have the "monotone likelihood ratio property" (MLRP, Karlin and Rubin, 1956). For the binary setting we've been considering, this property reduces to a restriction that $P(B|A) \geq P(B| \neq A)$. The likelihoods that satisfy this property are those in the facets on or above the 45-degree line of Figure 7.2. This restriction is equivalent to saying that no one updates in the "wrong" direction. Falsifying the claim that everyone is a Bayesian with likelihood functions that follow the MLRP is possible – all we'd have to do is demonstrate negative treatment effects of persuasive information on attitudes. If we found such evidence, we wouldn't have evidence against Bayesianism itself, just against the particular version of Bayesianism in which all likelihood functions follow the MLRP! Conversely, If we *don't* find evidence of negative updating (and we don't in the vast majority of cases), we certainly can't affirm that the Bayesian model is true.

For me, the overwhelming evidence of persuasion in parallel doesn't mean that people are Bayesians. Instead, the relative lack of treatment effect heterogeneity means instead that if people are Bayesians, they have qualitatively similar likelihood functions.

3. *Motivated reasoning*

Motivated reasoning is a body of theory built around a model of human information processing based on goals (motivations). Humans are hypothesized to be endowed with motivations that govern information search, evaluation, and interpretation. Within social psychology, a wide variety of goals has been articulated, such as self-esteem, cognitive consistency, and belief in a just world (Psyzcynski and Greenberg, 1987). Motivated reasoning theory within political science typically partitions goals into two sets: accuracy motivations and directional motivations. By and large, the accuracy motivations are what drive people to strive for unbiased reasoning and the directional motivations divert them from that course.

The accuracy motivation propels people to hold correct beliefs or correct attitudes. It is clear enough what "correct" means for beliefs about factual matters. A belief about a fact is correct if it matches the true state of the world. For beliefs about facts, we assume there exists a truth of the matter so beliefs can be correct or incorrect. It is less clear what it means to hold a correct attitude about a political attitude object. At a minimum, it means that we presume that

some policies are better than others (at least from an individual's own perspective), so holding a correct attitude would mean evaluating better policies more positively than worse policies. This conceptualization of holding correct attitudes immediately runs into some trouble if two different individuals disagree whether one policy is better than another. They may disagree despite both being motivated by accuracy goals.

Scholars have tried to demonstrate the pull of accuracy motivations in experiments that aim to "activate" these motivations. For example, a stream of studies has shown that offering experimental subjects financial incentives for correct answers increases accuracy: Prior and Lupia (2008), Prior, Sood and Khanna (2015), Bullock et al. (2015), and Khanna and Sood (2018) all find clear evidence that paying for correct answers causes subjects to come up with correct answers more often. As with most treatment effects, we don't know *why* this treatment works, we just know that it does. It could be that subjects' latent motivation for accuracy is activated by the prospect of earning extra money, or it could be that subjects' utility is increasing in money, so they take the survey more seriously and try harder. Financial incentives aren't the only treatments that increase subject accuracy about political facts; giving them extra time also works (Prior and Lupia, 2008), as does simply asking people to provide accurate answers (Prior, Sood and Khanna, 2015). These surveys about political facts are like pop quizzes; I imagine that we could increase subjects' score on them the way we increase undergraduate students' scores: by teaching them the right answer.

The directional motivation, by contrast, is posited by motivated reasoning theorists to be the force that propels people not to reach an *accurate* conclusion, but instead to reach a *congenial* conclusion. For beliefs about facts, the directional motivation compels people to believe that they want to be true, regardless of the facts of the matter. For attitudes, directional goals motivate people to hold more positive attitudes about things they like and more negative attitudes about things they don't like. This conceptualization of having directional goals about attitudes is also a little funny – of course people hold more positive attitudes about things they like, that's what it means to like something! More charitably, however, the directional goal is presumed to motivate people to preserve and defend this attitude against alternatives *for its own sake*, and not because they would evaluate the attitude object highly in the absence of prior attitudes.

Distinguishing motivated reasoning from cognitive or pseudo-Bayesian models of reasoning has been difficult from the very start. Kunda's highly influential essay *The Case for Motivated Reasoning* opens with the plain admission that "The major and most damaging criticism of the motivational view was that all research purported to demonstrate motivated reasoning could be reinterpreted in entirely cognitive, nonmotivational terms" (Kunda, 1990; pg. 480). Kunda's rebuttal is to claim that information processing is cognitive, precisely because motivations exert their influence through cognitive processes. She writes, "People rely on cognitive processes and representations to arrive at their desired conclusions, but motivation plays a role in determining which of these will be used on a given occasion." Personally, I don't find this defense convincing since the whole difficulty is that both motivations and cognitive processing itself are unobservable. The pattern of inputs (information) and outputs (attitudes) that we do observe are, as granted by Kunda, remain consistent with either the motivational or nonmotivational account.

Taber and Lodge's 2006 article brought motivated reasoning theory into political science full-force. They enumerated three main cognitive biases through which motivated reasoning is supposed to exert its causal effect on attitudes and beliefs: *Biased assimilation* (also referred to as the prior-attitude effect) refers to individuals' predisposition to evaluate information that contradicts their priors more negatively than information that confirms their priors. The term biased assimilation is somewhat misleading since it implies that, if an argument is negatively evaluated, it will not be "assimilated" into an individual's beliefs. For example, the measure of biased assimilation in Lord, Ross and Lepper (1979) (the study discussed at length in Chapter 2) is subjects' subjective ratings of the pro- and counter attitudinal articles about capital punishment. In any case, readers should keep in mind that this term of art refers to subjective evaluations of the arguments, not how the arguments are incorporated into post-treatment attitudes. *Disconfirmation bias* refers to individuals' proclivity to counterargue counterattitudinal information more than proattitudinal information. Disconfirmation bias more generally means that counterattitudinal information is subject to greater scrutiny than proattitudinal information, presumably because people spend more time criticizing evidence they find to be low quality than they do evidence they find to be high quality. In this sense, disconfirmation bias is an extension of biased assimilation – individuals think counterattitudinal information

is low quality, so they criticize it. It would indeed be odd to counterargue information with which one *agrees*. Finally, *confirmation bias* refers to the tendency of individuals to preferentially seek out information that confirms their priors. This bias is not about information processing *per se*, but it refers to how individuals encounter information in the first place. Supposing that individuals have directional goals, seeking out attitude-confirming information is a prime way to achieve that goal.

3.1 Bayesian interpretation of biased assimilation

As admitted in (Kunda, 1990), each of these three putative biases – biased assimilation, confirmation bias, and disconfirmation bias – and could be given a non-motivational account. This section articulates how exactly the same empirical patterns could be generated by Bayesians without motivations. The purpose of this section is not to affirm the Bayesian model over motivated reasoning but rather to argue that evidence of biased assimilation, confirmation bias, or disconfirmation does not in itself provide evidence in favor of the motivated reasoning model.

Biased assimilation is the name given to the phenomenon in which people evaluate counterattitudinal evidence negatively. Let's return to the example from above about the group of climate change skeptics that we expose to a scientific climate change study. Before we were considering the effect of the study on their posterior belief that climate change is real, but now we're interested in their evaluations of the study itself. Their task is to infer whether the study is high quality ($Q = 1$) or low quality ($Q = 0$) on the basis of everything they know about the study, including its conclusion C. Studies that conclude climate change is real have $C = 1$ and studies that conclude it's all a hoax have $C = 0$. Skeptics think that studies that claim $C = 1$ are lower quality than studies that claim $C = 0$. In our setup, we can express these beliefs as $P(Q = 1|C = 1) < P(Q = 1|C = 0)$.

As above, in order to calculate posteriors, we need three numbers: the prior belief that a study is high quality $P(Q = 1)$, the probability of the study finding that climate change is real if the study is high quality $P(C = 1|Q = 1)$, and if it's low quality ($P(C = 1|Q = 0)$). Suppose that this group of skeptics thinks most published studies are low quality, so they start with a low prior: $P(Q = 1) = 0.05$. Although the skeptics are undoubtedly well and truly deceived on the subject of climate change, they share their low opinion of most scientific studies with

many academics (e.g., Ioannidis, 2005), though presumably they have different reasons. What about their likelihood functions? The distinctive feature of the climate change skeptics is that they *do not think* climate change is real, so they think it is unlikely that a high quality study would find that it is. If we grant that the skeptics *actually believe* climate change is a hoax, then it's easy to further grant that they think that high quality studies would confirm their beliefs, since that's part of what it means to hold a belief. Let's imagine they put probability of a high quality study concluding climate change is real at 1%, so $P(C = 1|Q = 1) = 0.01$. The skeptics have also probably heard that most climate change studies conclude that climate change is real. Reconciling this with their other beliefs means they have to believe that the probability of weak studies concluding climate change is real is pretty high: $P(C = 1|Q = 0) = 0.95$. Now we have everything we need to calculate the posterior probabilities of the study being high quality.

$$\begin{aligned} P(Q = 1|C = 1) &= \frac{P(C = 1|Q = 1) * P(Q = 1)}{P(C = 1|Q = 1) * P(Q = 1) + P(C = 1|Q = 0) * P(Q = 0)} \\ &= \frac{0.01 * 0.05}{0.01 * 0.05 + 0.95 * 0.95} \\ &\approx 0.0005 \end{aligned}$$

Upon seeing that the study concluded that climate change is real, our 100% Bayesian group of skeptics concluded that the probability the study was high quality was a meager 0.05% – that's one-twentieth of a percent. The skeptics doubt the study is worth the paper it's printed on. What if the study had concluded climate change was a hoax? Plugging in $1 - P(C = 1|Q = 1)$ for $P(C = 0|Q = 1)$ and $1 - P(C = 1|Q = 0)$ for $P(C = 0|Q = 0)$, we find that the skeptics are far more likely to think the study that confirms their prior to be high quality.

$$\begin{aligned} P(Q = 1|C = 0) &= \frac{P(C = 0|Q = 1) * P(Q = 1)}{P(C = 0|Q = 1) * P(Q = 1) + P(C = 0|Q = 0) * P(Q = 0)} \\ &= \frac{0.99 * 0.05}{0.99 * 0.05 + 0.05 * 0.95} \\ &\approx 0.51 \end{aligned}$$

We find that their posterior belief that the study is high quality is a little better than 50/50 at 51%. Since 0.05% is far less than 51%, these skeptics definitely engage in biased assimilation. By the same token, of course, people who accept the truth that climate change is real *also* engage in biased assimilation. Doubtless they would rate a putatively scientific study that claimed climate change is a hoax as being of far lower quality than those that reaffirm the consensus. The fact that climate skeptics think studies that agree with them are higher quality than those that don't doesn't mean their capacity to reason is broken. Instead, they could just have bad likelihood functions and priors that are very wrong.

Disconfirmation bias – the tendency of people to spend more cognitive effort criticizing arguments with which they disagree than arguments with which they agree – can be understood in the same way as the foregoing analysis of biased assimilation. Which study would you spend more time and effect criticizing: the study you think has a 0.05% chance of being high quality or the one you think has a 51% probability of being correct? Disconfirmation bias isn't so much a bias as a straightforward consequence of thinking that some arguments are stronger than others. Arguments that appear to be stronger receive less criticism – well, because they are stronger!

The third mechanism through which motivated reasoning is posited to operate is confirmation bias, or the tendency of individuals to seek out arguments that confirm their priors and to avoid counter-attitudinal arguments. In the motivated reasoning framework, individuals are motivated by directional goals. They would like to conclude that their priors are correct, so they proactively seek information in furtherance of that goal. Strictly speaking, this bias isn't about information processing, but instead is about information search.

A Bayesian interpretation of confirmation bias is also straightforward to construct. Suppose that it is costly to acquire information, so people have to be choosy about what information they gather. All else being equal, they prefer correct information to incorrect information, but they don't know which is which. They therefore gather information that, in their view, is more likely to be correct than false on the basis of signals of the information's quality: its source, sponsor, and, when easily available (as in a headline), its conclusion. Bayesians are likely to think that information that agrees with their priors is more likely to be correct, so they are likely to select it.

Whatever their implications for theories of information processing, these three behavioral patterns – biased assimilation, disconfirmation bias, and confirmation bias – have a strong empirical basis. It is indeed true that people rate counterattitudinal evidence more negatively than proattitudinal evidence. It is also true that they produce more arguments against counterattitudinal evidence than against proattitudinal evidence. And it is clear that people seek out congenial information at higher rates than information with which they disagree. However, these behaviors are not evidence that people engage in motivated reasoning, as the same patterns of behavior could plausibly be generated without positing that people are motivated to arrive at attitudinal goals.

3.2 *Discussion and critique of Taber and Lodge (2006)*

This section is an extended discussion and critique of Taber and Lodge (2006), which argued that these three mechanisms are responsible for a fourth (and much more pernicious) pattern of behavior. That article claims that jointly, biased assimilation, disconfirmation bias, and confirmation bias lead to *attitude polarization*, or the tendency of individuals to strengthen their views when encountering counterattitudinal evidence. Like Lord, Ross and Lepper (1979) before, Taber and Lodge (2006) claim to find evidence of attitude polarization. In Chapter 2, I offered a critique of the Lord, Ross and Lepper (1979) study that focused on its weak measurement strategy and the lack of random assignment. Here, I'll present a critique of Taber and Lodge (2006), which in my view suffers from different design flaws. At its heart, the reason I don't find Taber and Lodge (2006)'s evidence of attitude polarization convincing is the lack of random assignment to information, though the precise problems this creates in the study are more complicated than a standard selection story.

Taber and Lodge (2006) present two separate studies, each conducted twice on undergraduate laboratory subjects. The first study is about confirmation bias and attitude polarization; subjects were randomly assigned to participate either in the affirmative action or gun control versions of the study. The second study is about biased assimilation and disconfirmation bias; subjects who saw the affirmative action version of the first study saw the gun control version of the second study and vice-verse. For simplicity, I'll focus only on the gun control versions only, but for complexity, I'll discuss the studies in opposite order.

In the second study (about biased assimilation and disconfirmation bias), subjects were asked to rate four pro-gun control arguments and four anti-gun control arguments. Consistent with biased assimilation, gun control proponents rated the pro-gun control arguments more highly and gun control opponents did the opposite. Consistent with disconfirmation bias, subjects spent more time reading the counterattitudinal arguments and, when given the opportunity in a thought listing task, spent more effort denigrating the counterattitudinal arguments than bolstering the proattitudinal ones. So far, so good. These descriptive patterns are consistent with previous work (and indeed the experiments in this book) that people do not like evidence with which they disagree.

Let's turn now to the first study (about confirmation bias and attitude polarization). Before any exposure to information, Time 1 measures of gun control attitudes were estimated from a series of questions combined into a scale. Next, subjects participated in an information board task, in which subjects were prompted to read and rate arguments from one of four sources: the Republican party, the National Rifle Association, the Democratic Party, or Citizens Against Handguns. Importantly, subjects were not randomly assigned to participate in the information board *or not*; all subjects were exposed to the same options. Consistent with confirmation bias, Taber and Lodge show that people with more pro-gun attitudes at time 1 were more likely to choose to read arguments from the Republicans and the NRA and people with more anti-gun attitudes are more likely to read arguments from the Democrats and Citizens Against Handguns. Up to this point, I have no issues with the study. Taber and Lodge provide convincing evidence of biased assimilation, confirmation bias, and disconfirmation bias. I would dispute that findings are evidence of motivated reasoning, since nonmotivational models predict them as well, but at least we agree about what happened in the study.

The larger problem arises in the analysis of attitude polarization. The authors' claim is that exposure to mixed information caused pro-gun control subjects to become more pro-gun control and anti gun control subjects to become more anti. The authors assessed attitude polarization by regressing Time 2 attitude extremity on Time 1 extremity. They interpret regression slopes greater than 1 as evidence of attitude polarization. This approach suffers from two main weaknesses. First, the difference between time 1 and time 2 attitudes is not obviously the causal effect of treatment; if it were, there would never be a need for random assignment of

treatments, because we could rely exclusively on pre-post designs. The very act of measuring attitudes a second time could itself cause subjects to engage with the survey questions differently, among myriad other threats to inference. But suppose for the moment that we grant that the differences do represent the causal effects effects of treatment; the difficulty now is understanding what the treatment *is*. The authors describe the treatment as exposure to balanced information. However, *by design*, subjects were allowed to self-select into the information treatments. Indeed, in the section on confirmation bias, the authors convincingly show that the gun control proponents selected into pro gun control arguments and gun control opponents selected into anti-gun control arguments. Instead of attitudes polarizing, it's entirely possible that all subjects update in the direction of the balance of information *that they saw*, but because of the study design, there was a correlation between subjects time 1 attitudes and the treatments they selected into.

To summarize, Taber and Lodge (2006) provide strong empirical evidence for the behavior patterns they label as biased assimilation, disconfirmation bias, and confirmation bias. My main challenge to the interpretation of their results concerns the claim they make about how these three mechanisms lead to attitude polarization. In my view, the study was not well-designed to measure attitude polarization in response to balanced information because it did not randomly assign subjects to the treatment. A follow-up study reported in Redlawsk, Civettini and Emmerson (2010) does just that. The information board that each subject sees contains a randomly-assigned dosage of counter-attitudinal information. When the data are analyzed according to the random assignment (rather than according to the information clicked on by each subject, which risks inducing post-treatment bias), we see small, statistically insignificant effects at low doses of counterattitudinal information and then somewhat larger, statistically significant effects in the direction of information at higher doses. The authors interpret this as an "affective tipping point" past which motivated reasoners finally succumb to reality, but a more straightforward interpretation – people update their views in the direction of information – fits the data just as well and is more parsimonious.

4. Summary

Stepping back from the idiosyncrasies of particular studies and even particular theories of cognition, I think it's important to consider what even the best designed randomized studies can tell us about information processing. We are able to manipulate the input (exposure to information) and we can measure the output (survey responses). From this design, we can estimate the effect of the thing we manipulate on the outcome we measure. But we have a huge amount of trouble understanding *why* the thing we manipulate changes the thing we measure.

Why is that? The main reason is that our theories of information posit intermediate variables that we can't directly manipulate. In Bayesian reasoning, this intermediate variable is the likelihood, or the relative probabilities of seeing the evidence we saw, depending on the state of the world. Experimenters can't *set* this probability directly – they have to settle for changing it indirectly, by manipulating what evidence is seen. We would love to know if changing a likelihood changed a posterior, *holding exposure to evidence* constant, since that would provide direct evidence for the Bayesian model. But we can't, because likelihood functions are imaginary constructs that we posit exist in people's minds.

A similar critique holds for theories of motivated reasoning. The main causal agents in such theories are motivations themselves, which are the putative drivers of biased information processing. To my knowledge, no study has attempted to *set* a subject's directional motivation. Occasionally, studies of motivated reasoning "activate" a motivation with some treatment, which raises the question of whether the treatment that does the activation could have operated via some other channel than motivation. The difficulty in setting motivations also makes sense if we presuppose that motivations are part of deep-seated psychological processes that are not particularly well suited to experimental manipulation.

Both Bayesian reasoning and motivated reasoning, then, have in common that crucial parts of their theoretical underpinnings resist empirical verification. We might instead turn to how well these two theories explain behavior. There again we have the problem that both theories can accommodate any pattern of evidence. With the right mix of likelihood functions, any pattern of updating in response to persuasive information can be called Bayesian. With the

right mix of directional and accuracy goals, any pattern of updating is possible too.

All I know for sure is that motivated reasoning theories often predict negative treatment effects, whether that idea is called backlash or backfire or attitude polarization. I don't think the claims of negative treatment effects offered in Lord, Ross and Lepper (1979) and Taber and Lodge (2006) are correct. I do think that we have strong evidence that on many issues, many different kinds of people update their views in the direction of information. Whether that makes them Bayesians or motivated reasoners or something else entirely, I don't know.