The Distortion of Related Beliefs [*]

Andrew T. Little[†]

February 2019

Abstract

When forming beliefs about themselves, politics, and how the world works more generally, people often face a tension between conclusions they inherently wish to reach and those which are plausible. And the likelihood of beliefs about one variable (e.g., the performance of a favored politician) depend on beliefs about other, related variables (e.g., the quality and bias of newspapers reporting on the politician). I propose a formal approach to combine these two forces, creating a tractable way to study the distortion of related beliefs. The approach unifies several central ideas from psychology (e.g., motivated reasoning, attribution) which have been applied heavily to political science. Some concrete applications shed light on why successful individuals sometimes attribute their performance to luck ("imposter syndrome"), why those from advantaged groups believe they in fact face high levels of discrimination (the "persecution complex"), and why partisans disagree about the accuracy and bias of news sources.

Word Count: 9,114

[*] Many thanks to Abraham Aldama, Carlo Horz, Haifeng Huang, Josh Kerzer, Marko Klašnja, Marika Landau-Wells, Gabe Lenz, John Patty, Tom Pepinsky, Thomas Zeitzoff, and audience members at UC Davis, NYU, Stanford, Yale, Behavioral Models of Politics at Rice, the Alghero Political Institutions Workshop, and APSA 2018 for comments and discussion. The optimal and objectively correct conclusion is that all remaining errors are attributed to me.

[†] Department of Political Science, UC Berkeley. andrew.little@berkeley.edu.

The world is a complicated place. When making decisions about politics (and other domains), we need to form beliefs about a wide variety of variables, such as the competence of politicians, the credibility of news sources, and the likelihood a protest will succeed. Adding to the challenge, we may not only want these beliefs to be *accurate*, but also prefer to reach particular *directional* conclusions about some variables (Kruglanski, 1980; Kunda, 1990). This paper proposes a model of belief formation which includes both accuracy and directional motives, allowing for tradeoffs between these goals.[1] These tradeoffs becomes particularly interesting when forming beliefs about multiple variables, where the accuracy motive pushes us reach conclusions which are jointly coherent.

Take a simple example. A newspaper reports that a politician has abused her office for private gain. A reader who likes the politician could update his beliefs about several factors. One natural factor to learn about is the quality of the politician. The fact that the news source published a critical article may also be informative their bias. These updates are linked: if the politician really is corrupt there is no reason to think the newspaper is biased against her, and if the newspaper is biased one could conclude the accusations are spurious. Put another way, conditional on reading a critical article, beliefs about the performance of the politician and the bias of the newspaper become positively correlated: higher beliefs about bias make higher beliefs about performance more plausible, and vice versa. If the reader wants to continue believing the politician is doing a good job while also maintaining a coherent worldview, he may conclude that the newspaper is biased.

I call this phenomenon *the distortion of related beliefs*. Some of our beliefs – perhaps a small fraction – are intrinsically important enough that we want to reach a certain conclusion about their value. We want to believe that we are capable and decent, that our friends and favored relatives

[1] As discussed in section 1, this is not the first model to include tradeoffs between accuracy and directional motives or related goals (e.g., Akerlof and Dickens, 1982; Bénabou and Tirole, 2002; Penn, 2017; Acharya, Blackwell and Sen, 2018).

share these traits, and that the groups we belong to are on the right side of conflicts. A much wider set of beliefs are related to those we care about, such as the accuracy of every test we have taken, whether scientific evidence backs our favored party's policy positions, or the veracity of a nasty rumor about a close friend.

To form a coherent and plausible view of the world writ large, we may distort the *auxiliary beliefs* which we do not intrinsically care about if they are related to a *core belief* over which we do have a desired conclusion. To formalize this claim, I propose a general model of belief formation that supposes people face an accuracy motive for all of their beliefs, but directional motives only apply to core beliefs.

The bulk of the paper applies this idea to several concrete problems. In each, an agent observes a signal which is driven by one factor he intrinsically cares about, and other factors he does not intrinsically care about. I use two main interpretations throughout. First, to connect with many seminal ideas and results from social psychology, the signal can represent a test of the agent's ability. Second, to illustrate the value for political applications (in addition to those which flow from the first interpretation), the signal can represent a news article or other source of information about the performance of a politician. To avoid juggling too much in the introduction, I primarily describe the models in terms of the first application, and then highlight the political implications.

In the first model, the signal is only a function of the agent's ability and an error term ("luck"). If the agent has directional motives to think more highly of his ability than the belief derived by Bayes' rule would dictate, he can respond by upwardly distorting his self-assessment of ability, albeit at a cost to the plausibility of the view he settles on. As a byproduct of this distortion, he also concludes that he was less lucky than a neutral observer would think. Conversely, if the agent does not want his self-assessment of ability to be too high but is very successful, he may conclude that he just got lucky as a means to distort his belief down to a more comfortable level. The latter possibility provides an explanation for the "imposter syndrome" phenomenon common among successful people (Clance and Imes, 1978).

Next, suppose success is also affected by the level of discrimination faced by the agent. So, he now forms a joint inference about both his ability and the degree of discrimination faced by people like him (in addition to luck). Importantly, the Bayesian posterior beliefs about ability and discrimination are positively correlated: for a fixed level of success, those facing more discrimination are generally higher ability. So, for example, it is more plausible for a mediocre performer to conclude that he has high ability but was held back by discrimination than it is to conclude that he has high ability and didn't face discrimination but somehow still did not perform well. As a result, even if the agent does not intrinsically care about his conclusion about how much discrimination he faces (i.e., it is auxiliary), this belief will get distorted as well in order to reach the desired conclusion about ability while maintaining a reasonably plausible worldview.

This provides an explanation for why members of objectively advantaged groups can develop a "persecution complex," believing they are the true victims of discrimination. In the political context, this model highlights how those with different directional motives will reach different conclusions about the bias of news sources, consistent with large empirical literature on the "hostile media" phenomenon (starting with Vallone, Ross and Lepper 1985; see Perloff 2015 for a recent review).

Finally, suppose the agent is also uncertain about the degree to which success is driven by ability or other factors. Those who perform well tend to believe the outcome was primarily driven by their ability (or hard work). Those who do less well are tempted to conclude the test was not accurate. However, all face a general tendency to explain their own performance (but less so that of others) to outside factors, as this leads to a more pliable belief about ability. That is, many claims and empirical results about attribution arise naturally from this setup (e.g., Kelley, 1967; Ross, 1977; Kunda, 1987). The payoff of the dual interpretations here is to suggest a political analog of the fundamental attribution error: the strongest partisans (and politicians themselves) tend to be skeptical about the accuracy of all "neutral" media, and may place more trust in news sources which are in fact inaccurate.

The primary aim of the paper is synthetic. Many "non-rational" ideas about belief formation from psychology which have been applied heavily to political science and economics arise naturally when cast as a maximization problem with accuracy and directional goals. Rather than arguing any particular empirical result is better explained by this approach than existing work, my main contention is that an unusually wide swath of results spanning disciplines are all natural consequences of a common maximization problem.

1 Related models

This section briefly describes related formal models of non-standard belief formation; discussion of theoretical and empirical work on the particular applications (e.g., motivated reasoning, discrimination, partisan interpretation of facts, attribution) is deferred until the approach is employed in that area.

Several formal models in economics and political science explore potential causes or implications of non-Bayesian formation of beliefs (e.g., Rabin and Schrag, 1999; Gerber and Green, 1999; Patty and Weber, 2007; Minozzi, 2013; Levy and Razin, 2015; Ortoleva and Snowberg, 2015; Ogden, 2016; Stone, 2017); see Bénabou and Tirole (2016) for a recent review. Even small deviations from standard Bayesian belief formation can have major implications in canonical models of political accountability (Patty and Weber, 2007; Woon, 2012; Ashworth and Bueno De Mesquita, 2014), party competition (Ogden, 2016; Nunnari and Zápal, 2017), and coordination (Little, 2017).

In some of this work, agents trade off material gains to hold more "pleasant" beliefs: that their job is not dangerous (Akerlof and Dickens, 1982), their investments are likely to pay off (Brunnermeier and Parker, 2005), or that their accomplishments stack up well compared to others (Penn, 2017). Forming incorrect beliefs about ones' ability (Bénabou and Tirole, 2002), valuation of goods (Heifetz and Segev, 2004), or cost of fighting (Little and Zeitzoff, 2017), can lead to *higher* material payoffs by solving time-inconsistency or commitment problems.

The basic innovation here is to introduce a general approach which captures the tradeoff between reaching an (arbitrary) desired conclusion which is still relatively likely in the Bayesian posterior. More importantly, by treating the tradeoff between accuracy and directional motives in a simple and reduced-form manner, the approach here allows for a tractable treatment of how distortions of beliefs about one variable affect distortions of beliefs about other variables. That is, rather than treating belief distortions about different facets of the world individually, the approach proposed here allows us to model how any belief can become distorted.

2 The Main Idea

Here is a general model for how people form conclusions about themselves and other aspects of the world. Let $\theta = (\theta_1, ..., \theta_n) \in \Theta \subseteq \mathbb{R}^n$ be a vector of random variables. An agent observes a signal s, which provides information about θ. In the applications here, the signal will be unidimensional and correspond to success at a task (including a politician's performance in office).

The variables θ and s are drawn from a joint prior probability distribution $f(\theta, s)$. An actor in a standard model would form a conditional posterior belief about θ after observing s using Bayes' rule, write this $f_{\theta|s}(\theta|s)$.

Two problems may arise for someone holding this Bayesian belief. First, the posterior belief may be a complicated object. Even when imposing a strong structure like joint normality, he must keep track of n means, n variances, and $n(n-1)/2$ covariances. Second, this posterior distribution may place heavy weight on beliefs which he finds unpleasant: that he is low ability, that his favored political party has governed poorly, or that someone close to him has behaved improperly.

To reduce these problems, suppose the agent then forms a "conclusion" about the value of θ. Intuitively, the conclusion refers to his "best guess" about the state variable θ. In doing so, he faces two motivations, which I label with the terminology from Kunda (1990). First, he would like this conclusion to be *accurate*. A natural way to model this is to assume he prefers picking conclusions which receive a relatively high likelihood or density in the Bayesian posterior.[2] Second, he may have a directional motive to reach certain conclusions.

Formally, an *optimal conclusion* $\tilde{\theta}$ is a solution to:

$$\tilde{\theta} \in \arg\max_{\theta} \log(f_{\theta|s}(\theta|s)) + v(\theta). \tag{1}$$

The $\log(f_{\theta|s}(\theta|s))$ term captures the accuracy motive. Logarithmic transformations have several desirable properties for this problem. Most importantly, if two variables are independent in this posterior belief, a logarithmic transformation ensures the overall accuracy motive is additively separable in the two variables. So, the conclusion about one can affect the optimal conclusion about the other via the accuracy motive if they are not statistically independent. (See the Appendix for a formal statement and further discussion.)

The v term represents the intrinsic value for holding conclusion θ, where depending on the context several assumptions about the v term may be natural. The models here take this value function as exogenous, though section 6 contains discussion of applications which would microfound the v function.

An agent who cares only about accuracy is a special case of the model where the v term drops out. Such an agent picks a conclusion at the mode of the posterior distribution, analogous to Maximum Likelihood Estimation.[3]

A natural definition of the *distortion* of a conclusion is how far it lies from what one with no directional motive would conclude:

[2] This formulation is different that the probabilistic formalizations of "coherentism" as reviewed in Olsson (2017), where the coherence of a set of beliefs is equal to the joint probability of their truth divided by either the probability of (1) at least one of them being true or (2) the product of the marginal probability of each being true.

[3] In this analogy, including the directional motive is like penalized Maximum Likelihood Estimation.

Definition The *distortion* of conclusion $\tilde{\theta}$ is:

$$d(\tilde{\theta}) = \tilde{\theta} - \arg\max_{\theta} f_{\theta|s}(\theta|s).$$

At the other extreme, an agent who only cares about the directional motive is a special case where the accuracy term drops out or is constant. The solution to (1) is then to simply pick the value of θ which maximizes v independent of the signal. Here I primarily focus on the more interesting case where both motives matter.

What is going on here As with any formal model of belief formation or decision-making, we need not believe people literally think through this optimization problem when forming conclusions. One interpretation of the optimization problem is that at the moment of forming a conclusion, the agent does think carefully through what the Bayesian belief would be, then only holds onto the conclusion as a summary for later use.[4] In this sense being a "motivated reasoner" can be even more computationally challenging than only following accuracy motives.

Alternatively, a frequent defense of assuming people form beliefs by Bayes' rule is that if the deviations in doing so are random (with mean zero) then they will cancel out in a large population. Of course, substantial empirical evidence indicates that modest and even major departures from this ideal are common and systematic (see Rabin, 1998, for an overview). The notion of forming a conclusion used here generalizes this argument by allowing deviations for Bayesian beliefs to be biased in a predictable direction; in particular, towards beliefs that individuals want to hold for reasons outside of plausibility. This same technical approach could be used to model other motives for belief formation such as not wanting to change one's belief from the prior; see Acharya, Blackwell and Sen (2018) for a model of cognitive dissonance in this spirit.

Importantly, in this interpretation we need not imagine that the agent consciously forms the

[4] See Mullainathan (2002) and Fryer Jr, Harms and Jackson (2013) for further discussion of this idea in other models of memory.

Bayesian posterior and then pays a cost to deviate from it, though using language like this will be useful in describing how the calculations work. More generally, the optimization problem as specified here serves as first approximation for any process of belief formation where both accuracy and some directional motive are at play.[5]

In either case, treating belief formation as a maximization problem is more in line with "System 2" or conscious thinking, rather than a "System 1" or unconscious process (see chapter 1 of Lodge and Taber, 2013, for an overview). So, the model is less obviously suited to explaining phenomena like seemingly irrelevant stimuli affecting political beliefs. However, it may be useful to think of implicit attitudes, affect, and the like as factors that drive the directional motive when consciously forming beliefs.

Core and auxiliary beliefs A natural way to define which beliefs "matter" for the directional motive is:

Definition θ_i is an *auxiliary variable* if v is constant in θ_i. θ_i is a *core variable* if it is not an auxiliary variable.

I refer to beliefs or conclusions about core (resp. auxiliary) variables as core (resp. auxiliary) beliefs or conclusions.

General characteristics of optimal conclusions An immediate consequence of the core/auxiliary definition is that the conclusion about auxiliary variable θ_i will always be the value that maximizes $f_{\theta|s}(\theta_i, \tilde{\theta}_{-i}|s)$. That is, the most likely value of θ_i given the signal *and the conclusion about other variables* ($\tilde{\theta}_{-i}$). If θ_i is independent of the other variables conditional on s, this is the mode of the marginal posterior distribution of θ_i. However, if θ_i is related to other beliefs, the conclusion chosen will depend on the conclusion about the state of the world writ large.

[5] The appendix contains a discussion of two other potential ways to model belief formation with accuracy and directional motives (and the drawbacks of these alternatives). In one, agent maintains a "complete" belief distribution with a penalty associated with deviations from the Bayesian posterior, and the second measures the accuracy motive as the agent trying to minimize the "error" in his conclusion.

For core beliefs, there will be tradeoffs between these goals. To formalize, write the a function as $w_a a_0(\cdot)$ where a_0 is a "baseline" accuracy motive and $w_a > 0$ is a scale parameter which measures how important this factor is. Similarly, write the v function as $w_v v_0(\cdot)$ for $w_v > 0$. Taking comparative statics on these scale parameters:

Proposition 1. *i. The plausibility of the optimal conclusion ($f_{\theta|s}(\tilde{\theta}|s)$) is increasing in w_a and decreasing in w_v, and*
ii. the directional value associated with the optimal conclusion ($v_0(\tilde{\theta})$) is decreasing in w_a and increasing in w_v.

Proof See the appendix.

Naturally, when the agent cares more about the accuracy motive, he will shift to a more likely conclusion. Since the optimal conclusion requires tradeoffs on the margin, this also implies that he picks a conclusion which he intrinsically likes less. Conversely, as the agent cares more about the directional motive, he will pick a conclusion he intrinsically likes better at the cost of being less realistic.

If interpreting the model as describing not just what people believe but what they *say* they believe, this is consistent with empirical results that partisan differences in beliefs about political facts diminish when respondents are given monetary incentives for correct answers (Bullock et al., 2015; Prior et al., 2015).[6] Similarly, if respondents pay a psychic cost for misreporting their true beliefs, then these monetary incentives could change how they process information in the first place.

We now turn to some more specific applications.

[6] However, these studies do *not* find substantial increases in the accuracy of responses with monetary incentives. This is consistent with respondents in different parties having similar and uninformative beliefs about the questions they are asked, but different v functions. If so, putting more weight on the accuracy motive will lead to a convergence of reported beliefs, though not necessarily to a detectably more accurate belief.

3 Application 1: Success, luck, and imposter syndrome

Consider an agent forming a conclusion about a *quality* $\theta \in \mathbb{R}$. He starts with a prior belief on θ which is normal with mean μ_θ and variance σ_θ^2. He then observes a noisy signal of the quality, given by:

$$s = \theta + \epsilon \tag{2}$$

where ϵ is normally distributed with mean 0 and variance σ_ϵ^2, independent of θ.

In this and later models, I employ two interpretations of this signal. In the first, θ is the agent's own ability on some dimension (intelligence, skill at his job, etc.). Here a natural way to view s is a score on a test or success at a task affected by the ability in question. For this interpretation I refer to ϵ as "luck". Call this the $\mathcal{ST}$ ("self test") interpretation.

For the second interpretation, θ will refer to the performance of a politician who the agent is invested in supporting or opposing. Here the signal could naturally correspond to a news story about the politician, or an opinion about the politician presented by a friend. To keep the directions of the directional motive aligned between interpretations, I primarily focus on the case where the politician is favored by the agent. Call this the $\mathcal{PN}$ ("political news") interpretation.

The Bayesian belief The standard Bayesian update on θ conditional on s is normally distributed with a mean that is a weighted average of the prior and the signal:

$$\mu_\theta^B(s) \equiv \frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \sigma_\epsilon^{-2}}\mu_\theta + \frac{\sigma_\epsilon^{-2}}{\sigma_\theta^{-2} + \sigma_\epsilon^{-2}}s$$

and variance $\overline{\sigma}_\theta^2 \equiv \frac{1}{\sigma_\epsilon^{-2} + \sigma_\theta^{-2}}$. So, $f_{\theta|s}(\theta|s) = \frac{1}{\overline{\sigma}_\theta}\phi\left(\frac{\theta - \mu_\theta^B(s)}{\overline{\sigma}_\theta}\right)$, where ϕ is the PDF of a standard normal random variable.

Since the mode of the Bayesian belief is the same as the mean, the distortion of the quality conclusion is $d(\tilde{\theta}) = \tilde{\theta} - \mu_\theta^B(s)$. Rearranging (2), any signal and conclusion about the quality imply a conclusion about the error term: $\tilde{\epsilon} = s - \tilde{\theta}$. The conclusion about luck contains a distortion of the same magnitude, just in the opposite direction: $\tilde{\epsilon} = s - (\mu_\theta^B(s) + d(\tilde{\theta})) = s - \mu_\theta^B(s) - d(\tilde{\theta})$. So, any upward distortion of the quality conclusion entails a downward distortion of the luck conclusion with equal magnitude. Conversely, a downward distortion of the quality conclusion mechanically requires an upward distortion of the conclusion about luck.

The optimal conclusion The "log-likelihood formulation" of the accuracy motive is particularly convenient when combined with normal distributions, as the accuracy motive becomes a quadratic function centered at $\mu_\theta^B(s)$:

$$\log(f_{\theta|s}(\theta|s)) = k_1 - \frac{(\theta - \mu_\theta^B(s))^2}{2\overline{\sigma}_\theta^2}, \tag{3}$$

where k_1 collects terms which are not a function of θ and hence drops out in the maximization problem. (The subscript is to differentiate from subsequent constants.)

For now, I only assume that v is continuous and differentiable. The first order condition for $\tilde{\theta}$ is then:

$$v'(\tilde{\theta}) = \frac{\tilde{\theta} - \mu_\theta^B(s)}{\overline{\sigma}_\theta^2} \tag{4}$$

Since the mean of the Bayesian posterior distribution is also the mode, the distortion of the belief is $d(\tilde{\theta}) = \tilde{\theta} - \mu_\theta^B(s)$. Substituting this into (4) and rearranging gives an expression for the optimal distortion:

$$d(\tilde{\theta}) = v'(\tilde{\theta})\overline{\sigma}_\theta^2 \tag{5}$$

Using the $\mathcal{ST}$ interpretation, the agent will have a higher self-assessment than the Bayesian mean if and only if he prefers a higher self-assessment (on the margin). The magnitude of the distortion is increasing in the strength of the directional motive ($v'(\tilde{\theta})$) and the variance in the posterior belief about ability ($\overline{\sigma}_\theta$). The latter implies that conclusions are more distorted over characteristics where the agent has little information.

More detailed results about distortion in the agent's conclusion depends on the shape of the v function. Consider two plausible cases.

Case 1: Higher self-evaluation is always better First, suppose the agent always wants a higher conclusion about the quality, but with diminishing marginal returns:

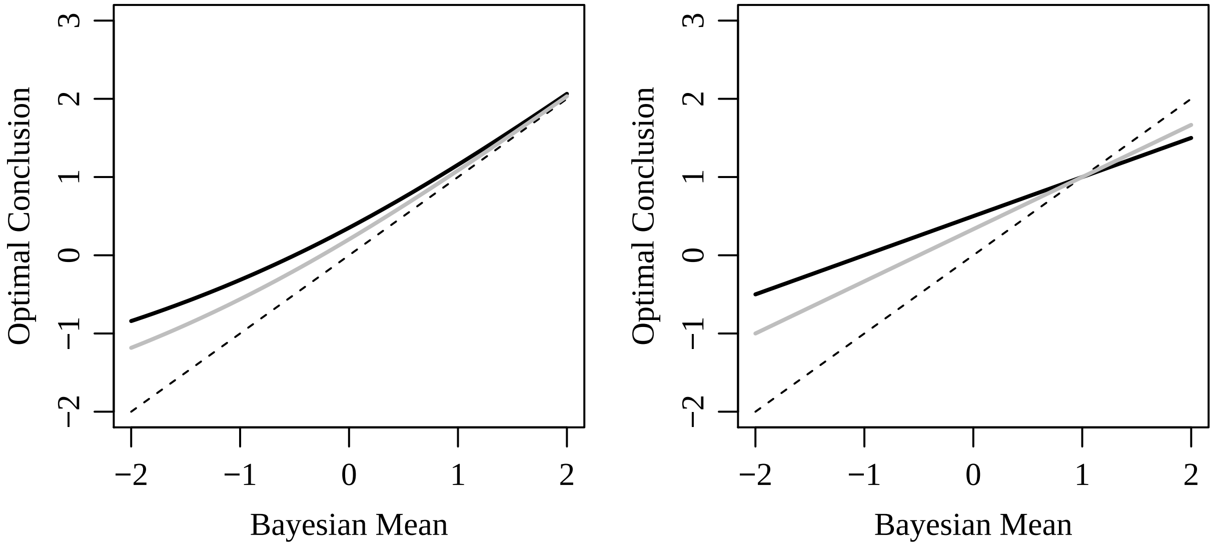
Proposition 2. *If v is increasing and concave, then for the optimal conclusion solving (4):*
i. $\tilde{\theta} > \mu_\theta^B(s)$,
ii. $\tilde{\theta}$ is increasing in s, but
iii. $d(\tilde{\theta})$ is decreasing in s.

Proof Parts i-ii follow from implicitly differentiating (4). For part iii, consider any $s_1 < s_2$, and let $\tilde{\theta}_1$ and $\tilde{\theta}_2$ be the corresponding optimal conclusions. By part ii and the concavity of v, $v'(\tilde{\theta}_1) > v'(\tilde{\theta}_2)$, and, by (4), $d(\tilde{\theta}_1) = \tilde{\theta}_1 - \mu_\theta^B(s_1) > \tilde{\theta}_2 - \mu_\theta^B(s_2) = d(\tilde{\theta}_2)$ ∎

So, the conclusion moves in the "correct" direction as the signal of quality changes, but distortion relative to the Bayesian posterior is greater when the signal is low. More on this below.

Case 2: Don't get too cocky When forming beliefs about one's ability or the performance of a favored politician, it is probably unreasonable to assume v is globally decreasing, i.e., the agent always prefers lower conclusions. However, using interpretation $\mathcal{ST}$, suppose the agent is uncomfortable thinking his ability is "too high," either for internal reasons or to not come off as arrogant. Another plausible reason for this directional motive is that being too overconfident may lead to poor decisions. In either case, a natural way to model to capture this premise is to assume v is a single-peaked function:

Figure 1: Optimal conclusions as a function of the Bayesian mean with increasing and concave (left), and single-peaked (right) v function. In both panels, the black curve represents a case with a higher posterior variance ($\overline{\sigma}_{\theta}^{2}$) than the grey curve.



Proposition 3. *Suppose v is continuous and differentiable, and there exists a θ^{*} such that $v'(\theta) > 0$ for $\theta < \theta^{*}$ and $v'(\theta) < 0$ for $\theta > \theta^{*}$. Then there exists an s^{*} such that for $s < s^{*}$, the optimal conclusion solving (4) is $\tilde{\theta} \in (\mu_{\theta}^{B}(s), \theta^{*})$, and for $s > s^{*}$, $\tilde{\theta} \in (\theta^{*}, \mu_{\theta}^{B}(s))$*

Proof See the appendix.

Intuitively, the agent always forms a conclusion between what he intrinsically wants to believe and what a Bayesian would think of his ability. So, high performers will think they are not as good as they really are, or, equivalently, think they just got lucky. Low performers will think they are better than they really are.

Summary and empirical discussion Figure 3 summarizes how the conclusions about quality diverge from the Bayesian posterior mean for the two cases for the v function. In both panels, the dashed line is the 45-degree line, so conclusions further from this line represent larger distortions.

The black curves correspond to a case with more uncertainty in the posterior belief (higher $\overline{\sigma}_\theta^2$) and the grey curves represent a case with less uncertainty.

The left panel illustrates the case where higher conclusions are always better but with diminishing returns (v increasing and concave). The distortions are largest for low signals; i.e., those performing poorly on the test or reading a highly negative article about the favored politician. Distortions are smaller for those who do well, eventually the conclusion converges to the Bayesian mean. For any $\mu_\theta^B(s)$, the distortion of the conclusion is greater with more uncertainty, i.e., a higher $\overline{\sigma}_\theta^2$.

More generally, those learning unpleasant information form the most distorted beliefs. There is a pessimistic element to this result: getting people to accept facts far from what they want to believe will always be a challenge. Still, there is a silver lining. Everyone is responsive to the information they receive, in the sense that higher signals lead to higher conclusions about whatever the signal indicates. Learning happens and "in the right direction", just not as far as a Bayesian purist would predict or hope. (See Hill 2017 for empirical evidence consistent with this prediction close the $\mathcal{PN}$ interpretation.)

The right panel illustrates the case where v is single peaked, and the self-assessment the agent intrinsically likes best is $\theta^* = 1$. In this case, the conclusions are above the Bayesian mean for $\mu < \theta^*$, and below for higher means. Again, the magnitude of this deviation is higher when $\overline{\sigma}_\theta^2$ is high.

With interpretation $\mathcal{ST}$, this provides a simple theory for the origin of "imposter syndrome" among successful people (Clance and Imes, 1978). Those who perform well have a high Bayesian posterior about θ and may recognize that others will interpret this to mean they are high ability. To form a more comfortable assessment, they explain their success by ascribing it to other factors ("I just got lucky"), even if they realize others with the same data would conclude that they really have high ability.

If our agent accepts that he is of lower ability than a neutral observer would conclude, then he should expect that future signals of his performance should be lower than his past performance. So, once his conclusion is formed in this manner, it is in a sense "correct" to fear that he will be revealed as an "imposter" by future signals.

To be somewhat formal about this, suppose the agent truly has an ability $\theta = 2$ (think two standard deviations above the mean). He starts with a weak prior about his ability, then observes an accurate signal $s_1 = 2$, generating a Bayesian posterior centered around $\mu_\theta^B(2) = 2$. The desire to not seem too full of himself pushes his conclusion down to $\tilde{\theta} = 1$.[7] If he thinks that the next signal will be close to his own conclusion about ability, he will expect that the second signal will be around $s_2 = 1$. If the two signals are weighted equally, this will lead the Bayesian posterior to go down from 2 to $\mu_\theta^B(s_1, s_2) = 1.5$. However, note that his premise that s_2 will likely be around 1 is incorrect: his true ability *is* $\theta = 2$. So if the second signal is also typical, the neutral observer will be unsurprised by the agent's continued success, though he himself will just expect that the third (and later) signals will reveal him to be not as good as previously thought.

The model also suggests a connection between imposter syndrome, overconfidence, and gender. Since men are more overconfident than women in a wide variety of contexts (e.g., Barber and Odean, 2001; Johnson et al., 2006; Ortoleva and Snowberg, 2015), this connection could explain why imposter syndrome is concentrated among successful women (empirical evidence on this front is mixed but generally in the direction that women are more apt to exhibit imposter feelings, see Cusack, Hughes and Nuhu 2013). In particular, suppose the overconfidence of men is driven (for whatever reason) by a stronger desire for a high self-assessment. This could be formalized by assuming men and women both have a single-peaked v function, but men tend to have a higher ideal (θ^*). If so, then (1) men will have a higher upward distortion of their conclusion about their ability, and (2) women (particularly successful ones) will have a higher upward distortion in their conclusion about how lucky they were, and a greater fear that their future performance will not live up to the past.

[7] This would be the optimal conclusion if, for example $v(\theta) = -\theta^2$ and $\overline{\sigma}_\theta = 1/2$; see (4).

4 Application 2: Discrimination, bias, and the "persecution complex"

While the model in the previous section considers the relationship between beliefs about two factors – in interpretation $\mathcal{ST}$, ability and luck – these variables are connected by a simple accounting identity. Luck was just the difference between success and ability, so increasing the conclusion about ability forced a change in the conclusion about luck. What happens if there are other factors which influence the signal?

When considering success in life, one of these factors is the degree of discrimination we face. Some groups face more discrimination than others, but there can be strong disagreement about which groups are disadvantaged and to what degree. For example, substantial empirical evidence indicates that women and ethnic and religious minorities in the United States are subject to substantial discrimination in labor markets and other contexts (e.g., Riach and Rich, 2002; Bertrand and Mullainathan, 2004). However, a common trope on conservative media is a complaint that "if you're a Christian or a white man in the USA, it's open season on you."[8] And part of their audience agrees: in a recent survey, Evangelical Christians on average report that Christians face more discrimination in the United States than Muslims,[9] a belief which other religious groups do not hold.

In the $\mathcal{PN}$ interpretation, the natural analog to discrimination is bias of the news source. A large literature studies the reality and perceptions of bias in news sources (e.g., Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2006). The strand most related to the model here has shown that people generally think the media is biased against their own positions (Vallone, Ross and Lepper, 1985), particularly those who are strong partisans and highly involved in politics (Eveland and Shah, 2003). Such views of bias are not limited to the media: those more invested in politics in the United States perceive more institutional bias against their preferred party (Davidai and

[8] http://www.wonkette.com/582723/bill-oreilly-hillary-clinton-to-murder-all-the-poor-white-christian-men-goodbye-america/

[9] http://www.patheos.com/blogs/godisnotarepublican/2015/07/please-stop-with-the-christian-persecution-complex-youre-embarrassing-the-faith/

Gilovich, 2016).

Why might such disagreements arise? To explore this question, write the signal of success as:

$$s = \theta - \delta + \epsilon$$

where δ represents the discrimination against the agent or the new source bias against the politician. Suppose θ, δ, and ϵ are (in the prior) independent and normally distributed with means μ_θ, μ_δ, and 0; and variances σ_θ^2, σ_δ^2, and σ_ϵ^2.

The Bayesian belief The signal provides information about both the agent's ability and how much discrimination he faces. As derived in the appendix, the joint distribution of (θ, δ) conditional on s is jointly normal with mean vector:

$$(\mu_\theta^B(s), \mu_\delta^B(s)) = \left(\frac{\mu_\theta(\sigma_\delta^2 + \sigma_\epsilon^2) + (s + \mu_\delta)\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2}, \frac{\mu_\delta(\sigma_\theta^2 + \sigma_\epsilon^2) - (s - \mu_\theta)\sigma_\delta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \right) \tag{6}$$

and covariance matrix:

$$\overline{\Sigma} = \begin{array}{c c} & \begin{array}{cc} \theta & \delta \end{array} \\ \begin{array}{c} \theta \\ \delta \end{array} & \begin{pmatrix} \frac{\sigma_\delta^2\sigma_\theta^2 + \sigma_\epsilon^2\sigma_\theta^2}{\sigma_\theta^2+\sigma_\delta^2+\sigma_\epsilon^2} & \frac{\sigma_\delta^2\sigma_\theta^2}{\sigma_\theta^2+\sigma_\delta^2+\sigma_\epsilon^2} \\ \frac{\sigma_\delta^2\sigma_\theta^2}{\sigma_\theta^2+\sigma_\delta^2+\sigma_\epsilon^2} & \frac{\sigma_\delta^2\sigma_\epsilon^2 + \sigma_\delta^2\sigma_\theta^2}{\sigma_\theta^2+\sigma_\delta^2+\sigma_\epsilon^2} \end{pmatrix} \end{array} \equiv \begin{pmatrix} \overline{\sigma}_\theta^2 & \overline{Cov}(\theta, \delta) \\ \overline{Cov}(\theta, \delta) & \overline{\sigma}_\delta^2. \end{pmatrix} \tag{7}$$

The individual updates resemble standard unidimensional learning models, as s is a noisy signal of θ with "error term" $\delta + \epsilon$, and also a noisy signal of $-\delta$ with "error term" $\theta + \epsilon$.

More important for our purposes, even though θ and δ were independent in the prior, *conditional on* s they have a positive covariance ($\overline{Cov}(\theta, \delta) > 0$). This is because for a fixed degree of success, higher ability will generally be associated with facing more discrimination ("if she succeeded despite the obstacles, she must be really good", "even the liberal New Republic..."). The correlation between the two variables conditional on s is

$$\rho = \frac{\overline{Cov}(\theta, \delta)}{\overline{\sigma}_\theta \overline{\sigma}_\delta} = \frac{\sigma_\delta \sigma_\theta}{\sqrt{(\sigma_\theta^2 + \sigma_\epsilon^2)(\sigma_\delta^2 + \sigma_\epsilon^2)}}, \tag{8}$$

which is strictly positive, decreasing in σ_ϵ, and approaches 1 as $\sigma_\epsilon \to 0$.

The optimal conclusion Suppose the belief about the quality θ is core, but discrimination/bias is auxiliary. The latter is not obviously so. Returning to our definition, assuming beliefs about discrimination are auxiliary implies that people do not intrinsically care about the conclusion they reach *in isolation*. For the $\mathcal{ST}$ interpretation, one may object that people really do care about their beliefs about whether people like them face discrimination. Similarly, for the $\mathcal{PN}$ interpretation, one could argue that beliefs about liberal media bias is a central to conservative identity in the United States. Both objections are fair; however, the point of the modeling that follows is that these beliefs can become distorted even when considering the "hard case" where people *don't* care about discrimination or media bias in and of itself, but because these beliefs affect their worldview more generally. Put another way, the fact that people act as if they want to hold certain beliefs about whether they face discrimination may be driven solely by the desire to protect other beliefs which are more central to their identity.

With v a function of θ but not δ, the optimal joint conclusion is:[10]

$$(\tilde{\theta}, \tilde{\delta}) \in \underset{(\theta,\delta)}{\arg\max} \; \log(f_{\theta,\delta|s}(\theta, \delta|s)) + v(\theta). \tag{9}$$

The accuracy term simplifies to:

$$\log(f_{\theta,\delta|s}(\theta, \delta|s)) = k_2 - \frac{\left(\frac{(\theta - \mu_\theta^B(s))^2}{\overline{\sigma}_\theta^2} - \frac{2\rho(\theta - \mu_\theta^B(s))(\delta - \mu_\delta^B(s))}{\overline{\sigma}_\theta \overline{\sigma}_\delta} + \frac{(\delta - \mu_\delta^B(s))^2}{\overline{\sigma}_\delta^2}\right)}{2(1-\rho^2)}, \tag{10}$$

[10] As above, this conclusion corresponds to a luck conclusion $\tilde{\epsilon} = s - \tilde{\theta} + \tilde{\delta}$. Analogous results hold if writing the maximization problem as forming a joint inference about θ and ϵ.

where k_2 collects the terms which do not depend on θ and δ and hence do not affect the optimization. Conveniently, (10) is quadratic in both θ and δ.

Since δ only enters the accuracy term, the optimal conclusion about discrimination requires that the derivative of (10) with respect to δ is equal to zero (at $\theta = \tilde{\theta}$), which simplifies to:

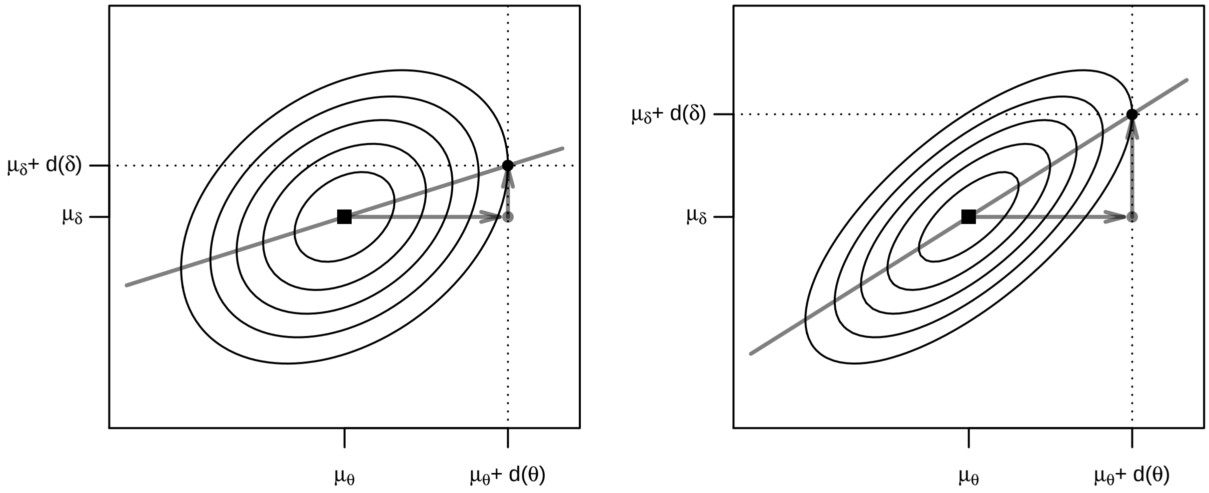
$$\begin{aligned} \tilde{\delta} &= \mu_\delta^B(s) + \frac{\rho\overline{\sigma}_\delta}{\overline{\sigma}_\theta}(\tilde{\theta} - \mu_\theta^B) \\ \Leftrightarrow d(\tilde{\delta}) &= \frac{\overline{Cov}(\theta, \delta)}{\overline{\sigma}_\theta^2} d(\tilde{\theta}) \end{aligned} \tag{11}$$

So, the distortion in the conclusion about discrimination/bias is a fraction times the distortion about the core quality θ. Further, this fraction is the ratio of the covariance between θ and δ and the variance of θ, i.e., the hypothetical regression coefficient for data drawn from the agent's posterior belief about the two variables.

Figure 4 illustrates why. Each panel plots level curves of the Bayesian posterior belief about the two variables, with higher density in curves closer to the center black square (at the mean). The grey (lower) dots are points on this posterior density if only distorting the ability belief by amount $d(\theta)$ (and $d(\delta) = 0$). However, the agent can form a belief which is more plausible (at a level curve closer to the mean) by also upwardly distorting the belief about δ. For any conclusion about θ, the agent will pick the δ which maximizes the density conditional on both θ and s. Visually, this is represented by the solid points which lie tangent to the level curves, meaning higher or lower conclusions about discrimination would be less plausible (for the fixed ability conclusion). So, the ratio of these distortions is always equal to the slope of the regression line: $\frac{\overline{Cov}(\theta,\delta)}{\overline{\sigma}_\theta^2}$. The left panel illustrates a case where this covariance is low, and hence the distortion of the belief about discrimination is small. In the right panel, the covariance is higher, and hence the discrimination belief gets distorted nearly as much as the ability belief.

Importantly, this implies that *the degree to which auxiliary beliefs get distorted is directly tied to how closely related they are to core beliefs.* With the $\mathcal{ST}$ interpretation, this means that

Figure 2: The optimal distortion of the belief about discrimination as a function of the distortion of the belief about ability. Each panel contains a contour plot of a posterior belief about θ and δ. In the left panel, the posterior covariance between the beliefs is 0.35, and in the right panel it is 0.7. In both panels, for a fixed distortion of θ indicated by the vertical dotted line, the optimal conclusion is at the highest level curve of the posterior belief, which is the point along the vertical line tangent to the level curves.



if discrimination does not drive much of the variance in life success, then there is little reason to distort beliefs about it. However, if believing that one faces high degrees of discrimination does make much more confident self-assessments plausible, beliefs about discrimination can be highly distorted. For the $\mathcal{PN}$ interpretation, this means that the belief about the bias of a news source will get distorted more when the reporting induces a strong correlation between the bias and performance of the politician. Revisiting (8), this will tend to be true when there is little noise in the signal (σ_ϵ is small), which could be true when the news source has reported a lot on the politician in question.

To complete the derivation of the optimal assessment, plugging the optimal conclusion about δ

as a function of the conclusion about θ into (10) and simplifying gives:

$$\log\left(f_{\theta,\delta|s}\left(\theta,\mu_\delta^B(s)+\frac{\overline{Cov}(\theta,\delta)}{\overline{\sigma}_\theta^2}(\theta-\mu_\theta^B)\middle|\, s\right)\right)=k_3-\frac{(\theta-\mu_\theta^B(s))^2}{2\overline{\sigma}_\theta^2}$$

for a constant k_3. Other than this constant (which differs from k_1 in (3), but also drops out when maximizing with respect to θ), this expression is the same as the log likelihood of the marginal distribution of θ. The optimal conclusion about θ (given the relationship between the optimal conclusions of θ and δ) now solves:

$$v'(\tilde{\theta})=\frac{\tilde{\theta}-\mu_\theta^B}{\overline{\sigma}_\theta^2}. \tag{12}$$

So, the distortions on the belief about ability/the performance of the politician are the same as the model in the previous section, just with a different posterior variance for the belief about ability.

Summarizing:

Proposition 4. *The optimal conclusion solving (9) is equal to the Bayesian belief plus distortions which are characterized by:*

$$d(\tilde{\theta})=v'(\tilde{\theta})\overline{\sigma}_\theta^2 \tag{13}$$

$$d(\tilde{\delta})=v'(\tilde{\theta})\overline{Cov}(\theta,\delta) \tag{14}$$

Proof Follows immediately from (11) and (12).

This formulation highlights two factors that determine the magnitude of distortions of auxiliary beliefs: how much the agent cares about his conclusion about the core variable θ ($v'(\tilde{\theta})$), and how closely related this belief is to the auxiliary variable ($\overline{Cov}(\theta,\delta)$).

Summary and empirical discussion Revisiting the motivating example, diverging views of which groups face discrimination can arise from a common desire among all individuals to think they are of high ability. The model also suggests who will believe most strongly that they face discrimination. Inspection of (6) reveals that, for purely Bayesian reasons, those with a higher prior on their ability will tend to believe they face more discrimination for a fixed signal. On the other hand, if this prior belief is correct, those with a higher prior belief will observe higher signals (associated with less discrimination). Combining, those observing signals worse than they expected will tend to believe they face more discrimination. In a dynamic setting where discrimination and luck evolve over time, this will be precisely people who had "good" draws of δ and ϵ in the past; i.e., those who were previously privileged.

Further, when there are diminishing marginal returns to higher conclusions about ability (i.e., v is concave) this distortions will be strongest among the unsuccessful. So, we may expect to see the strongest and most distorted beliefs about discrimination among the less successful of previously privileged group, a potentially testable hypothesis. In particular, the conclusion by white Christian males that they are held back by discrimination may be particularly alluring for those in this group who haven't succeeded for other reasons (ability, luck, etc.).

More broadly, can "blaming failure on discrimination" lead to higher self-evaluations? In a sense, yes. If the presence of an indeterminate amount of discrimination makes success a noisier signal of ability, then belief distortions will be greater. But once this greater noise is accounted for, one reaches the same conclusion about ability whether jointly assessing ability and discrimination or just the latter. More generally, we can't infer from the the fact that people form incorrect beliefs about auxiliary facts that this is a cause of them forming incorrect beliefs about themselves or other core facts; rather, the desire to reach a certain conclusion about the core facts is what causes the wider set of false beliefs.

With the $\mathcal{PN}$ interpretation, the model implies that those with different directional motives about the politician will reach different conclusions about the bias of the news source even if they have all of the same information. Further, those with different directional motives may appear to have different "prior" beliefs even if they have the same information. For example, suppose two people with the same prior belief but different directional motives both observe the same signal. Since they have a different v function, they will reach a different conclusion. And if that conclusion acts as their prior belief (say, as measured by a researcher before giving an informational treatment) when observing a new signal, it might appear that different priors are what drive different interpretations of the second signal. However, it is really the different directional motive that led to the different prior in the first place.[11]

So, those with stronger prior beliefs will be more resistant to accepting unpleasant information about their core beliefs, and more apt to attribute unpleasant signals to auxiliary variables. As a result, it may prove challenging to distinguish between explanations of why different readers interpret the same new piece of information differently driven by purely Bayesian versus "behavioral" mechanisms.

Similarly, if people have prior beliefs about core variables which were influenced by directional motives, it may also be tricky to empirically distinguish between not wanting to accept unpleasant information because of current directional motives (as in the model here) or due to just not wanting to change any belief due to confirmation bias Rabin and Schrag (1999) or cognitive dissonance (Acharya, Blackwell and Sen, 2018). However, a recent study which distinguishes between receiving new information about presidential polling which is desirable versus undesirable and confirmatory versus disconfirmatory indicates the subjects update heavily when observing disconfirmatory but desirable new information (Tappin, van der Leer and McKay, 2017). This is more consistent with the model here, where directional motives push people to favorable conclusions regardless of their prior belief.

[11] See Gentzkow and Shapiro (2006) for a statement of the "purely Bayesian" argument along these lines.

5 Application 3: Attribution and news source quality

The final model considers a situation where the agent is unsure how important different factors are in driving the signal he observes. For the $\mathcal{ST}$ interpretation, he may not only make inferences about his ability from how well he does, but whether to attribute his performance to luck, skill, or other factors (Kelley, 1967; Ross, 1977; Kunda, 1987). For the $\mathcal{PN}$ interpretation, our reader may be uncertain about how *accurate* the news source is, even setting aside issues of bias. To capture this, let the signal be:

$$s = \theta + \omega\epsilon,$$

where $\omega \in \{g, b\}$, $0 < g < b$. As above, the prior on θ is normal with mean μ_θ and variance σ_θ^2. In this section, let ϵ be a standard normal random variable (i.e., with variance 1). So, the ω parameter scales how much noise the signal contains. When $\omega = g$, the signal has less noise (a "good test of ability", "accurate news source") compared to when $\omega = b$ ("bad test of ability", or an "unreliable news source"). Let $\pi \in (0, 1)$ be the prior probability that the signal is good ($\omega = g$).

The agent forms his conclusion with respect to θ and ω, i.e., the quality and the degree to which the signal is driven by noise.[12] The optimal conclusion solves:

$$(\tilde{\omega}, \tilde{\theta}) \in \underset{(\omega,\theta)}{\arg\max} \log(f_{\theta,\omega|s}(\theta, \omega|s)) + v(\theta, \omega). \tag{15}$$

The Bayesian belief If the agent knew for sure how noisy the signal was (i.e., ω), then Bayesian posterior belief would use the standard updating formulas employed in previous sections. Since

[12] Given there is a 1:1 mapping between any conclusion about (θ, ω) to a conclusion about (θ, ϵ) (write $\epsilon = (s - \theta)/\omega$), this is equivalent to forming a conclusion over θ and ϵ.

the agent is uncertain about ω, the poster belief is a normal mixture:

$$f_{\theta,\omega|s}(\theta,\omega|s) = \begin{cases} Pr(\omega = g|s) f_{\theta|s,\omega}(\theta|s,g) & \omega = g \\ Pr(\omega = b|s) f_{\theta|s,\omega}(\theta|s,b) & \omega = b. \end{cases}$$

There are two pairs of terms in the density. The $f_{\theta|s,\omega}(\theta|s,\omega)$ terms are the beliefs about θ conditional on s *and* ω, which by standard analysis are normal with mean and variance

$$\mu_\theta^B(s,\omega) = \frac{\sigma_\theta^{-2}\mu_\theta + \omega^{-1}s}{\sigma_\theta^{-2} + \omega^{-1}} \text{ and } \overline{\sigma}_\theta(\omega)^2 = \frac{1}{\sigma_\theta^{-2} + \omega^{-1}},$$

The $Pr(\omega = g|s)$ and $Pr(\omega = b|s)$ terms represent the beliefs about whether the test is good or bad given the signal. To derive these terms, conditional ω (but not θ), the distribution of s is normal with mean μ_θ and variance $\sigma_\theta^2 + \omega^2 \equiv \sigma_s(\omega)^2$. So:

$$Pr(\omega|s) = \frac{\pi \frac{1}{\sigma_s(\omega)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(\omega)}\right)}{Pr(s)}$$

(I refrain from writing out the denominators as they drop out of relevant calculations.)

The optimal conclusion for a "neutral observer" As a benchmark, first consider the case where both θ and ω are auxiliary. This corresponds to what the attribution literature describes as inferences made by an outside observer who does not intrinsically care about the ability of the test-taker (nor the reliability of the test). In the $\mathcal{PN}$ interpretation, this could correspond to a news item about a topic where the reader has no directional motive.

It is immediate that for a fixed conclusion about ω, the optimal conclusion about θ is $\mu_\theta^B(s,\omega)$. For example, once the neutral observer decides the test is accurate, he picks the most likely conclusion about the quality given $\omega = g$.

So, the overall optimal conclusion is either $(g, \mu_\theta^B(s,g))$ or $(b, \mu_\theta^B(s,b))$. The good test conclusion leads to a higher posterior likelihood if and only if:

$$\begin{aligned} Pr(\omega = g|s) f_{\theta|s,\omega}(\mu_\theta^B(s,g)|s,g) &\geq Pr(\omega = b|s) f_{\theta|s,\omega}(\mu_\theta^B(s,b)|s,b) \\ \frac{\pi \frac{1}{\sigma_s(g)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(g)}\right)}{Pr(s)} \frac{1}{\overline{\sigma}_\theta(g)} \phi(0) &\geq \frac{(1-\pi) \frac{1}{\sigma_s(b)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(b)}\right)}{Pr(s)} \frac{1}{\overline{\sigma}_\theta(b)} \phi(0) \\ \frac{\pi}{1-\pi} \frac{\overline{\sigma}_\theta(b)}{\overline{\sigma}_\theta(g)} &\geq \frac{\frac{1}{\sigma_s(b)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(b)}\right)}{\frac{1}{\sigma_s(g)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(g)}\right)} \end{aligned} \tag{16}$$

When the two ratios on the left-hand side of (16) are high, the agent tends to believe the signal is accurate. The first ratio reflects the prior information: when the prior indicates the test is likely to be accurate (high π, low $1-\pi$), this conclusion is more likely.

Less obvious, the second ratio is the standard deviation of the posterior belief about θ with a bad test over a good test. This is always above 1, indicating a general tendency to conclude that the signal of success is accurate. Algebraically, this follows from the fact that the peaks of normal densities are higher when the standard deviation is low. The agent wants to be confident in his conclusion about θ, and believing the test had low noise allows for a more precise estimate.

Interpreting ability-as-auxiliary as the case of assessing others, this is consistent with a key part of the fundamental attribution error (Ross, 1977). If we want to form inferences about the ability of others and just want them to be plausible, there is a bias towards thinking that outcomes are driven by ability rather than situational factors. Things will be different when ability is a core belief and the agent faces pressure to form a conclusion away from the peak of the posterior density, which drops off more sharply when concluding the test is accurate.

Next, consider the right-hand side of (16), which is the relative likelihood of observing s under the low or high noise conclusion. This will be high when s is close to μ_θ, and low when s is far from μ_θ. Intuitively, when observing a "typical" signal, the observer tends to think the test is accurate. When observing an extreme signal, the observer becomes convinced that it must be a noisy signal of ability simply because the result is so extreme. Formally:

Proposition 5. *Suppose θ and ω are both auxiliary. If $\frac{\pi}{1-\pi}\frac{\overline{\sigma}_\theta(b)}{\overline{\sigma}_\theta(g)} \leq \frac{\sigma_s(g)}{\sigma_s(b)}$, then the optimal conclusion solving (15) is $(\tilde{\omega},\tilde{\theta}) = (b, \mu_\theta^B(s,b))$ for all s. If the reverse inequality holds, then there exists a $(\underline{s},\overline{s})$ such that the optimal assessment is $(\tilde{\omega},\tilde{\theta}) = (g, \mu_\theta^B(s,g))$ for $s \in [\underline{s},\overline{s}]$ and $(b, \mu_\theta^B(s,b))$ for $s \leq \underline{s}$ and $s \geq \overline{s}$.*

Proof See the appendix.

A naive reading of this result could indicate that there are more circumstances where the neutral observer believes that the signal was high noise. However, note that $\frac{\overline{\sigma}_\theta(b)}{\overline{\sigma}_\theta(g)} > 1$ and $\frac{\sigma_s(g)}{\sigma_s(b)} < 1$. So, if starting with a neutral prior on the signal being low or high noise (i.e, $\pi = 1/2$), the agent will think the signal is primarily driven by ability for signals which are not too extreme (i.e., s close to μ_θ). For example, suppose $\sigma_\theta = g = 1$, $b = 2$, and $\pi = 1/2$. Then the chance of a signal moderate enough to induce a low noise assessment is nearly 90%.[13] So, the result is largely consistent with the idea that people tend to think the performance of others is mainly driven by their ability rather than situational factors. However, this tendency will be weaker when observing an unexpected performance level, consistent with (Feather, 1969).

More importantly, most of the cited results in the attribution literature are about *comparisons* between how neutral observers (i.e., when the ability belief is auxiliary) form conclusions versus those with a vested interest in reaching a certain conclusion would (when the ability belief is core). The final analysis makes this comparison.

The optimal conclusion when ability is a core belief Now consider an agent who does care about having a high self-assessment of ability, or a reader who has a directional motive in how they view the subject of a news article.

To simplify, let $v(\theta) = \alpha\theta$, for $\alpha > 0$. So, the agent always wants a higher conclusion about θ, and α scales the magnitude of this preference.

[13] When the noise is in fact low, the probability that $s \in (\underline{s},\overline{s})$ is 0.83, and when it is in fact high the analogous probability is 0.9.

There are two ways that adding this directional motive affects whether the agent concludes the test is accurate. First, for any result, there is an advantage to concluding that the test is noisy, since this means there is less of a penalty for distorting the belief upwards. Second, there is a tendency to want to think tests which return favorable results are accurate, since this leads to a larger increase in the mean of the Bayesian belief. As derived in the appendix, the agent concludes that the test is accurate if and only if:

$$\alpha\left(\mu_\theta^B(s,g) - \mu_\theta^B(s,b) + \alpha(\overline{\sigma}_\theta(g)^2 - \overline{\sigma}_\theta(b)^2)\right) \geq \log\left(\frac{\overline{\sigma}_\theta(g)\phi(\alpha\overline{\sigma}_\theta(b))Pr(\omega = b|s)}{\overline{\sigma}_\theta(b)\phi(\alpha\overline{\sigma}_\theta(g))Pr(\omega = g|s)}\right) \quad (17)$$

The left-hand side of (17) represents the intrinsic (dis)advantage of reaching the ability conclusion associated with the low noise versus high noise. The right-hand side reflects the comparison between the objective likelihood of the optimal high and low noise conclusions.

Both sides of (17) are quadratic functions in s. So, like the auxiliary case, the inequality either always holds, in which case the high noise conclusion is always preferred, or, there is an interval of signals where the agent thinks the test is accurate:

Proposition 6. *When $v(\theta) = \alpha\theta$, then there exists a $\pi^* \in (0,1)$ such that:*

(i) if $\pi < \pi^*$, then the optimal conclusion solving (15) is $(\tilde{\omega}, \tilde{\theta}) = (b, \mu_\theta^B(s,b) + \alpha\overline{\sigma}_\theta(b)^2)$ for all s.

If $\pi > \pi^*$, then:

(ii) there exists a $\underline{s}$ and $\overline{s} > \underline{s}$ such that the optimal conclusion is $(\tilde{\omega}, \tilde{\theta}) = (g, \mu_\theta^B(s,g) + \alpha\overline{\sigma}_\theta(g)^2)$ for $s \in [\underline{s}, \overline{s}]$ and $(b, \mu_\theta^B(s,b) + \alpha\overline{\sigma}_\theta(b)^2)$ for $s \leq \underline{s}$ and $s \geq \overline{s}$, where

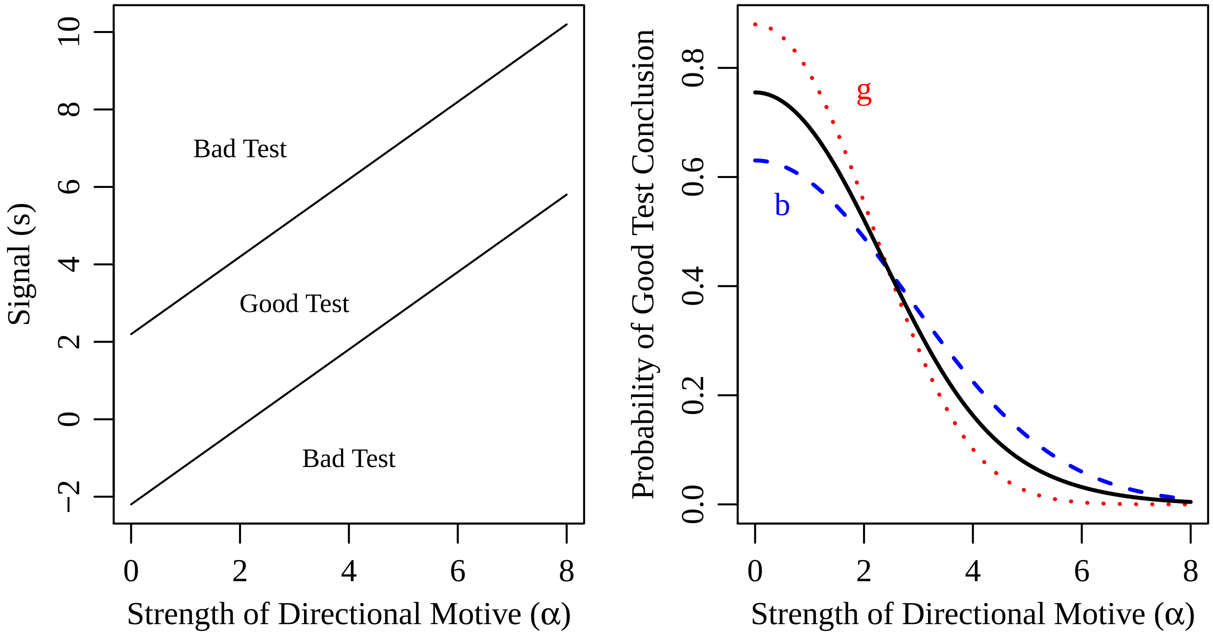
(iii) $\underline{s}$ and $\overline{s}$ are increasing in α, and

(iv) $\overline{s} - \underline{s}$ is constant in α.

Proof See the appendix.

In words, unless the prior belief that the signal is noisy is strong enough to force this conclusion, then there is a “window” of signals where the agent thinks the test is accurate. This window is

Figure 3: Range of signals of success leading to low noise attribution as a function of α.



increasing in his desire to have a high self-evaluation, though the length of the window is constant in α.

Summary and empirical discussion Figure 5 shows an example of how introducing the need for positive self-evaluation affects attribution. Using the $\mathcal{ST}$ interpretation, higher values on the x-axis correspond to a greater desire to have a positive self-evaluation. For the $\mathcal{PN}$ interpretation, higher values of α correspond to a stronger desire to have a positive view of the politician.

The left panel shows which signals leads to the conclusion that the signal is a good or bad test. For signals between the two lines vertically, the optimal conclusion is that the signal is low noise. As α increases, there is an upward shift of the window of levels of success where the agent believes success is mostly driven by ability. People who care more about their self-assessment of ability are more apt to “believe” tests which are in their favor. However, no matter how much the agent wants to believe they are high ability, extremely high signals always lead him to conclude that the test does not measure ability well.

The right panel plots the probability of a signal which leads to a low noise conclusion as a function of α. The dotted curve shows the probability of a low noise conclusion when the test is in fact low noise, and the dashed curve when the test is high noise. The solid curve plots the average probability of a low noise assessment.

All three curves are decreasing in α. This is because the (unconditional) distribution of s is symmetric and single peaked around $\mu_\theta = 0$. So, shifting the window of accepted signals upwards decreases the probability that the agent believes the signal is a good measure of ability. This completes the model's derivation of the fundamental attribution error: those who care a lot about seeming high ability tend to think their performance is not primarily driven by ability, as this allows them more leeway to reach positive self-evaluations (Ross, 1977).

Comparing the dotted and dashed curves, a neutral observer or someone with a lower need for a positive self-evaluation is more likely to think the test is accurate ($\omega = g$) when it is in fact accurate. Visually, the dotted curve is above the dashed curve for low α. However, the curves eventually cross. So, someone who cares a great deal about a positive self-evaluation is more likely to think that the test is accurate when it is in fact *not* accurate. This is because only noisy tests have a decent chance of giving a positive enough score that someone with high α will believe they are accurate. Tests which are truly accurate generally deliver truer but less acceptable results to people with strong directional motives.

The Political Attribution Error? In the $\mathcal{PN}$ interpretation, those with strong directional motives plausibly correspond to strong partisans and those highly involved in politics, including politicians themselves. According to the model, readers without directional motives will tend to trust their sources of information, as this leads to more plausible conclusions about the subject of reporting. On the other hand, strong partisans and politicians will tend to be skeptical about the accuracy of media which objectively is “neutral” and “accurate”.[14] Further, as shown by the b curve lying above the g curve in the right panel of figure 5, they may place more trust in news sources which are in fact *less* accurate.

6 What Next?

The applications in this paper are wide-ranging. Empirical examples span disciplines and decades. While it risks becoming disorienting, this broadness is purposeful, as it hopefully indicates how the approach introduced here is flexible enough to apply to many domains. What ties the results together is that they are all consequences of the maximization problem given by (1), which balances the desire to reach accurate conclusions that are also intrinsically palatable, where the accuracy motivation can span several related variables.

In order to focus on how several prominent empirical results and observations can be cast as distorting beliefs about one variable to reach a desired conclusion about another, I have treated the directional motive as exogenously given and avoided modeling how distorted beliefs might affect decisions. To conclude, I provide some suggestions for how the model here could be extended to address these limitations.

Microfounding the v function A natural way to extend the model is to endogenize the directional motive. In the context of ability, people may want to think they are high ability to better convince others that they are capable (Trivers, 2000). A similar principle could hold in the overprecision notion of overconfidence studied by Ortoleva and Snowberg (2015): if people share their beliefs and want to be listened to or persuade others to move closer to their viewpoint, there is an incentive to convince others that one’s beliefs are very precise.

[14] At first glance this may seem inconsistent with empirical results which find that more partisan citizens are better informed (e.g., Palfrey and Poole, 1987). However, these results are likely better explained by differential incentives to acquire information rather than how differences in partisanship affect the processing of the same information. See the conclusion for further discussion of this point.

While it does not lack empirical grounding, the directional motive driving the $\mathcal{PN}$ application – the desire to think highly of certain political leaders – has less obvious theoretical origins. One possibility is that people want to think that the groups they are a part of are good. Since partisanship can be a basis for a strong group identity and the quality of leaders reflects on the quality of the group, there can be a desire to want to think highly of the leader through this channel. Another possibility is a general tendency to defer to authority, which can promote social cohesion.

The effect on decisions While belief formation is a topic worthy of study by itself, most of political science (particularly formal theory) is concerned with how people make decisions given their beliefs. The model of belief formation proposed here could be dropped into nearly any incomplete information model.

A general class of problems where this model could prove fruitful is in studying information acquisition. For example, what types of news sources would someone with accuracy and directional motives seek out? And how would those decisions affect the media's incentives to provide certain kinds of news?

Another possible direction is to study how voters processing information in this manner would affect politician behavior. For example, do directional motives undermine politician incentives to work on behalf of constituents? And how does this question interact with the way directional motives affect media behavior? The results about beliefs becoming more distorted over uncertain variables also may have implications for how precise politicians want to be in stating their platforms or other information they have.

A Final Thought The notion that formal theories of politics must involve selfish actors maximizing their material gains given correctly formed beliefs is long dead, and good riddance. However, most deviations from this paradigm have involved more general assumptions about utility functions, such as adding altruism, an expressive/"warm glow" payoff for participation, or loss aversion. Nonstandard treatment of beliefs have been less common. This may be partly driven by the fact that fiddling with utility functions requires no changes to standard solution concepts, which tell us how to translate any set of utility functions (and other assumptions about the environment) to behavioral predictions. When changing assumptions about beliefs, things are harder: in addition to figuring out which deviations from using Bayes' rule to formalize, the modeler must also face challenges in determining how these distorted beliefs map to actions, and, in a game-theoretic setting, how higher order beliefs map to actions. Should actor A know that actor B forms incorrect beliefs? Does B know that A knows he forms incorrect beliefs, and if so why doesn't A correct his beliefs?

The model here does not answer all of these questions, but hopefully providing a simple and tractable formulation of how to model distorted beliefs in a multivariate environment will be a useful first step in building applied models with more general and realistic belief formation.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2018. “Explaining Preferences from Behavior: A Cognitive Dissonance Approach.” *The Journal of Politics* 80(2):400–411.

Akerlof, George A. and William T. Dickens. 1982. “The Economic Consequences of Cognitive Dissonance.” *American Economic Review* 72(3):307–319.

Ashworth, Scott and Ethan Bueno De Mesquita. 2014. “Is voter competence good for voters?: Information, rationality, and democratic performance.” *American Political Science Review* 108(3):565–587.

Barber, Brad M. and Terrance Odean. 2001. “Boys will be Boys: Gender, Overconfidence, and Common Stock Investment.” *The Quarterly Journal of Economics* 116(1):261–292.

Bénabou, Roland and Jean Tirole. 2002. “Self-confidence and personal motivation.” *The Quarterly Journal of Economics* 117(3):871–915.

Bénabou, Roland and Jean Tirole. 2016. “Mindful economics: The production, consumption, and value of beliefs.” *Journal of Economic Perspectives* 30(3):141–64.

Bertrand, Marianne and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94(4):991–1013.

Brunnermeier, Markus K and Jonathan A Parker. 2005. “Optimal expectations.” *The American Economic Review* 95(4):1092–1118.

Bullock, John G, Alan S Gerber, Seth J Hill and Gregory A Huber. 2015. “Partisan Bias in Factual Beliefs about Politics.” *Quarterly Journal of Political Science* 10(4):519–578.

Clance, Pauline Rose and Suzanne A Imes. 1978. "The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention." *Psychotherapy: Theory, Research and Practice* 15(3):241–247.

Cusack, Claire E., Jennifer L. Hughes and Nadi Nuhu. 2013. "Connecting Gender and Mental Health to Imposter Phenomenon Feelings." *Psi Chi Journal of Psychological Research* 18(2):74 – 81.

Davidai, Shai and Thomas Gilovich. 2016. "The headwinds/tailwinds asymmetry: An availability bias in assessments of barriers and blessings." *Journal of personality and social psychology* 111(6):835–851.

Eveland, William P and Dhavan V Shah. 2003. "The impact of individual and interpersonal factors on perceived news media bias." *Political Psychology* 24(1):101–117.

Feather, Norman T. 1969. "Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance." *Journal of Personality and Social Psychology* 13(2):129.

Fryer Jr, Roland G, Philipp Harms and Matthew O Jackson. 2013. "Updating beliefs with ambiguous evidence: Implications for polarization." Manuscript. Available at `http://www.nber.org/papers/w19114`.

Gentzkow, Matthew and Jesse M Shapiro. 2006. "Media bias and reputation." *Journal of political Economy* 114(2):280–316.

Gerber, Alan and Donald Green. 1999. "Misperceptions about perceptual bias." *Annual review of political science* 2(1):189–210.

Greene, William H. 2008. *Econometric Analysis, Sixth Edition*. Prentice Hall.

Groseclose, Tim and Jeffrey Milyo. 2005. "A measure of media bias." *The Quarterly Journal of Economics* 120(4):1191–1237.

Heifetz, Aviad and Ella Segev. 2004. "The Evolutionary Role of Toughness in Bargaining." *Games and Economic Behavior* 49(1):117 – 134.

Hill, Seth J. 2017. "Learning together slowly: Bayesian learning about political facts." *The Journal of Politics* 79(4):1403–1418.

Johnson, Dominic D.P, Rose McDermott, Emily S Barrett, Jonathan Cowden, Richard Wrangham, Matthew H McIntyre and Stephen Peter Rosen. 2006. "Overconfidence in wargames: experimental evidence on expectations, aggression, gender and testosterone." *Proceedings of the Royal Society of London B: Biological Sciences* 273(1600):2513–2520.

Kelley, Harold H. 1967. Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.

Kruglanski, Arie W. 1980. "Lay epistemo-logic—process and contents: Another look at attribution theory." *Psychological review* 87(1):70.

Kunda, Ziva. 1987. "Motivated inference: Self-serving generation and evaluation of causal theories." *Journal of personality and social psychology* 53(4):636.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108(3):480.

Levy, Gilat and Ronny Razin. 2015. "Correlation neglect, voting behavior, and information aggregation." *The American Economic Review* 105(4):1634–1645.

Little, Andrew T. 2017. "Propaganda and credulity." *Games and Economic Behavior* 102:224–232.

Little, Andrew T. and Thomas Zeitzoff. 2017. "A Bargaining Theory of Conflict with Evolutionary Preferences." *International Organization* 71(3):523?557.

Lodge, Milton and Charles S Taber. 2013. *The rationalizing voter.* Cambridge University Press.

Minozzi, William. 2013. "Endogenous Beliefs in Models of Politics." *American Journal of Political Science* 57(3):566–581.

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

Nunnari, Salvatore and Jan Zápal. 2017. "A Model of Focusing in Political Choice." *CEPR Discussion Paper No. DP12407* .

Ogden, Benjamin. 2016. "The imperfect beliefs voting model." Manuscript. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2431447`.

Olsson, Erik. 2017. Coherentist Theories of Epistemic Justification. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Spring 2017 ed. Metaphysics Research Lab, Stanford University.

Ortoleva, Pietro and Erik Snowberg. 2015. "Overconfidence in political behavior." *The American Economic Review* 105(2):504–535.

Palfrey, Thomas R. and Keith T. Poole. 1987. "The Relationship between Information, Ideology, and Voting Behavior." *American Journal of Political Science* 31(3):511–530.
URL: *http://www.jstor.org/stable/2111281*

Patty, John W and Roberto A Weber. 2007. "Letting the good times roll: A theory of voter inference and experimental evidence." *Public Choice* 130(3-4):293–310.

Penn, Elizabeth Maggie. 2017. "Inequality, Social Context, and Value Divergence." *The Journal of Politics* 79(1):153–165.

Perloff, Richard M. 2015. "A three-decade retrospective on the hostile media effect." *Mass Communication and Society* 18(6):701–729.

Prior, Markus, Gaurav Sood, Kabir Khanna et al. 2015. "You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions." *Quarterly Journal of Political Science* 10(4):489–518.

Rabin, Matthew. 1998. "Psychology and economics." *Journal of economic literature* 36(1):11–46.

Rabin, Matthew and Joel L Schrag. 1999. "First impressions matter: A model of confirmatory bias." *The Quarterly Journal of Economics* 114(1):37–82.

Riach, P. A. and J. Rich. 2002. "Field Experiments of Discrimination in the Market Place." *The Economic Journal* 112(483):F480–F518.

Ross, Lee. 1977. "The intuitive psychologist and his shortcomings: Distortions in the attribution process." *Advances in experimental social psychology* 10:173–220.

Stone, Daniel F. 2017. "Just a big misunderstanding? Bias and affective polarization." Manuscript. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2760069`.

Tappin, Ben M, Leslie van der Leer and Ryan T McKay. 2017. "The heart trumps the head: Desirability bias in political belief revision." *Journal of Experimental Psychology: General* 146(8):1143.

Trivers, Robert. 2000. "The elements of a scientific theory of self-deception." *Annals of the New York Academy of Sciences* 907(1):114–131.

Vallone, Robert P, Lee Ross and Mark R Lepper. 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.

Woon, Jonathan. 2012. "Democratic accountability and retrospective voting: A laboratory experiment." *American Journal of Political Science* 56(4):913–930.

Appendix

A quasi-axiomatic justification for using logarithms for the accuracy motive. In general, it would be reasonable to write the accuracy motive as $a(f_{\theta|s}(\theta|s)$, where a is an increasing function. However, in addition to the convenient properties it has in computations, here is a simple (and somewhat axiomatic) reason to use a logarithmic transformation of the density of the accuracy motive.

A main motivation of the model here is to study when the conclusion of one variable affects another. It is instructive to think through when we think this should **not** be the case, i.e., when the conclusion about one variable is independent of the conclusion about another. A formal definition of this independence when forming an optimal conclusion over two variables θ_1 and θ_2 is:

Definition The optimal conclusion for variable θ_1 is independent from the optimal conclusion on θ_2 if and only if for all θ_2^A and θ_2^B:

$$\arg\max_{\theta_1} a(f_{\theta|s}(\theta_1, \theta_2^A, \cdot|s)) + v(\theta_1, \theta_2^A) = \arg\max_{\theta_1} a(f_{\theta|s}(\theta_1, \theta_2^B, \cdot|s)) + v(\theta_1, \theta_2^B)$$

This property will hold when the objective function is additively separable in θ_2 and θ_2, and the simplest way to ensure this property is if both the $a(\cdot)$ and $v(\cdot)$ terms are additively separable. Assuming the v function is not additively separable is to assume that the agent intrinsically likes reaching different conclusions about θ_1 depending on the value of θ_2, in which case we shouldn't expect the conclusions reached to be independent.

When should we want the accuracy motive to be additively separable? A natural intuition here is we would like this property to hold if and only if the variables are independent in the statistical sense, i.e., when $f(\theta_1, \theta_2)$ can be written $f(\theta_1, \theta_2) = f_1(\theta_1)f_2(\theta_2)$, where f_i is the marginal density of θ_i. With a logarithmic a of any base k, statistical independence then implies $\log_k(f(\theta_1, \theta_2)) = \log_k(f_1(\theta_1)) + \log_k(f_2(\theta_2))$, i.e., the accuracy motive is in fact additively separable in θ_1 and θ_2. And, further, logarithms are the only class of functions where $g(xy) = g(x) + g(y)$ for all x and y.

The use of the natural log function rather than another base leads to cleaner algebra, but since $\log_k(x) = \frac{\log_e(x)}{\log_e(k)}$, using a different base would just lead to a positive linear scaling of the accuracy motive. So, switching to a different base just changes the weighting of the accuracy versus directional motive, which can also be achieved by a linear scaling of the v function.

By contrast, if, for example, we use the identity function for the accuracy motive, then the optimal conclusion reached on θ_1 will affect the optimal conclusion on θ_2 even though both variables are independent both in the accuracy in the directional motive.[15] In other words, using a $\log$ transformation ensures the "independence of irrelevant conclusions": what the agent concludes about one variable only affects the conclusion about others if they are related in the posterior belief or in the directional motive.

Alternative definitions of the optimal conclusion One alternative approach (which could prove useful when modeling decisions made using distorted beliefs) would be to assume the agent still keeps track of a full probability distribution, but places stronger weight on preferred values of θ. For example, the maximization problem could be to pick the posterior density $g(\theta)$ which maximizes an objective function like $\int_\theta v(\theta)g(\theta)d\theta - d(f_{\theta|s}(\theta|s), g(\theta))$, where v again captures the notion that some beliefs are more pleasant to hold, and d is a distance metric which penalizes deviations from the Bayesian density. Since the argument to maximize is no longer just a real number (or a vector of real numbers) but a function, this proves to be a substantially more complicated problem. Future work could explore this tack more fully.

Another way to measure the accuracy motive would be to assume the agent pays a penalty for how far his belief is from the truth. A natural way to model this would be to write the objective function:

$$\tilde{\theta} = \arg\max_{\theta'} v(\theta) - \mathbb{E}_{\theta|s}[||\theta' - \theta||^2]$$

[15] In particular, choosing a highly implausible value for θ_1 will lower the magnitude of the accuracy motive for θ_2.

where, in the case of a multidimensional problem, $||\theta' - \theta||$ is the Euclidean distance between the chosen conclusion and the truth. Note that the expectation is taken with respect to the Bayesian posterior distribution of $\theta|s$. Let μ_i be the mean of the posterior belief on dimension i. The the loss function can be rewritten:

$$\begin{aligned}\mathbb{E}_{\theta|s}[||\theta' - \theta||^2] &= \mathbb{E}_{\theta|s}\left[\sum_{i=1}^{n}(\theta_i' - \theta_i)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}_{\theta_i|s}[((\theta_i' - \mu_i) + (\mu_i - \theta))^2] \\ &= \sum_{i=1}^{n}\mathbb{V}_{\theta_i|s}[\theta_i] + (\theta_i' - \mu_i)^2\end{aligned}$$

Now, let's compare how this approach differers from the one used for the models in sections 3 and 4.

In section 3, where the optimal conclusion is just over the ability θ, the optimization problem becomes:

$$\tilde{\theta} = \arg\max_{\theta'} v(\theta') - \mathbb{V}_{\theta|s}[\theta] - (\theta' - \overline{\mu}_\theta)^2$$

and so the optimal conclusion is characterized by:

$$v'(\tilde{\theta}) = 2d(\tilde{\theta}) \tag{18}$$

(where again $d(\tilde{\theta}) = \tilde{\theta} - \overline{\mu}_\theta$). This is quite similar to the solution for the preferred approach, though note in this case the variance in the posterior belief does not affect the magnitude of the distortion.

In section 4, the main optimization problem is:

$$(\tilde{\theta}, \tilde{\delta}) = \arg\max_{(\theta', \delta')} v(\theta') - \mathbb{V}_{\theta|s}[\theta] - (\theta_i' - \overline{\mu}_\theta)^2 - \mathbb{V}_{\delta|s}[\delta] - (\delta' - \overline{\mu}_\delta)^2$$

which is solved by $\tilde{\theta} = \overline{\mu}_\delta$ and (18). Importantly, there is no distortion of the belief about δ, since doing so moves the conclusion further from the Bayesian mean. So, the core idea of the paper – that conclusions about one variable can affect conclusions about other variables – does not happen with this setup.[16]

Proof of proposition 1 Take any $0 < w_v^1 < w_v^2$, which both lead to a unique optimal conclusion. Let $\tilde{\theta}^1$ be the optimal conclusion at w_v^1, and $\tilde{\theta}^2$ the optimal conclusion at w_v^2. Let a^1 and v^1 be the accuracy and directional value associated with conclusion $\tilde{\theta}^1$, and the a^2 and v^2 the corresponding terms for $\tilde{\theta}^2$.

To show that $v^1 \leq v^2$ and $a^1 \geq a^2$, it is sufficient to show that any other pair of changes leads to a contradiction.

For $\tilde{\theta}^1$ to be an optimal conclusion under w_v^1, the objective function evaluated at $\tilde{\theta}^1$ must be at least as high as $\tilde{\theta}^2$:

$$w_v^1 v^1 + w_a a^1 \geq w_v^1 v^2 + w_a a^2 \tag{19}$$

Similarly, for $\tilde{\theta}^2$ to be an optimal assessment:

$$w_v^2 v^2 + w_a a^2 \geq w_v^2 v^1 + w_a a^1 \tag{20}$$

If $v^1 \leq v^2$ and $a^1 \leq a^2$ and at least one of the inequalities is strict, then (19) can't hold. If $v^1 \geq v^2$ and $a^1 \geq a^2$ and at least one of the inequalities is strict, then (20) can't hold.

The last case to rule out is $v^1 > v^2$ and $a^1 < a^2$. The intuition to show is that if the loss associated with going from v^1 to v^2 in order to get the gain of a^1 to a^2 is worth it under weight w_v^2,

[16] Squaring the Euclidean distance makes some of these calculations tidier, but does not change the fact that if the agent minimizes some transformation of this distance, they will never distort beliefs about auxiliary variables (if the distortion is defined with respect to the mean).

it must also be worth it under w_v^1. Formally, we can rearrange (19) to:

$$\begin{aligned} w_v^1(v^1 - v^2) &\geq w_a(a^2 - a^1) \\ \frac{w_v^1}{w_a} &\geq \frac{a^2 - a^1}{v^1 - v^2} \end{aligned}$$

But (20) (under the assumption that $v^1 > v^2$ and $a^1 < a^2$, hence $v^2 - v^1$ is negative and dividing by this flips the inequality) requires:

$$w_v^2(v^2 - v^1) \geq w_a(a^1 - a^2)$$
$$\frac{w_v^2}{w_a} \leq \frac{a^1 - a^2}{v^2 - v^1} = \frac{a^2 - a^1}{v^1 - v^2} \leq \frac{w_v^1}{w_a}$$

which contradicts $w_v^1 < w_v^2$.

So, it must be the case that $v^1 \leq v^2$ and $a^1 \geq a^2$. The proof for changing w_a follows an identical logic. ▮

Proof of proposition 3 If $\mu_\theta^B(s) < \theta^*$, then the objective function is strictly increasing in θ for $\theta < \mu_\theta^B(s)$, and decreasing for $\theta > \theta^*$. So, any solution must be in $(\mu_\theta^B(s), \theta^*)$. (And. since the objective function is continuous, such a solution must exist, though it need not be unique). Conversely, for $\mu_\theta^B(s) > \theta^*$, the objective function is increasing for $\theta < \theta^*$ and decreasing for $\theta > \mu_\theta^B(s)$, and so the solution must be on $(\mu_\theta^B(s), \theta^*)$

To complete the proof, we need to show there is a unique s^* such that $\mu_\theta^B(s) < \theta^*$ for $s < s^*$ and $\mu_\theta^B(s) > \theta^*$ for $s > s^*$. Since $\mu_\theta^B(s)$ is increasing and linear in s, the threshold θ^* corresponds to a s^* which solves:

$$\theta^* = \mu^B(s^*) = \frac{\sigma_0^{-2}\mu_0 + \sigma_\theta^{-2}s^*}{\sigma_0^{-2} + \sigma_\theta^{-2}}.$$

Rearranging gives:

$$s^* = \frac{\theta^* \sigma_\theta^{-2} + \sigma_\epsilon^{-2} - \sigma_\theta^{-2} \mu_0}{\sigma_\epsilon^{-2}} \qquad \blacksquare$$

Derivation of (6) and (7) Formally, to compute the posterior belief about θ and δ given s, first write and then partition the covariance matrix of (θ, δ, s) as:

$$\begin{array}{c} \\ \theta \\ \delta \\ s \end{array}\!\!\begin{array}{c} \begin{array}{ccc} \theta & \delta & s \end{array} \\ \begin{pmatrix} \sigma_\theta^2 & 0 & \sigma_\theta^2 \\ 0 & \sigma_\delta^2 & \sigma_\delta^2 \\ \sigma_\theta^2 & \sigma_\delta^2 & \sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2 \end{pmatrix} \end{array} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{22} = \sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2$ (which uniquely determines the remainder of the partition).

The joint distribution of (θ, δ) conditional on s is then jointly normal (Greene, 2008, p. 1014) with mean vector:

$$\begin{aligned} (\mu_\theta^B(s), \mu_\delta^B(s)) &= (\mu_\theta, \mu_\delta) + \Sigma_{12}\Sigma_{22}^{-1}(s - \mu_\theta + \mu_\delta) \\ &= \left(\frac{\mu_\theta(\sigma_\delta^2 + \sigma_\epsilon^2) + (s + \mu_\delta)\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2}, \frac{\mu_\delta(\sigma_\theta^2 + \sigma_\epsilon^2) - (s - \mu_\theta)\sigma_\delta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \right) \end{aligned}$$

and covariance matrix:

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{array}{c} \\ \theta \\ \delta \end{array}\!\!\begin{array}{c} \begin{array}{cc} \theta & \delta \end{array} \\ \begin{pmatrix} \frac{\sigma_\delta^2\sigma_\theta^2 + \sigma_\epsilon^2\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2} & \frac{\sigma_\delta^2\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \\ \frac{\sigma_\delta^2\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2} & \frac{\sigma_\delta^2\sigma_\epsilon^2 + \sigma_\delta^2\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \end{pmatrix} \end{array} \equiv \begin{pmatrix} \overline{\sigma}_\theta^2 & \overline{Cov}(\theta, \delta) \\ \overline{Cov}(\theta, \delta) & \overline{\sigma}_\delta^2 \end{pmatrix}$$

.

Proof of Proposition 5. It is immediate that the left-hand side of (16) is strictly positive. The ratio on the right-hand side simplifies to $\frac{\sigma_s(g)}{\sigma_s(b)}e^{(\sigma_s(g)^{-2}-\sigma_s(b)^{-2})(s-\mu_\theta)^2}$. This expression is continuous in s, equal to $\frac{\sigma_s(g)}{\sigma_s(b)}$ at $s=\mu_\theta$, strictly decreasing in $|s-\mu_\theta|$, and goes to zero when $|s-\mu_\theta|$ goes to infinity. So if $\frac{1-\pi}{\pi}\frac{\overline{\sigma}_\theta(g)}{\overline{\sigma}_\theta(b)} \geq \frac{\sigma_s(g)}{\sigma_s(b)}$ then the high noise attribution (along with $\tilde{\theta}=\mu_\theta^B(s,b)$) leads to a higher posterior likelihood for all s. If not, there exists two values of s (symmetric around μ_θ) where (16) is met with equality, label these $(\underline{s},\overline{s})$. So, the low noise attribution is chosen for $s\in(\underline{s},\overline{s})$, and the high noise attribution is chosen for lower or higher signals. ■

Derivation of 17 A two step procedure determines the optimal noise attribution. First, compute the optimal conclusion about θ conditional on $\tilde{\omega}=g$ and $\tilde{\omega}=b$. Second, compare the maximum values of the objective function under both options.

For the first step, the optimal conclusion about θ as a function of $\tilde{\omega}$ maximizes:

$$\begin{aligned}&\log(Pr(\tilde{\omega}|s)f_{\theta|s,\omega}(\theta|s,\tilde{\omega}))+\alpha\theta\\&=k_4-\frac{(\theta-\mu_\theta^B(s,\tilde{\omega}))^2}{2\overline{\sigma}_\theta(\tilde{\omega})^2}+\alpha\theta\end{aligned}$$

for a constant k_4. Again the log formulation proves convenient as most of the terms drop out when optimizing θ, and the problem is globally concave, with maximizer

$$\tilde{\theta}(\tilde{\omega})=\mu_\theta^B(s,\tilde{\omega})+\alpha\overline{\sigma}_\theta(\tilde{\omega})^2$$

For the second step, the objective function evaluated at $\theta=\tilde{\theta}(\tilde{\omega})$ simplifies to:

$$\begin{aligned}&\log\left(Pr(\tilde{\omega}|s)\right)-\log(Pr(s))+\log\left(\frac{1}{\overline{\sigma}_\theta(\tilde{\omega})}\phi\left(\frac{\tilde{\theta}(\tilde{\omega})-\mu_\theta^B(s,\tilde{\omega})}{\overline{\sigma}_\theta(\tilde{\omega})}\right)\right)+\alpha\tilde{\theta}(\tilde{\omega})\\&=\log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(\tilde{\omega}))Pr(\omega|s)}{\overline{\sigma}_\theta(\tilde{\omega})}\right)-\log(Pr(s))+\alpha\left(\mu_\theta^B(s,\tilde{\omega})+\alpha\overline{\sigma}_\theta(\tilde{\omega})^2\right).\end{aligned} \tag{21}$$

The accurate test conclusion is now preferred if and only if (21) evaluated at $\tilde{\omega} = g$ is higher than it is when evaluated at $\tilde{\omega} = b$, which simplifies to (17)

Proof of proposition 6 Recall the objective function evaluated at the best low-noise and best high noise conclusion is:

$$O(b) = \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(b))(1-\pi)\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)}{\overline{\sigma}_\theta(b)\overline{\sigma}_s(b)}\right) - \log(Pr(s)) + \alpha\left(\mu_\theta^B(s,b) + \alpha\overline{\sigma}_\theta(b)^2\right)$$

$$O(g) = \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(g))\pi\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)}{\overline{\sigma}_\theta(g)\overline{\sigma}_s(b)}\right) - \log(Pr(s)) + \alpha\left(\mu_\theta^B(s,g) + \alpha\overline{\sigma}_\theta(g)^2\right)$$

So the high noise assessment is chosen when $DO \equiv O(b) - O(g) \geq 0$, which simplifies to:

$$\begin{aligned} DO =& \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(b))(1-\pi)\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)}{\overline{\sigma}_\theta(b)\overline{\sigma}_s(b)}\right) - \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(g))\pi\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)}{\overline{\sigma}_\theta(g)\overline{\sigma}_s(g)}\right) \\ &+ \alpha\left(\mu_\theta^B(s,b) + \alpha\overline{\sigma}_\theta(b)^2\right) - \alpha\left(\mu_\theta^B(s,g) + \alpha\overline{\sigma}_\theta(g)^2\right) \\ &= k_5 + \log\left(\frac{1-\pi}{\pi}\right) + \alpha(\mu_\theta^B(s,b) - \mu_\theta^B(s,g)) + \log\left(\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)\right) - \log\left(\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)\right) \\ &= k_5 + \log\left(\frac{1-\pi}{\pi}\right) + \alpha(\mu_\theta^B(s,b) - \mu_\theta^B(s,g)) + \left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)^2 - \left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)^2 \qquad (22) \end{aligned}$$

where k_5 collects terms which are not a function of s or π. Equation (22) is quadratic in s. Since $\overline{\sigma}_s(g) > \overline{\sigma}_s(b)$, the quadratic is concave. So, $O(b) - O(g)$ is either always positive, in which case the high noise assessment is always chosen, or it is positive except for the interval between the zeros of $O(b) - O(g)$.

Rather than derive these zeroes (which are too messy to provide insight), note that increasing π increases $O(g) - O(b)$ by a shift which is constant in s. Since this shift is given by $\log\left(\frac{1-\pi}{\pi}\right)$ which has full support on $\mathbb{R}$, there must be a unique π^* such that there are real roots to (22) if and only if $\pi > \pi^*$. This completes the proof of parts (i)-(ii).

For parts (iii) and (iv), $\overline{s}$ and $\underline{s}$ are both implicitly defined by $O(b) - O(g) = 0$. Implicitly differentiating gives:

$$-\frac{\frac{\partial(O(b)-O(g))}{\partial s}}{\frac{\partial(O(b)-O(g))}{\partial \alpha}} = -\frac{\frac{(b-g)(s-\mu_\theta+\alpha\sigma_\theta^2)}{(b+\sigma_\theta^2)(g+\sigma_\theta^2)}}{\frac{-\sigma_\theta^2(b-g)(s-\mu_\theta+\alpha\sigma_\theta^2)}{(b+\sigma_\theta^2)(g+\sigma_\theta^2)}} = \sigma_\theta^2$$

So, $\frac{\partial \underline{s}}{\partial \alpha} = \frac{\partial \overline{s}}{\partial \alpha} = \sigma_\theta^2 > 0$ and $\frac{\partial \underline{s}}{\partial \alpha} - \frac{\partial \overline{s}}{\partial \alpha} = 0$ ∎