# The Distortion of Related Beliefs*

Andrew T. Little†

August 2018

**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 framework 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.

*Many thanks to Carlo Horz, Haifeng Huang, Josh Kerzer, Marko Klašnja, Gabe Lenz, Tom Pepinsky, Thomas Zeitzoff, and audience members at UC Davis, Yale, Behavioral Models of Politics at Rice, and the Alghero Political Institutions Workshop 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.

This paper introduces a formal approach to model how people distort beliefs about themselves and the world around them. The central premise draws on a canonical distinction from social psychology between reasoning motivated by accuracy versus the aim to reach a certain conclusion (Kruglanski, 1980; Kunda, 1990). Rather than treating belief formation as following one motivation or the other, I propose a model which includes both accuracy and directional motives, allowing us to study the tradeoffs between these goals.

The framework proves particularly powerful for analyzing belief formation over multiple variables which are *related*. 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 *how good of a job is the politician doing*? Another possibility to update on is *does this newspaper exhibit bias against the politician's party*? These updates are linked: if the politician really is corrupt there is no reason to think the newspaper is biased, 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, 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 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.

To demonstrate the value of this approach, the bulk of the paper applies it 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"). A "standard" agent only concerned with accuracy would use Bayes' rule to form a posterior belief about his ability. If the agent has directional motives to think more highly of his ability than the standard belief 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 discrim-

ination 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). Their conclusions will become particularly distorted when (1) the agent has a strong directional motive about the performance of the politician, and (2) the beliefs about the quality of the politician and the bias of the newspaper are closely related, e.g., if the source has published a lot of negative coverage.

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 core ideas 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 innacurate.

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. In addition, the models in sections 3 and 4 are (to my knowledge) the first formal theories of imposter syndrome and persecution complexes.

The rest of the paper is organized as follows. Section 1 describes related work, and section 2 introduces the general modeling approach. Sections 3-5 present the applications which make up the core of the paper. Section 6 summarizes and makes suggestions for future work.

# 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 framework is employed in that area.

Several formal models in economics and political science explore potential causes or implications of non-Bayesian (or at least not-fully Bayesian) formation of beliefs (e.g., Rabin and Schrag, 1999; Minozzi, 2013; Levy and Razin, 2015; Ortoleva and Snowberg, 2015; Cheng and Hsiaw, 2017); see Bénabou and Tirole (2016) for a recent review. 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).[1]

Forming incorrect beliefs about ones' ability (Bénabou and Tirole, 2002),[2] valuation of goods

[1] These tradeoffs also resemble models of attitude formation (Acharya, Blackwell and Sen, 2018) and responses to survey questions Bullock et al. (2015).

[2] On beliefs about one's ability or knowledge, Ortoleva and Snowberg (2015) and Levy and Razin (2015) show how

(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.[3]

Another idea which plays a key role here and shows up in past work is the notion that people only store summary information about variables in memory (Mullainathan, 2002; Fryer Jr, Harms and Jackson, 2013). These papers also generally focus on the downstream implications of this kind of information processing, while here the summary statistic formed (the "conclusion") takes center stage.[4]

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 framework proposed here allows us to model how any belief can become distorted.

## 2 General framework

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 which an agent forms a conclusion about. The agent observes a set of signals $s = (s_1, ..., s_m) \in S \subseteq \mathbb{R}^m$, which provide information about $\theta$. In the applications here, the signal will be unidimensional and correspond to success at a task (including a politician's performance in office).

voters can be come overconfident if they neglect the correlation between their information sources. However, these papers are mostly concerned with *over-precision* of beliefs rather than a self-assessment of how good one is.

[3]See also Lipnowski and Mathevet (2017), who study agents who intrinsically prefer to hold some beliefs (and what information they would like to receive as a result), but form their beliefs in a standard Bayesian fashion.

[4]Another recent paper which focuses on belief distortion by itself is Cheng and Hsiaw (2017), who show that not correctly doing joint updates on the credibility of information sources and the state they report on can lead to disagreement even among those seeing the same information (in a different order).

The variables $\theta$ and $s$ are drawn from a prior joint probability distribution $f(\theta, s)$. An actor in a standard model would form a conditional posterior belief about $\theta$ 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", or best guess about the state variable $\theta$. 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. Second, he may have a directional motive to reach certain conclusions.[5]

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

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

The $a(f_{\theta|s}(\theta|s))$ term represents the accuracy motive, where $a : \mathbb{R}^+ \mapsto \mathbb{R}$ is a strictly increasing function.[6] The second term represents the intrinsic value for holding conclusion $\theta$, 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 (or is constant in $\theta$). Such an agent picks a conclusion at the mode of the posterior distribution.

[5] Bénabou and Tirole (2016) similarly classify belief formation being driven by accuracy and *desirability*.

[6] Of course one could not explicitly model the signal process, and just write the Bayesian density of the joint distribution at the time of the conclusion as $f(\theta)$, and hence the accuracy term as $a(f(\theta))$.

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

**Definition** The *distortion* of conclusion $\tilde{\theta}$ is:[7]

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

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

**Comments on the setup** 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 about the world. 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.[8]

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.[9] More generally, the optimization problem as specified here

[7] If $f_{\theta|s}(\theta|s)$ is multi-modal then there are multiple potential ways to define the distortion, an issue which will not arise in any of the applications here.

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

[9] Another 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

serves as first approximation for any process of belief formation where both accuracy and some directional motive are at play.

**Core and auxiliary beliefs** Many results hinge on a distinction between beliefs that "matter" in the $v$ function. Formally:

**Definition** $\theta_i$ is an *auxiliary variable* if $v$ is constant in $\theta_i$. $\theta_i$ 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 $\theta_i$ will always be the value that maximizes $f_{\theta|s}(\theta_i, \tilde{\theta}_{-i}|s)$. That is, the most likely value of $\theta_i$ given the signal *and the conclusion about other variables* ($\tilde{\theta}_{-i}$). If $\theta_i$ is independent of the other variables conditional on $s$, this is the mode of the marginal posterior distribution of $\theta_i$. However, if $\theta_i$ is related to other beliefs, the conclusion chosen will depend on the conclusion about the state of the world writ large.

For core beliefs, a useful characterization in the applications is the following: Suppose $\theta_i \in \mathbb{R}$, and that $f_{\theta|s}$, $v$ and $a$ are continuous and differentiable.[10] Then any optimal conclusion $\tilde{\theta}_i$ must be a $\theta_i$ which solves the first order condition

$$\underbrace{\frac{\partial v(\theta_i, \tilde{\theta}_{-i})}{\partial \theta_i}}_{\text{Marginal Directional Motive}} = -\underbrace{\frac{\partial}{\partial \theta_i}\left[a(f_{\theta|s}(\theta_i, \tilde{\theta}_{-i}|s))\right]}_{\text{Marginal Accuracy Motive}} \tag{2}$$

$\theta$. 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. Here I stick with the choice of a single conclusion for clarity and tractability.

[10] If $\theta_i$ is bounded there can be a corner solution; without the continuity and differentiability assumptions then optimal conclusions may lie at discontinuities or kinks.

for each $i$. The main lesson (2) teaches is that an agent who prefers a higher conclusion about $\theta_i$ on the margin (the left-hand side is positive) must pick an a conclusion which would become more plausible if lowered ($f_{\theta|s}(\theta_i, \tilde{\theta}_{-i}|s)$ must be decreasing in $\theta_i$ at $\tilde{\theta}$). Similarly, if the agent intrinsically prefers a lower conclusion on the margin, a higher conclusion must be more plausible. So, as long as the agent has accuracy and directional motives which do not have the same optimal value, there must be a tradeoff between these goals.[11]

This leads to a general result about what happens when either the accuracy or directional motive becomes more important. 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. 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).[12]

[11] If, for example, the accuracy and directional motives are both maximized at a common $\theta^*$, then the optimal conclusion is $\theta^*$ and both sides of equation 2 are zero at this point.

[12] However, these studies do *not* find substantial increases in the accuracy of responses with monetary incentives.

**A natural $a$ function.** In all of the applications in this paper, I let the $a$ function be $a(\cdot) = \log(\cdot)$. That is, the accuracy motive is measured by the log-likelihood of the conclusion.

Using a logarithmic transformation will make the algebra work out nicely with the particular distributions used. However, doing so also has an important property which makes this transformation a natural choice regardless of the other distributional assumptions or application. Suppose $\theta_1$ and $\theta_2$ are independent is the posterior (conditional on $s$). Further, suppose they are also independent in the directional motive in the sense that they are additively separable; i.e., we can write the $v$ function as $v(\theta_1, \theta_2) = v_1(\theta_1) + v_2(\theta_2)$. In words, this means that the conclusion reached on $\theta_2$ does not affect the relative value of different conclusions on $\theta_1$.

If we let $a(\cdot) = \log(\cdot)$, then by the independence in the posterior belief the objective function becomes:

$$\begin{aligned}\log(f(\theta_1, \theta_2)) + v(\theta_1, \theta_2) &= \log(f_1(\theta_1) f_2(\theta_2)) + v_1(\theta_1) + v_2(\theta_2) \\ &= \log(f_1(\theta_1)) + v_1(\theta_1) + \log(f_2(\theta_2)) + v_2(\theta_2)\end{aligned}$$

Since all of the terms here are additively separable, these are essentially two independent maximization problems.

By contrast, if, for example, we use the identity function for the accuracy motive, then the optimal conclusion reached on $\theta_1$ will affect the optimal conclusion on $\theta_2$ even though both variables are independent both in the accuracy in the directional motive. In other words, using a $\log$ transformation ensures a kid of "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.

We now turn to some more specific applications of this modeling framework.

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 $\theta$ which is normal with mean $\mu_\theta$ and variance $\sigma_\theta^2$. He then observes a noisy signal of the quality, given by:

$$s = \theta + \epsilon \tag{3}$$

where $\epsilon$ is normally distributed with mean 0 and variance $\sigma_\epsilon^2$.

In this and later models, I employ two interpretations of this signal. In the first, $\theta$ 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 (e.g., job performance). For this interpretation I refer to $\epsilon$ as "luck". Call this the $\mathcal{ST}$ ("self test") interpretation.

For the second interpretation, $\theta$ 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.

For either interpretation, two key assumptions are implicit with this signal structure. First, that the signal is only driven by the quality $\theta$ and one other factor deemed "noise" or "luck". In section 4, I allow for other factors. Second, the agent knows the mapping between quality and noise to the signal. In particular, a unit increase in either quality or noise is associated with a unit increase in the signal. So, this formulation can not capture the core notion of attribution theory: that people must form inferences about not only the inputs which determine success (or any other outcome), but about which inputs are most important. This is addressed in section 5.

**The Bayesian belief** The standard Bayesian update on $\theta$ 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 $\phi$ is the PDF of a standard normal random variable.

Rearranging (3), any signal and conclusion about the quality imply a conclusion about the error term: $\tilde{\epsilon} = s - \tilde{\theta}$. Since the mode of the Bayesian belief is the same as the mean, the distortion of the quality conclusion is $d(\tilde{\theta}) = \tilde{\theta} - \mu_\theta^B(s)$. 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** Using the notation from the previous section, let $a(\cdot) = \log(\cdot)$. So, the optimal conclusion is given by

$$\tilde{\theta} = \arg\max_\theta \log f_{\theta|s}(\theta|s) + v(\theta) \tag{4}$$

For now, I only assume that $v$ is continuous and twice-differentiable.

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 $\mu_\theta^B(s)$. So, there are increasing marginal costs to shifting the conclusion away from the standard

Bayesian belief. In particular,

$$\log f_{\theta|s}(\theta|s) = k_1 - \frac{(\theta - \mu_\theta^B(s))^2}{2\overline{\sigma}_\theta^2} \tag{5}$$

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

The first order condition for $\tilde{\theta}$ is then:

$$v'(\tilde{\theta}) = \frac{\tilde{\theta} - \mu_\theta^B(s)}{\overline{\sigma}_\theta^2} \tag{6}$$

Recall the mean of the Bayesian posterior distribution is also the mode, so the distortion of the belief is $d(\tilde{\theta}) = \tilde{\theta} - \mu_\theta^B(s)$. Substituting this into (6) and rearranging gives an expression for the optimal distortion:

$$d(\tilde{\theta}) = v'(\tilde{\theta})\overline{\sigma}_\theta^2 \tag{7}$$

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. So, if there is a benefit to having the freedom to pick favorable conclusions, one might prefer to be less informed about their ability or the performance of politicians. People may be averse to seeking out information they worry will be negative not just because learning bad news is uncomfortable, but because it makes the tradeoffs in plausibility required to reach a more pleasant conclusion sharper.

In the $\mathcal{PN}$ interpretation, this result could provide a reason for politicians to make ambiguous promises to audiences pre-disposed to support them, a point which I revisit in section 6.

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:*
*i. $\tilde{\theta} > \mu_\theta^B(s)$,*
*ii. $\tilde{\theta}$ is increasing in $s$, but*
*iii. $d(\tilde{\theta})$ is decreasing in $s$.*

**Proof** Part i follows from (2), and part ii from implicitly differentiating the equilibrium condition. 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 (6), $d(\tilde{\theta}_1) = \tilde{\theta}_1 - \mu_\theta^B(s_1) > \tilde{\theta}_2 - \mu_\theta^B(s_2) = d(\tilde{\theta}_2)$ ∎

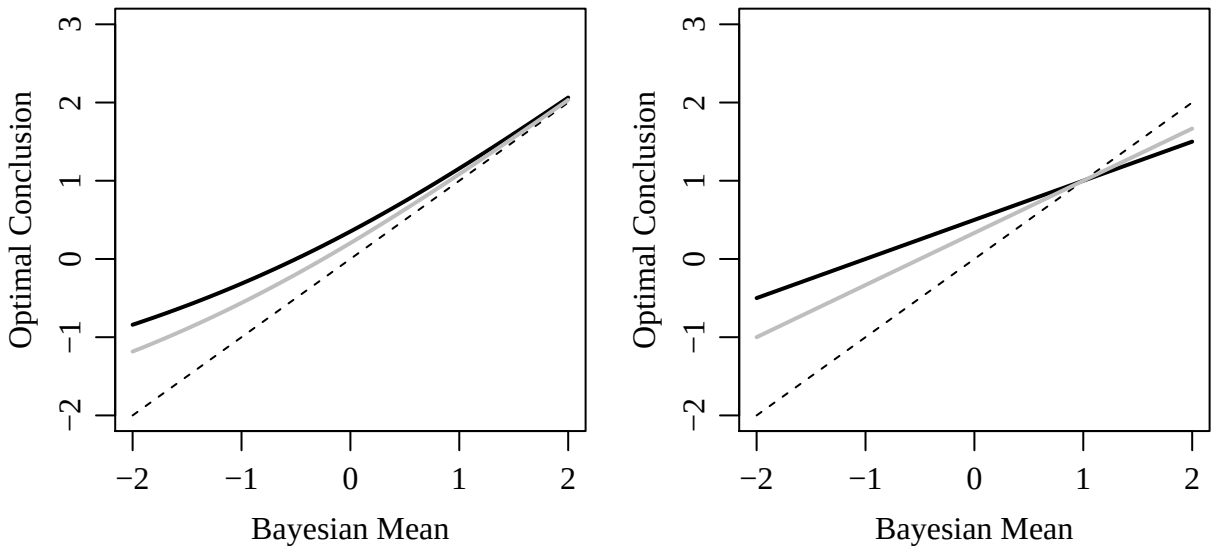
So, the conclusion moves in the "correct" direction as the signal of quality changes, but distortion relative to the Bayesian posterior is greater when the signal is low. More on this below.

**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. A natural way to model that is to suppose $v$ is a single-peaked function:

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

**Proof** See the appendix.

Figure 1: Optimal conclusions as a function of the Bayesian mean with increasing and concave (left), and single-peaked (right) $v$ function.




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 above this 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 $\theta$ 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 ("the error term"), 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 two standard deviations above the mean ($\theta = 2$). 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 $\tilde{\theta} = 1$.[13] 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* two standard deviations above the mean. So if the second signal is also typical, the neutral observer will be unsurprised by the agents 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 ($\theta^*$). 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.

## 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 account-

[13] This would be the optimal conclusion if, for example $v(\theta) = -\theta^2$ and $\overline{\sigma}_\theta = 1/2$; see (6).

ing identity. Luck was just the difference between success and ability, so increasing the conclusion about ability forced a change in the conclusion about luck.

Fortunately, the framework easily extends to belief distortion of less mechanically connected variables. This allows for a more satisfying model of how beliefs about multiple variables get distorted at the same time. Such a complication is important because success (and other observed outcomes) are generally driven by many factors.

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."[14] 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,[15] a belief which other religious groups do not hold.[16] In fact, another recent survey found that a majority of every racial group in the US believes they face discrimination.[17]

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

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

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

[16] Nor is this phenomenon limited to the United States. From Turkey: "another key element of Erdogan's religious narrative is the idea of victimization, playing on the fact that political Islamists were treated with prejudice by the Kemalist army in the past. Somehow, after over 12 years in power, Erdogan still manages to convince his constituency that they are a victimized group and that he is the most victimized of all. This narrative trick is used over and over." http://foreignpolicy.com/2015/08/12/how-president-erdogan-mastered-the-media/

[17] http://www.npr.org/sections/health-shots/2017/10/24/559116373/poll-most-americans-think-their-own-group-faces-discrimination

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). [18]

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

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

where $\delta$ represents the discrimination against the agent or the new source bias against the politician.[19] Suppose $\theta$, $\delta$, and $\epsilon$ are (in the prior) independent normals with means $\mu_\theta$, $\mu_\delta$, and $0$; and variances $\sigma_\theta^2$, $\sigma_\delta^2$, and $\sigma_\epsilon^2$.

**The Bayesian belief** The signal provides information about both the agent's ability and how much discrimination he faces. In isolation, these updates happen in a natural manner: succeeding more makes it more likely that one is high ability and does not face discrimination. More consequential is how the updates are related.

Formally, to compute the posterior belief about $\theta$ and $\delta$ given $s$, first write and then partition the covariance matrix of $(\theta, \delta, s)$ 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).

[18] See also study 2 in Davidai and Gilovich (2016) for evidence that those more invested in politics in the United States perceive more institutional bias against their preferred party.

[19] To be consistent with the notation in section 2 we could write the ability as $\theta_1$ and discrimination as $\theta_2$, but the connection to the general framework is clear and this notation allows for mnemonic interpretation.

The joint distribution of $(\theta, \delta)$ 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:

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

.

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

More important for our purposes, even though $\theta$ and $\delta$ 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

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

which is strictly positive, decreasing in $\sigma_\epsilon$, and approaches 1 as $\sigma_\epsilon \to 0$.[20]

[20] Intuitively, when $\sigma_\epsilon^2$ is very small, the agent knows that $s$ must be very close to $\theta + \delta$, and so conditional on $s$ any unit increase in the draw of $\theta$ must be associated with a one unit increase in $\delta$.

**The optimal conclusion** Suppose the belief about the quality $\theta$ 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 for those who *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.

As in the previous section let $a(\cdot) = \log(\cdot)$. Combined with the assumption that $\delta$ is auxiliary (and hence the directional term can be written $v(\theta)$), the optimal joint conclusion is:[21]

$$(\tilde{\theta}, \tilde{\delta}) \in \arg\max_{(\theta,\delta)} \log f_{\theta,\delta|s}(\theta, \delta|s) + v(\theta),$$

where $f_{\theta,\delta|s}$ is the bivariate normal density derived above, and so:

$$\log f_{\theta,\delta|s}(\theta, \delta|s) = k_2 - \frac{\left(\frac{(\theta-\mu_\theta^B(s))^2}{\overline{\sigma}_\theta^2} - \frac{2\rho(\theta-\mu_\theta^B(s))(\delta-\mu_\delta^B(s))}{\overline{\sigma}_\theta\overline{\sigma}_\delta} + \frac{(\delta-\mu_\delta^B(s))^2}{\overline{\sigma}_\delta^2}\right)}{2(1-\rho^2)}, \tag{8}$$

where $k_2$ collects the terms which do not depend on $\theta$ and $\delta$ and hence do not affect the optimization. Conveniently, (8) is quadratic in both $\theta$ and $\delta$.

[21] 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 $\theta$ and $\epsilon$.

The first order conditions for the optimal conclusion are:

$$v'(\tilde{\theta}) = -\left.\frac{\partial \log f_{\theta,\delta|s}(\theta,\delta|s)}{\partial \theta}\right|_{\theta=\tilde{\theta},\delta=\tilde{\delta}} \tag{9}$$

$$0 = \left.\frac{\partial \log f_{\theta,\delta|s}(\theta,\delta|s)}{\partial \delta}\right|_{\theta=\tilde{\theta},\delta=\tilde{\delta}}. \tag{10}$$

Writing out (10), solving for $\tilde{\delta}$, and bringing in the definition of the distortion inherent in $\tilde{\delta}$:

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

So, the distortion in the conclusion about discrimination/bias is a fraction of the distortion about the core quality $\theta$. Further, this fraction is the ratio of the covariance between $\theta$ and $\delta$ and the variance of $\theta$, i.e., the hypothetical regression coefficient for data drawn from the agent's posterior belief about the two variables. Why? For any conclusion about $\theta$, the agent will pick the $\delta$ which maximizes the density conditional on both $\theta$ and $s$: $f_{\delta|\theta,s}(\delta|\theta,s)$. This is the conditional mean of $\delta$ given $\theta$ and $s$. So, the joint conclusion about $\theta$ and $\delta$ always lies on the regression line.[22]

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 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

[22]This principle always holds if there are multiple core variables and one auxiliary variable. For a fixed conclusion about the core variables, the agent picks the value of the auxiliary variable that maximizes the posterior likelihood, which is the "predicted value" of the regression equation evaluated at the core variable values. However, for multiple variables, deriving the coefficients for this regression equation from the covariance matrix become less tidy, so the practical value of this observation diminishes.

and performance of the politician. This will tend to be true when the news source has reported a lot on the politician in question, so the authority of that reporting would make a neutral reader think what the news paper writes is highly informative about the politician.[23]

To complete the derivation of the optimal assessment, plugging the optimal conclusion about $\delta$ as a function of the conclusion about $\theta$ into (8) and simplifying gives:

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

for a constant $k_3$. Other than this constant (which differs from $k_1$ in (5), but also drops out when maximizing with respect to $\theta$), this expression is the same as the log likelihood of the marginal distribution of $\theta$. The optimal conclusion about $\theta$ (given the relationship between the optimal conclusions of $\theta$ and $\delta$) 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.[24]

Summarizing:

**Proposition 4.** *With uncertainty about both $\theta$ and $\delta$, the agent's optimal conclusion about these variables is equal to the Bayesian belief plus distortions which are characterized by:*

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

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

[23]Formally, the correlation will be high when $\sigma_\epsilon$ is low, i.e., the news source is very informative (if potentially biased) about the politician performance. The model in the next section considers the informativeness of news sources in more detail.

[24]The $\overline{\sigma}_\theta^2$ term in this section is different because it incorporates the noise associated with both $\epsilon$ and $\delta$.

**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 $\theta$ ($v'(\tilde{\theta})$), and how closely related this belief is to the auxiliary variable ($\overline{Cov}(\theta, \delta)$).

**Summary and empirical discussion** One way to interpret this comparison is that the agent does not form and incorrect belief about discrimination to help develop a better self-assessment, as he would reach the same conclusion if "ignoring" discrimination when forming a conclusion about ability. Rather, when trying to make sense of the joint facts about his ability and discrimination, his belief about discrimination gets "dragged along" by distortion of beliefs about ability. Similarly, disparate views of the reliability of news sources may be a consequence rather than a cause of divergent views about what they report on.

Revisiting the motivating example, diverging views of which groups face discrimination can arise from a common desire among all individuals to think are of high ability. And if there are diminishing marginal returns to higher conclusions about ability (i.e., $v$ is concave) this tendency will be strongest among the unsuccessful, a hypothesis which should be testable. In particular, the conclusion by White Christian males that they are held back by discrimination may be particularly alluring for those 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) before 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. So, 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.

## 5 Application 3: Attribution and news source quality

The final model shows how the framework can be used to model situations 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:

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

As above, the prior on $\theta$ is normal with mean $\mu_\theta$ and variance $\sigma_\theta^2$. In this section, let $\epsilon$ be a standard normal random variable (i.e., with variance 1). The $\omega$ parameter scales how much noise the signal contains. To simplify, suppose $\omega \in \{g, b\}$, $0 < g < b$. So, 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 $\theta$ and $\omega$, i.e., the quality and the degree to which the signal is driven by noise.[25] Continuing to use the log-likelihood formulation of belief plausibility, the optimal conclusion solves:

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

**The Bayesian belief** If the agent knew for sure how noisy the signal was (i.e., $\omega$), then Bayesian posterior belief would use the standard updating formulas employed in previous sections. Formally, conditional on $\omega = g$ or $\omega = b$, the belief about $\theta$ given $s$ is normal with mean

$$\mu_\theta^B(s,g) = \frac{\sigma_\theta^{-2}\mu_\theta + g^{-1}s}{\sigma_\theta^{-2} + g^{-1}} \quad \text{and} \quad \mu_\theta^B(s,b) = \frac{\sigma_\theta^{-2}\mu_\theta + b^{-1}s}{\sigma_\theta^{-2} + b^{-1}},$$

and the posterior variance

$$\overline{\sigma}_\theta(g)^2 = \frac{1}{\sigma_\theta^{-2} + g^{-1}} \quad \text{and} \quad \overline{\sigma}_\theta(b)^2 = \frac{1}{\sigma_\theta^{-2} + b^{-1}}.$$

Since the agent does now know for sure whether $\omega = g$ or $\omega = b$, the joint posterior belief about these variables is a normal mixture with density:

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

There are two pairs of terms in the density. The $f_{\theta|s,\omega}(\theta|s,\omega)$ terms are the beliefs about $\theta$ con-

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

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

$$Pr(\omega = g|s) = \frac{\pi \frac{1}{\sigma_s(g)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(g)}\right)}{Pr(s)} \quad \text{and} \quad Pr(\omega = b|s) = \frac{(1-\pi) \frac{1}{\sigma_s(b)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(b)}\right)}{Pr(s)}.$$

(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 $\theta$ and $\omega$ 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 $\omega$, the optimal conclusion about $\theta$ 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:

$$\begin{aligned} Pr(\omega = g|s) f_{\theta|s,\omega}(\mu_\theta^B(s,g)|s,g) &\geq Pr(\omega = b|s) f_{\theta|s,\omega}(\mu_\theta^B(s,b)|s,b) \\ \frac{\pi \frac{1}{\sigma_s(g)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(g)}\right)}{Pr(s)} \frac{1}{\overline{\sigma}_\theta(g)} \phi(0) &\geq \frac{(1-\pi) \frac{1}{\sigma_s(b)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(b)}\right)}{Pr(s)} \frac{1}{\overline{\sigma}_\theta(b)} \phi(0) \\ \frac{\pi}{1-\pi} \frac{\overline{\sigma}_\theta(b)}{\overline{\sigma}_\theta(g)} &\geq \frac{\frac{1}{\sigma_s(b)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(b)}\right)}{\frac{1}{\sigma_s(g)} \phi\left(\frac{s-\mu_\theta}{\sigma_s(g)}\right)} \end{aligned} \tag{15}$$

When the two ratios on the left-hand side of (15) 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 $\pi$, low $1-\pi$), this conclusion is more likely.

Less obvious, the second ratio is the standard deviation of the posterior belief about $\theta$ 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 $\theta$, and believing the test had low noise allows for a more precise estimate.

To clarify this intuition, it is instructive to ponder a discrete analog. Suppose ability is either below average ($\theta = -1$), average ($\theta = 0$), or above average ($\theta = 1$), with equal prior probability. An accurate test returns exactly $s = \theta$, while a bad test returns each result with equal probability. For any result, there are four possibilities: the test is accurate and therefore $\theta = s$, or the test is inaccurate and the ability level is any of the three possibilities. Since accurate tests are always correct, the posterior probability assigned to the test being accurate and $\theta = s$ is $\pi$.[26] The posterior probability assigned to the test being bad and each of the three possible ability levels is $(1-\pi)(1/3)$.[27] Rearranging, as long as $\pi > 1/4$, the accurate test conclusion is the most plausible. So, even if the *marginal* conclusion about $\omega$ lends more credence to a noisy test ($\pi < 1/2$), the most plausible *joint* conclusion about the test and the test-taker may be that the test is accurate and the takers' ability is exactly equal to what the test indicates.

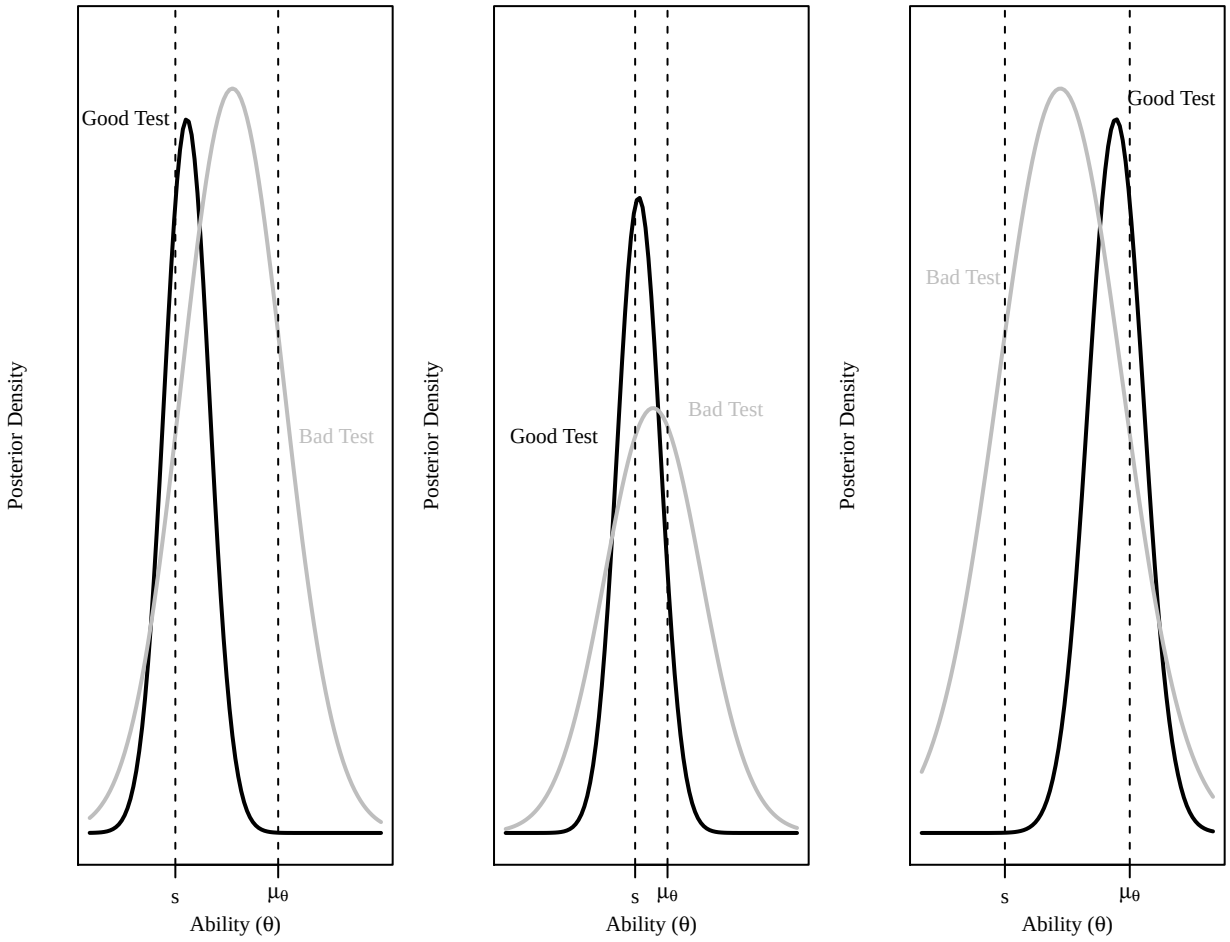
Interpreting ability-as-auxiliary as the case of assessing others, this is consistent with a key part of the fundamental attribution error (Ross, 1977). If we want to form inferences about the ability of others and just want them to be plausible, there is a bias towards thinking that outcomes are driven 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 (15), which is the relative likelihood of observing $s$ under

[26] Formally, $Pr(\omega = g, \theta = s|s) = \frac{(1/3)\pi}{(1/3)} = \pi$

[27] Formally, $Pr(\omega = b, \theta|s) = \frac{(1-\pi)(1/3)(1/3)}{1/3} = (1-\pi)(1/3)$

Figure 2: The Bayesian posterior belief in the attribution model with a low, medium, and high signal.




the low or high noise state. This will be high when $s$ is close to $\mu_\theta$, and low when $s$ is far from $\mu_\theta$. 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.

Figure 5 illustrates these observations for a case where the test is unlikely to be accurate in the prior. Each panel plots the Bayesian posterior belief about $\omega$ and $\theta$. The black curve represents the $\omega = g$ component of the mixture, and the grey curve the $\omega = b$ component. In each panel, the area

under the curve for the $\theta = b$ curve is greater, meaning if the agent only formed a conclusion about the test he would conclude it was noisy for each signal. The optimal conclusion is the highest point among the two curves. The fact that the "good test" curve is steeper illustrates the general tendency to conclude the test was accurate. However, in the left and right panels, this is outweighed by the fact that the signal is extreme (and the prior indicates it is not accurate), and so the optimal conclusion is that the test is bad, and the ability is at the peak of the grey curve. For the middle panel, where the signal is less extreme, the observer concludes the test was accurate and the ability is almost exactly what the signal indicates.

Formally:

**Proposition 5.** *Suppose $\theta$ and $\omega$ 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 assessment $(\tilde{\omega}, \tilde{\theta}) = (b, \mu_\theta^B(s, b))$ 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 $(b, \mu_\theta^B(s, b))$ otherwise.*

**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 $\mu_\theta$). 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%.[28] 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. This result can be interpreted as a continuous analog to the empirical finding that people are more apt to attribute success to ability and not luck when the results conform with expectations (Feather, 1969).

[28]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 $0.9$. Since $\pi = 1/2$, the average probability of a low noise assessment is around $.87$.

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 next 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 $a(\cdot) = \log(\cdot)$ and $v(\theta) = \alpha\theta$, for $\alpha > 0$. So, the agent always wants a higher conclusion about $\theta$, and $\alpha$ scales the magnitude of this preference.

A two step procedure determines the optimal overall conclusion. First, compute the optimal conclusion about $\theta$ 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 $\theta$ 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 $\theta$, and the problem is globally concave, with maximizer

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

The optimal conclusion is equal to the Bayesian belief plus a distortion. The distortion is larger when the agent cares a lot about forming a high belief (high $\alpha$) and there is more uncertainty about ability ($\overline{\sigma}_\theta(\tilde{\omega})^2$).

When is this conclusion higher under a belief that the test is accurate versus noisy? It depends.

The distortion is always greater under the high noise assessment, as it is easier to shift one's belief when more uncertain. However, the Bayesian belief is higher for a good test if and only if $s > \mu_\theta$. That is, conditional on doing well, there is a tendency to believe success is mostly driven by ability. Combining, the agent will only have a higher self-assessment of ability if assessing the test as low noise when doing substantially better than expected. But if the signal is too much higher than average, a good test becomes very implausible, and so the accuracy motive may lead back to a high-noise conclusion.

To see how these effects shake out, we need to complete the derivation of the optimal conclusion. The objective function evaluated at $\theta = \tilde{\theta}(\tilde{\omega})$ simplifies to:

$$\begin{aligned}
&\log\left(Pr(\tilde{\omega}|s)\right) - \log(Pr(s)) + \log\left(\frac{1}{\overline{\sigma}_\theta(\tilde{\omega})}\phi\left(\frac{\tilde{\theta}(\tilde{\omega}) - \mu_\theta^B(s,\tilde{\omega})}{\overline{\sigma}_\theta(\tilde{\omega})}\right)\right) + \alpha\tilde{\theta}(\tilde{\omega}) \\
&= \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(\tilde{\omega}))Pr(\omega|s)}{\overline{\sigma}_\theta(\tilde{\omega})}\right) - \log(Pr(s)) + \alpha\left(\mu_\theta^B(s,\tilde{\omega}) + \alpha\overline{\sigma}_\theta(\tilde{\omega})^2\right). \qquad (16)
\end{aligned}$$

The accurate test conclusion is now preferred if and only if (16) evaluated at $\tilde{\omega} = g$ is higher than it is when evaluated at $\tilde{\omega} = b$, which simplifies to:

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

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.

The log formulation again proves convenient, as both sides are quadratic functions of $s$ (see the Appendix for detail). 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 assessment 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 $(\underline{s}, \overline{s})$ such that the optimal assessment 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 $(b, \mu_\theta^B(s,b) + \alpha\overline{\sigma}_\theta(b)^2)$ otherwise, where*

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

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

**Proof** See the appendix.

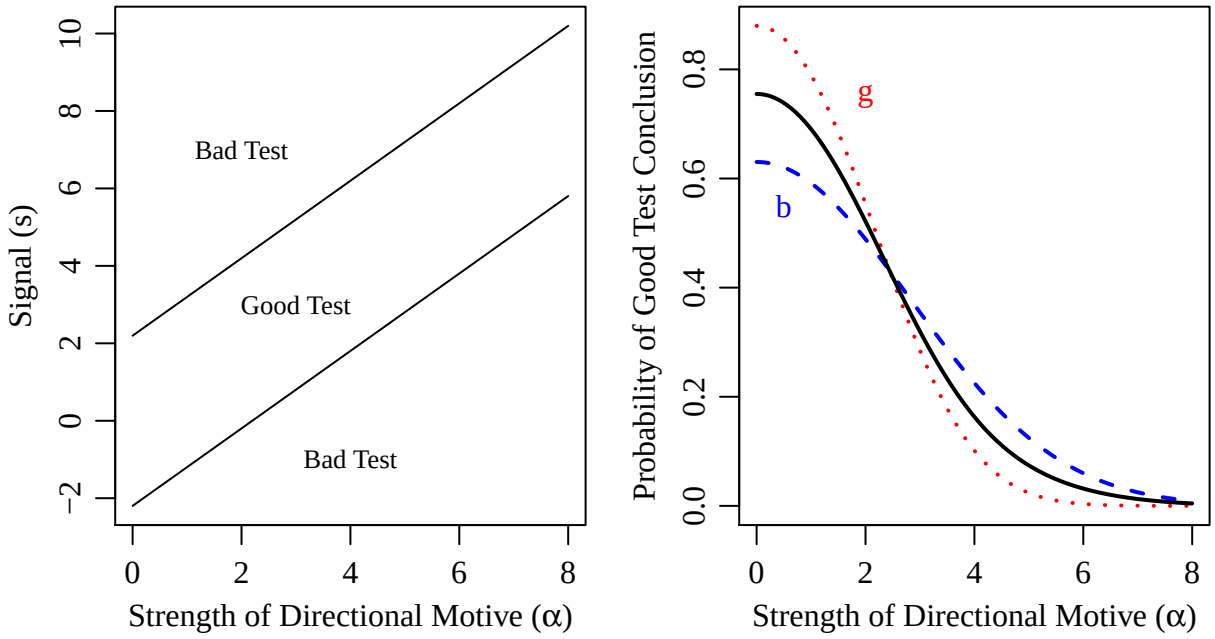
In words, unless the prior belief that the signal is noisy is strong enough to force this conclusion, then there is a "window" of signals where the agent thinks the test is accurate. This window is increasing in his desire to have a high self-evaluation, though the length of the window is constant in $\alpha$.

**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 $\alpha$ 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. At $\alpha = 0$, the agent does not care at all about $\theta$, i.e., the case where the ability conclusion is auxiliary. The range of signals accepted as low noise are symmetric around $\mu_\theta = 0$. Again, this illustrates the fact that, absent other considerations, people are apt to think tests accurately measure ability when their performance is close to expectations (Feather, 1969).

As $\alpha$ 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

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




agent wants to believe they are high ability, extremely high signals always lead him to conclude that the test does not measure ability well.[29]

The right panel plots the probability of a signal which leads to a low noise conclusion as a function of $\alpha$. 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 $\alpha$. 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

[29]Technically, this follows from the fact that the plausibility of the signal being low-noise falls faster than the advantage of believing the (positive) signal is accurate.

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 $\alpha$. 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 $\alpha$ 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". 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. This seems consistent with an informal observation that all politicians tend to complain about the accuracy of the media (and, given the model in the previous section, biased against them).

## 6 Overview and other ideas

The applications in this paper are admittedly scattershot. 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.

Rather than recap the disparate results again, here are some brief thoughts on how to extend or apply this technology to other questions.

**Ambiguity (in campaigns)** As alluded in the discussion surrounding equation 7, the ease of shifting beliefs about variables about which one is uncertain could imply an advantage to ambiguity. Suppose two candidates are competing in a primary, a situation where voters usually want to like whichever candidate wins. If one of the candidates makes their positions very clear, voters will be able to see differences between these positions and their preferred policy. This will be good for voters with very similar ideal points, but potentially harmful for those further from the candidate. However, voters who want to like the politician (or, perhaps, anticipate having to support her in the general election) may generally prefer candidates whose views are vague, making it easier distort their beliefs about her position closer to their own. As long as there is enough heterogeneity of beliefs among primary voters, the benefits of ambiguity in seeming palatable to a wider audience may outweigh the costs of not being strongly preferred by those with very similar preferences.[30] This notion is consistent with experimental results from Tomz and Van Houweling (2009), who find that "when voters encountered an ambiguous candidate from their own party, they expected the candidate to lean significantly in their own direction, instead of implementing a policy at the center of the candidate's ambiguous platform." In aggregate, a heterogenous audience behaving in this way may collectively prefer a more ambiguous candidate.

Similar dynamics could suggest why people may want to avoid collecting information when they think the implications may be unpleasant. For example, those nervous that a favored politician is not performing well may avoid informative news sources. More broadly, if one one cares a lot

[30]Perhaps an opposite effect applies to generally disliked politicians. For example, my reading of people's evaluation of Hilary Clinton's ideology is that many more left-wing democrats persuaded themselves that she is very centrist, while centrist democrats and conservatives seemed to genuinely think she is extremely liberal.

about reaching a particular conclusion, learning more may lead to an unacceptable risk.

**Attribution and downward spirals** Suppose an agent repeatedly chooses how much to invest in his success, and then success is driven by effort and luck. Those who perform poorly early in life may conclude that success is driven by luck, inducing them to exert even less effort in the future. This can create a dynamic process where those who are unlucky early in life perpetually exert low effort.

**Microfounding the $v$ function; What happens next?** Another way to extend the model would be to endogenize the $v$ function. 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 over-precision 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.

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 tied to the explanation for the directional motive on ability 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.

In either application, a natural next step would be to combine the model of belief formation here with either an endogenous directional motive, or more analysis of decisions made after beliefs are formed. Hopefully a rigorous and tractable formulation of the intermediate stage of belief formation provided here makes these directions plausible.

## Appendix

**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 $\tilde{\theta}^2$:

$$w_v^1 v^1 + w_a a^1 \geq w_v^1 v^2 + w_a a^2 \tag{18}$$

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

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

If $v^1 \leq v^2$ and $a^1 \leq a^2$ and at least one of the inequalities is strict, then (18) 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 (19) 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 $w_v^2$, it must also be worth it under $w_v^1$. Formally, we can rearrange (18) 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 (19) (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 $\theta$ 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 $\theta^*$ 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:

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

**Proof of Proposition 5.** It is immediate that the left-hand side of (15) 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 $\mu_\theta$) where (15) 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. ∎

**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:

$$\begin{aligned} DO =& \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(b))(1-\pi)\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)}{\overline{\sigma}_\theta(b)\overline{\sigma}_s(b)}\right) - \log\left(\frac{\phi(\alpha\overline{\sigma}_\theta(g))\pi\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)}{\overline{\sigma}_\theta(g)\overline{\sigma}_s(g)}\right) \\ &+ \alpha\left(\mu_\theta^B(s,b) + \alpha\overline{\sigma}_\theta(b)^2\right) - \alpha\left(\mu_\theta^B(s,g) + \alpha\overline{\sigma}_\theta(g)^2\right) \\ &= k_5 + \log\left(\frac{1-\pi}{\pi}\right) + \alpha(\mu_\theta^B(s,b) - \mu_\theta^B(s,g)) + \log\left(\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)\right) - \log\left(\phi\left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)\right) \\ &= k_5 + \log\left(\frac{1-\pi}{\pi}\right) + \alpha(\mu_\theta^B(s,b) - \mu_\theta^B(s,g)) + \left(\frac{s-\mu_\theta}{\overline{\sigma}_s(g)}\right)^2 - \left(\frac{s-\mu_\theta}{\overline{\sigma}_s(b)}\right)^2 \end{aligned} \tag{20}$$

where $k_5$ collects terms which are not a function of $s$ or $\pi$. Equation (20) 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 $\pi$ 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 $\pi^*$ such that there are real roots to (20) 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 $\frac{\partial \underline{s}}{\partial \alpha} - \frac{\partial \overline{s}}{\partial \alpha} = 0$ ∎

## 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.

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.

Cheng, Ing-Haw and Alice Hsiaw. 2017. "Distrust in Experts and the Origins of Disagreement." Manuscript.

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. Technical report National Bureau of Economic Research.

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

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.

Lipnowski, Elliot and Laurent Mathevet. 2017. "Disclosure to a Psychological Audience." Manuscript.

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

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.

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

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.

Tomz, Michael and Robert P Van Houweling. 2009. "The electoral implications of candidate ambiguity." *American Political Science Review* 103(1):83–98.

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.