BELIEF UPDATING WITH DISSONANCE REDUCTION

KYLE P. CHAUVIN

MAY 31, 2021

ABSTRACT. This paper models a dissonance-reducing learner. The learner suffers utility loss when faced with mutually conflicting information and, in response, distorts how she perceives different pieces of information to increase the apparent agreement among them. Relative to Bayesian updating, the learner always appears partisan. In two-sided debates, she champions one side by exaggerating arguments that support it and downplaying opposing arguments. Her choice of favored side is both influenced by objective evidence and skewed by hotly-debated issues. Her posterior belief is inordinately extreme. Applications to dynamic learning, networks, persuasion, and information-seeking further characterize the learner’s behavior.

JEL CLASSIFICATION: D03, D74, D80, D83

KEYWORDS: Motivated Reasoning, Cognitive Dissonance, Non-Bayesian Updating

Department of Economics, Harvard University. Email: kchauvin@hbs.edu. My thanks to Roland Bénabou, Christine Exley, Tristan Gagnon-Bartsch, Faruk Gul, Anna Hopper, Alex Jackobsen, Korhan Koçak, Kathleen McGinn, Matthew Rabin, Aditi Sengupta, Joshua Schwartzstein, Tomasz Strzalecki, Michael Thaler, Huseyin Yildirim, Ben Young as well as participants at various conferences for helpful comments.
1. Introduction

People often learn by aggregating information from multiple sources. In politics, science, business, ethics, etc., we garner information from newspapers, television channels, websites, social media, teachers, colleagues, family, friends, and other sources. Sometimes, different sources are mutually consistent; in other cases, they disagree. A Bayesian updater takes agreement and disagreement in stride, shifting probability mass from one state to another in accordance with the totality of evidence. People, however, dislike inconsistency. As documented by psychology research on cognitive dissonance, inconsistency ‘upsets us and it drives us to action to reduce our inconsistency. The greater the inconsistency we face, the more agitated we will be and the more motivated we will be to reduce it’ (Cooper, 2007). Moreover, one of the ways people redress inconsistency is by subjectively altering their beliefs.

This paper analyzes a model that incorporates dissonance reduction into Bayesian updating. It makes three principal contributions. Using tools developed in a concurrent paper (Chauvin, 2020), I provide a simple and portable way to quantify the agreement between sources and measure the distortion of information. Comparing a dissonance-reducing agent to a Bayesian updater illustrates a series of effects – partitioning one’s sources into friends and enemies, overweighting friendly sources, etc. – that collectively amount to a form of disciplined partisanship. Lastly, the paper analyzes the implications of dissonance reduction in several applied settings, including dynamic learning, information seeking/avoidance, persuasion, and social learning in networks.

A simple example illustrates the basic mechanics of the model. Suppose an agent learns about a contentious political issue with a left-right spectrum of positions, e.g. Does free trade benefit the domestic economy? or Is gun control effective? Initially the agent has no information on

---

1 Introduced in Festinger (1957), cognitive dissonance theory posits that individuals experience distress from the tension between incongruent actions and opinions and that people will change either or both in an attempt to mitigate the dissonance. The subsequent literature is too voluminous to be adequately reviewed here, but see Cooper (2007) for a comprehensive summary. However, it is worth noting one conceptual divide between the cognitive dissonance literature and the present paper: most studies consider tension between a person’s actions and their beliefs; because in this paper I restrict attention solely to the mechanics of learning, actions are not discussed, and tension exists strictly between pieces of information (‘arguments’). Notwithstanding, it should be recognized that each argument supports a particular conclusion, and thus implicitly recommends a particular course of action.

2 The framework builds from the fact that Bayesian updating can be represented with beliefs as points in Euclidean space and information as vectors that shift the space of beliefs. Vector scaling provides a useful way to conceptualize information distortion; an inner product quantifies agreement. Although the modeling approach is novel, certain components have precedent in prior literature, e.g. see Molavi, Tahbaz-Salehi and Jadbabaie (2018) and Benjamin, Bodoh-Creed and Rubin (2019) as instances of distortion via scaling.
the matter, but then she hears arguments from three friends. All of the arguments are of equal strength, but two of them push the agent’s belief to the right while one pushes leftwards. The agent dislikes the disagreement between the arguments, so she adjusts her perception of them. She spends more time with some friends, thinks less about some arguments, and convinces herself that some arguments are stronger than in reality, etc. These actions are costly, so she engages in a limited amount, and because the balance of arguments is right-leaning, it is cost-effective for her to exaggerate the right and dampen the left. She thus perceives two strong right-leaning arguments and a weak left-leaning one, and shifts her posterior belief on the matter rightwards by strictly more than a Bayesian would. In this way, dissonance reduction has caused the agent to act like a partisan, systematically favoring one side over the other, despite her ex-ante neutrality.

Section 2 presents the core model. As in the example, an agent reasons about an unobserved binary state. To simplify notation as well as provide key geometric intuition, I treat the agent’s beliefs as log odds-ratios, which means the set of all potential beliefs is represented by the real line. Each point on the line corresponds to a particular distribution over \( \{L, R\} \); the origin represents the uniform distribution and points further to the right correspond to greater weight on \( R \). According to Bayes’ rule, each unit of information, or argument, changes the agent’s beliefs by shifting her along the spectrum some amount to the right or to the left.\(^3\) Like a vector, each argument is defined by its direction and magnitude. The agreement between two arguments is proportional to the product of their magnitudes; it is positive if the arguments share the same direction and negative if they have opposite directions.

The agent starts at the origin and is presented with a set of arguments. As determined by her distaste for dissonance, she distorts her perception of the information by scaling each argument by some real-valued factor. Scaling an argument by 1 leaves it unchanged; scaling up by a factor greater than 1 increases the argument’s magnitude while maintaining its direction; scaling down below 1 decreases magnitude; scaling below 0 reverses direction. The agent incurs a cognitive cost for choosing a scaling factor further away from the neutral value of 1. Her net utility is the sum

\(^3\)As used here, an argument is equivalent to the realization of a Bayesian signal. Because it implicitly supports a particular course of action, an argument operates like a ‘cognition’ as understood in the cognitive dissonance literature as well as a ‘reason’ (Shafir, Simonson and Tversky, 1993) for making decisions. Among others, Shiller (2017) and Bénabou, Falk and Tirole (2019) use ‘narratives’ to refer to ‘stories people tell themselves . . . to make sense of human experience’; note that an argument is broader in scope, because it is purely informational and not necessarily a tool of persuasion.
of perceived agreement among all argument pairs minus the sum of scaling costs. As discussed in Appendix B, the functional form of utility both satisfies intuitive properties inherent in the everyday concept of agreement and facilitates tractable analysis.

The paper’s main results characterize the agent’s optimal choice of scaling factors and her subjective posterior belief. First, for any profile of arguments, the agent acts as a partisan: she identifies with either the left or the right; she scales up arguments on her chosen side; and she scales down those on the opposing side. Arguments from sources with lower scaling costs are scaled more dramatically; in the extreme, arguments from the opposing side provoke backlash and are scaled negatively. The agent’s posterior, equal to the composition of the distorted arguments, is strictly more extreme than the Bayesian posterior. Despite these departures from objective learning, the agent’s chosen side always coincides with the objective balance of evidence. This is because, being ex-ante neutral, the agent seeks the most cost-effective way to increase agreement, and that involves amplifying the largest subset of arguments already in mutual agreement.

Section 3 extends the model to a dynamic framework. In each period, the agent re-evaluates her current belief in light of a newly arrived argument. This process leads to more dramatic scaling for arguments received earlier in time, causing the agent’s beliefs to follow a path-dependent process that eventually places full weight on one of the two states. Conditional on the true state, both the correct and incorrect conclusions have strictly positive probability ex-ante. There are particular signal distributions such that the agent learns the wrong state with arbitrarily high probability. In contrast to confirmation bias (Rabin and Schrag, 1999), the marginal effect of a new argument is independent of the agent’s current belief, and the dissonance-reducing agent is subject to more dramatic conversions. For new arguments which oppose her current belief but have small magnitude, the agent advances in the direction of her current belief, but for a sufficiently large opposing argument, the agent instead advances in the direction of that argument. In this way, the agent converts from being one side’s partisan to being the other’s, resembling a real-life voter who switches political parties and accordingly shifts her source of news content, or a religious convert who subsequently discounts any information that opposes the tenets of her adopted belief.

Section 4 considers a version of the static model in which the agent reasons simultaneously about multiple issues. The agent’s belief is now represented as a point in higher dimensional Euclidean space, with arguments acting as vectors of the same dimension. The main results in one
dimension generalize, but an additional effect called *issue skewing* is also at work. To illustrate, suppose arguments debate both the merit of free trade as well as a frivolous political scandal. Although the agent’s criterion for calculating agreement places low weight on the scandal topic, the arguments feature extreme, heated positions on that dimension, having a large impact on the level of agreement. Consequently, the agent’s determination of which sources to exaggerate hinges disproportionately on the scandal. With respect to trade policy, the agent’s posterior is governed not by the direction of the balance of evidence, as in the one-dimensional case, but by the skewing influence of the scandal topic.

Section 5 studies the consequences of dissonance reduction in three areas of application. First, I consider a simple version of the network learning model as studied by, for example, DeMarzo, Vayanos and Zwiebel (2003) and Golub and Jackson (2010). Although dissonance-reducing agents endowed with conflicting initial arguments may initially scale down what they hear from their peers, repeatedly hearing and re-scaling one’s neighbors’ opinions leads to asymptotic convergence as is typically found in the literature. Unlike most network models, however, the fact that different agents have different numbers of neighbors and scale up what they hear in different proportions leads to variation in the speed of convergence towards the consensus. Next, I show that an agent always gains in expected utility from acquiring new information. In contrast to cases of information avoidance (Kőszegi, 2006; Oster, Shoulson and Dorsey, 2013; Golman, Hagmann and Loewenstein, 2017) in which people fear a bad state of nature, the agnosticism of a purely dissonance-reducing agent means she anticipates molding future arguments to support *some* conclusion, and therefore exhibits curiosity (Ely, Frankel and Kamenica, 2015; Golman and Loewenstein, 2015) for new information. Finally, I analyze the decision of a strategic principal seeking to influence a dissonance-reducing agent towards a particular goal belief. When constrained to provide the agent with an argument of a fixed, small magnitude, the principal sends a compromise argument that, while supporting the goal belief on each issue, is scaled in accordance with the agent’s existing belief.4

This paper contributes foremost to the literature on *motivated reasoning* (Abelson, 1986; Kunda, 1990), which studies how preferences over beliefs distort learning. The agent featured here is

---

4That the principal capitalizes on the agent’s bias by presenting him with something that both advances the principal’s cause while appealing to the agent echoes analogous results in other models of persuasion of a behavioral agent. For example, in Schwartzstein and Sunderam (2019) and Eliaz, Spiegler and Thysen (2020), principals strategically choose ways of framing data that agents find highly plausible.
nonetheless distinguished from the majority of prior studies in being indifferent to the underly-
ing state of the world, lacking any desire to believe in things such as her own ability (Kőszegi,
2006; Gottlieb, 2014), moral character (Bénabou and Tirole, 2011), future happiness (Caplin and
Leahy, 2001; Brunnermeier and Parker, 2005; Mayraz, 2011; Bénabou, 2012; Bracha and Brown,
2012), or the correctness of past decisions (Eyster, 2002; Yariv, 2005). Indeed, the most prominent
treatments of cognitive dissonance reduction in economics (Akerlof and Dickens, 1982; Rabin, 1994;
Konow, 2000) feature agents who wish to take a particular action – e.g. foregoing workplace safety
equipment or eating unethically treated animals – but also want to believe that the action is wise
or moral.

Relative to the motivated reasoning literature, a distaste for dissonance per se both reproduces
several non-Bayesian features while predicting a high degree of factual discipline. Apart from
confirmation bias-like behavior, discussed above, the agent also exhibits over-precision, the most
consistently documented form of overconfidence (Kahneman and Tversky, 1979; Moore and Healy,
2008). However, the mechanism is distinct: instead of a desire to view oneself as competent,
the agent effectively has a taste for strong conclusions. The agent shows partisan bias (Sunstein
et al., 2016), the over-trusting of information from one’s preferred side, but this comes about as a
byproduct of the taste for consistency. In contrast to a fixed partisan status, and unlike models
of rational quality-evaluation of information sources (Gentzkow and Shapiro, 2006), the agent
can transition from amplifying one side to emphasizing the other. The agent’s seeming revealed
preference for additional information is not out of innate curiosity (Golman and Loewenstein, 2015;
Ely, Frankel and Kamenica, 2015) but rather due to the agent anticipating being able to use
additional information as a way to bolster the agreement among existing arguments. The agent
does not exhibit a ‘good news / bad news’ effect (Eil and Rao, 2011; Zimmermann, 2020); when
the sides of a debate have positive and negative valence, the agent is just as eager to embrace
pessimistic voices when they predominate. As discussed in Section 6, the agent’s preference to
perceive agreement among sources is mathematically equivalent to a motive for conformity to her
sources.

Several recent studies also feature endogenous over- and under-weighting of information sources.
In Gentzkow, Wong and Zhang (2018), people with different priors about state will diverge in their
beliefs about the trustworthiness of different sources, further exacerbating disagreements on issues.
Relative to their quasi-Bayesian approach, the present paper demonstrates how ostensible trust and
distrust can emerge without exogenous disagreements between the learner and her sources. Cheng
and Hsiaw (2018) study an agent processing a source’s signal by first ‘pre-screening’ to update
about the source’s credibility given the agent’s current beliefs and then updating on the signal
content having fixed an opinion about the source’s credibility. This also leads to overconfidence
and confirmation bias in a sequential setting, but the agent in the present paper differentiates
sources even in a static framework.

2. Model

2.1. Mechanics

An agent learns about an unobserved state $\omega \in \{L, R\}$ (‘left’, ‘right’). She initially places equal
weight$^5$ on both states, and then she receives information from sources $i = 1, \ldots, N$. I assume
that no source delivers fully revealing information and that the sources deliver (conditionally)
independent signals. Below, I first discuss how an unbiased agent would process the information
and then consider the effect of dissonance reduction.

**Baseline: Unbiased Agent.** The agent’s beliefs are expressed as log odds-ratios, a notational
feature that provides critical intuition. The real line serves as a left-right spectrum of beliefs; belief
$\pi$ corresponds to placing weight $1/(1 + e^\pi)$ on state $R$, so that $\pi \rightarrow \infty$ corresponds to full weight
on $R$ and $\pi \rightarrow -\infty$ corresponds to full weight on $L$. By starting with equal weight on both states,
the agent has a prior log odds-ratio of $\pi_{\text{prior}} = 0$. The information communicated by each source is
called its *argument*, identified by a log likelihood-ratio $a_i \in \mathbb{R}$. Per Bayes’ rule, the unbiased agent
transitions to the objective posterior

$$
\pi_{\text{post}}^{\text{obj}} = \pi_{\text{prior}} + a_1 + \cdots + a_N = \sum_{i=1}^N a_i.
$$

Expressing Bayes’ rule in this form interprets information as vectors that push the agent along the
spectrum of beliefs. Each argument $a_i$ has magnitude $|a_i|$ and is either left-leaning ($a_i < 0$) or

---

$^5$Section 2.6 discusses how a non-neutral prior affects the main results. In brief: an agent’s non-neutral prior can
be understood as the consequence of information she received before participating in the model. If the agent is
able to scale her pre-model information at the same time she scales new arguments, dissonance reduction leads to a
reinforcement of the pre-model information. Section 3 further explores this effect in an extension of the main model
to a sequential framework.
right-leaning ($a_i > 0$). It can be ‘large’ or ‘small’ depending on how much it moves the agent’s belief.

**Bias.** A *dissonance-reducing* agent is motivated by a distaste for conflicting information. Intuitively, she experiences utility loss from receiving a mix of right-leaning and left-leaning arguments, and this is exacerbated when the arguments are larger. Formally, the level of *agreement* between two arguments $a_i$ and $a_j$ is defined by $\langle a_i, a_j \rangle = a_i \cdot a_j$. Two arguments have positive agreement if and only if they share the same direction, and the absolute value of their agreement is proportional to the magnitude of each argument. The total net agreement that the agent experiences on account of the sources’ arguments is the sum $\sum_{i,j=1}^{N} \langle a_i, a_j \rangle$. As discussed in Appendix B, this functional form both satisfies intuitive properties inherent in the everyday concept of agreement and facilitates tractable analysis. In particular, it guarantees that the perceived agreement among arguments is unchanged when one or more arguments are decomposed into pieces. For example, if source $i$ sends $a'_i$ and $a''_i$ instead of $a_i$, where $a'_i + a''_i = a_i$, the agent still perceives the same level of agreement. One notable consequence of this property is that the agent perceives agreement between each argument and itself.

In order to ameliorate the utility loss from dissonant arguments, the agent misperceives the sources’ information. Instead of perceiving argument $a_i$, the agent chooses a *scaling* factor $s_i \in \mathbb{R}$ and consequently perceives $s_ia_i$. Setting $s_i = 1$ corresponds to treating source $i$ objectively. If the agent sets $s_i > 1$, she amplifies source $i$’s argument by hearing a more powerful version of its objective counterpart. If she sets $0 \leq s_i < 1$, the agent dampens $i$’s argument. Finally, if the agent sets $s_i < 0$, then she reverses $i$’s argument, perceiving an objectively right-leaning argument as a left-leaning one or visa-versa. For a given profile of scaling factors $s_1, \ldots, s_N$, the agent’s *subjective posterior* aggregates the distorted arguments.

$$\pi_{\text{post}} = \sum_{i=1}^{N} s_ia_i.$$  

The cost of misperception is specified by a quadratic loss function, which models the psychological effort needed to delude oneself. There is no cost associated with scaling an argument by the neutral value of $s_i = 1$, so the $i$’th component of the cost function is proportional to $(1 - s_i)^2$. The proportional coefficient is specified to include two different components: a source specific *malleability* parameter $\gamma_i \in (0, 1)$ and a normalizing constant $c_i(a)$ which captures the agent’s cognitive
load. Modeling the idea that larger arguments are more difficult to distort, \( c_i(\cdot) \) increases in the magnitude of each argument. The \( i \)'th cost term is multiplied by \( \gamma_i^{-1} \), so that sources with higher values of \( \gamma_i \) are easier to distort. Incorporating the agent’s preference for perceiving mutually agreeing arguments as well as her costs of subjective scaling, her utility given argument profile \( a = (a_1, \ldots, a_N) \) and scaling profile \( s = (s_1, \ldots, s_N) \) is

\[
U(s|a) = \sum_{i,j=1}^{N} \langle s_ia_i, s_ja_j \rangle - \sum_{i=1}^{N} c_i(a) \cdot \gamma_i^{-1} \cdot (1 - s_i)^2.
\]

As stated formally in Proposition 1 below, specifying \( c_i(a) = |a_i| \cdot \sum_{j=1}^{N} |a_j| \) guarantees the agent’s utility function is both strictly concave and scale-invariant, facilitating analysis of the agent’s optimal distortion. See Appendix B for further discussion of this functional form.

**Proposition 1.** If malleability \( \gamma_i \in (0, 1) \) for all sources \( i = 1, \ldots, N \), then \( U(\cdot|a) \) is strictly concave for all argument profiles \( a \). If \( \gamma_{i^*} > 1 \) for some source \( i^* \), then there exists an argument profile \( a^* \) such that \( U(\cdot|a^*) \) is nowhere concave and admits no optimum. *Proof in the appendix.*

Note that \( U \) is homogenous of degree 2 in argument profiles: \( U(s|ka) = k^2 \cdot U(s|a) \) for \( k > 0 \). Hence, the degree of scaling of arguments only depends on the relative magnitudes and agreements between arguments.

### 2.2. Optimal Distortion

The agent’s optimal choice of scaling factors follows from her utility function’s first order conditions. In order to express the optimal factors in a concise and intuitive manner, I first introduce several summary statistics of the set of arguments. Let \( V(a) = \sum_{i=1}^{N} |a_i| \) denote the volume of the set of arguments, and let

\[
G(\gamma; a) = \sum_{i=1}^{N} \left( \frac{|a_i|}{V(a)} \right) \gamma_i
\]

denote the magnitude-weighted average malleability factor among all the sources. Let

\[
\rho(a_i, \pi_{\text{obj}}) = \frac{\langle a_i, \pi_{\text{obj}} \rangle}{|a_i| \cdot |\pi_{\text{obj}}|} = \begin{cases} 
1 & \text{if } \langle a_i, \pi_{\text{obj}} \rangle > 0 \\
-1 & \text{if } \langle a_i, \pi_{\text{obj}} \rangle < 0 \\
0 & \text{if } \langle a_i, \pi_{\text{obj}} \rangle = 0
\end{cases}
\]
indicate whether argument $a_i$ and the objective posterior $\pi_{\text{obj}} = \sum_{i=1}^{N} a_i$ have the same or opposite directions.

**Fact 1.** The agent’s subjective posterior is

$$\pi_{\text{post}} = \frac{1}{1 - G(\gamma; a)} \pi_{\text{obj}}$$

and her optimal scaling of source $i$ is

$$s_i^* = 1 + \rho(a_i, \pi_{\text{obj}}) \cdot \frac{\gamma_i}{1 - G(\gamma; a)} \cdot \frac{||\pi_{\text{obj}}||}{V(a)}.$$

These formulas exhibit several properties discussed below.

**Partisan Treatment of Sources.** The most immediate feature of the agent’s optimal behavior is that she scales up each argument that agrees with her subjective posterior and scales down each argument that disagrees:

$$s_i^* \geq 1 \text{ if and only if } \langle a_i, \pi_{\text{post}} \rangle \geq 0.$$

In this way, the agent effectively separates the sources into friends, who with their concordant arguments are given greater prominence, and enemies that are shunned for their dissonant statements.

**The Agent’s Chosen Side Matches the Totality of Evidence.** When all sources are in mutual agreement, the agent necessarily identifies positively with all of them, but when the sources disagree, with some sending right-leaning and others sending left-leaning arguments, the agent identifies only with a single side. As identified in Fact 1, the subjective posterior is a positively scaled version of the objective, and thus the chosen side agrees with the objective posterior. Intuitively, this happens because the agent does not identify with either side ex-ante, but instead seeks the most cost-effective method of reducing dissonance. For her, scaling up any argument $a_i$ consequently increases the marginal benefit to scaling up any argument that agrees with $a_i$. Hence the agent identifies with the side whose neutrally perceived arguments have the greater agreement among each other, and this side corresponds to the neutral posterior $\pi_{\text{obj}}$.

**Extent of Scaling.** How severely the agent scales any given source’s argument depends on three factors. The first is the source’s individual malleability factor $\gamma_i$. As $\gamma_i \to 0$, the source becomes rigid, and all misperception disappears. The second component is the average malleability value of $G(\gamma; a)$. If $G(\gamma; a) \to 0$, then all sources are treated without distortion; if $G(\gamma; a) \to 1$, then
all sources are scaled arbitrarily extremely. Finally, the third component is the ratio $|\pi_{post}^{obj}|/V(a)$, which describes how united the sources are in their objective arguments. If all sources are left-leaning or all sources are right-leaning, the ratio is 1, but as the extent of disagreement between the sources increases (keeping fixed their neutral aggregate $\pi_{post}^{obj}$), the ratio approaches 0. In the extreme, the agent reacts to sources that disagree with her subjective posterior by scaling them negatively: $s_i^* < 0$ if $\langle a_i, \pi_{post} \rangle < 0$ and

$$\gamma_i > \frac{1 - G(\gamma; a)}{|\pi_{post}^{obj}|/V(a)}.$$ 

There are two ways to interpret this backlash. In a literal reading, the agent actually misperceives the direction of the source’s argument, but it is equally valid to interpret the agent framing the source as a foe with whom disagreement is preferable.

**OVER-LEARNING.** A final observation about the formulae in Fact 1 is that the agent always over-learns, in that her posterior $\pi_{post}$ is more extreme than its objective counterpart by a factor of $1/(1 - G(\gamma; a))$. As the average malleability $G(\gamma; a)$ approaches its maximum value of 1, the agent’s subjective posterior becomes arbitrarily extreme ($\pi_{post} \rightarrow \infty$ or $\pi_{post} \rightarrow -\infty$), meaning the agent becomes arbitrarily certain of either $\omega = L$ or $\omega = R$.

### 2.3. Examples

**SINGLE ARGUMENT.** The simplest possible case of the model involves a single argument $a_1$ with malleability $\gamma_1$. To compute the agent’s scaling and posterior, first consider the intermediate quantities discussed in the previous subsection, the calculation of which are trivial in this particular case. The volume of the arguments $V(a)$ equals the magnitude of $a_1$, the average malleability factor $G(\gamma; a)$ is equal to $\gamma_1$, the objective posterior is simply $a_1$, and the directional indicator $\rho(a_1, \pi_{post}^{obj})$ between $a_1$ and the objective posterior is necessarily equal to 1. Following Fact 1, the agent scales the argument by a factor of

$$s_1^* = 1 + \rho(a_1, \pi_{post}^{obj}) \cdot \frac{\gamma_1}{1 - G(\gamma; a)} = 1 + 1 \cdot \frac{\gamma_1}{1 - \gamma_1} \cdot \frac{|a_1|}{|a_1|} = \frac{1}{1 - \gamma_1}.$$ 

Her subjective posterior is $\pi_{post} = a_1/(1 - \gamma_1)$. Note how this simple example highlights the agent’s propensity to over-learn. With a single argument, there is no room for disagreement among sources; the lone argument must agree with itself. The agent likes this and wants more of it, so she is guaranteed to scale up the argument and arrive at an exaggerated version as her posterior.
(An alternative perspective: were the single argument decomposed into two or more components in mutual agreement, then the agent would enjoy the agreement between the components, and ultimately all of them would be scaled just as the original argument was.)

**Figure 1.** Dissonance Reduction of Two Arguments. *The agent receives a weak right-leaning argument and a strong left-leaning argument, scales down the right, scales down the left, and arrives at a subjective posterior further to the left of the Bayesian posterior.*

**TWO ARGUMENTS.** The next simplest case is that of two arguments. Suppose first that arguments $a_1$ and $a_2$ share the same malleability factor $\gamma = \gamma_1 = \gamma_2$. If $a_1$ and $a_2$ agree with each other, then the agent treats both arguments just as though she had received each in isolation, scaling them by $s^* = 1/(1 - \gamma)$. If $a_1$ and $a_2$ disagree, then the agent sides with the stronger of the two, exaggerating that one while downplaying the other. See Figure 1 for a graphical depiction. Finally, suppose that $a_1$ and $a_2$ share the same direction, but $a_1$ is more malleable: $\gamma_1 > \gamma_2$. In this case, the average malleability factor $G(\gamma; a)$ lies between $\gamma_1$ and $\gamma_2$. The agent’s subjective posterior is scaled up by $1/(1 - G)$, but the scaling factors for the individual arguments reflect the differential malleabilities:

$$s_1^* = 1 + \frac{\gamma_1}{1 - G(\gamma; a)} > 1 + \frac{\gamma_2}{1 - G(\gamma; a)} = s_2^*.$$  

**MULTIPLE ARGUMENTS WITH FIXED MAGNITUDE, MALLEABILITY.** As a last illustrative example, suppose that the agent is faced with $k_R$ right-leaning arguments and $k_L$ left-leaning arguments, where all arguments share the same magnitude $|a|$ and the same malleability $\gamma$. The agent sides with whichever side has the majority of the arguments. Without loss of generality suppose it is the right; then the agent scales all of the arguments by

$$s_i^* = 1 \pm \frac{\gamma}{1 - \gamma} \cdot \frac{k_R - k_L}{k_R + k_L}.$$
This illustrates how, although the agent appears partisan, in fact she simply wants to back the ‘winning’ side, as she perceives greater agreement by amplifying the majority. The expression for $s_i^*$ also highlights a wedge between the agent’s subjective posterior and the scaling of any single argument. Her subjective posterior is $(k_R - k_L)|a|/(1 - \gamma)$, which depends only on the net number of right-leaning arguments. Her scaling of any single argument depends both on that value as well as the total number of arguments. Holding fixed the net number of right-leaning arguments, an increase in the total number of arguments reduces the extremity of scaling of any single argument, but that is evenly counterbalanced by the increasing number of arguments so that the agent’s subjective posterior remains the same.

2.4. Comparative Statics

The fundamentals of the model can vary in several ways. First, consider a marginal strengthening of a particular argument $a_i \neq 0$. The effect of an increase in the magnitude $|a_i|$ on the average malleability factor $G(\gamma; a)$ is

$$\frac{\partial G(\gamma; a)}{\partial |a_i|} = \frac{\gamma_i - G(\gamma; a)}{V(a)}.$$ 

That is, $G(\gamma; a)$ increases in $|a_i|$ whenever the malleability of source $i$ exceeds $G(\gamma; a)$, and decreases whenever $i$’s malleability is lower than $G(\gamma; a)$. Using this formula, we can express the comparative statics of $|a_i|$ on the subjective posterior and the individual scaling factors.

**Fact 2.** The marginal effect of $|a_i|$ on the agent’s subjective posterior is:

$$\frac{\partial \pi_{post}}{\partial |a_i|} = \frac{a_i}{1 - G(\gamma; a)} + \left(\frac{1}{1 - G(\gamma; a)}\right)^2 \left(\frac{\partial G(\gamma; a)}{\partial |a_i|}\right) \pi_{post}^{obj}.$$ 

and the marginal effect of $|a_i|$ on the scaling of source $j$ is:

$$\frac{\partial s_j^*}{\partial |a_i|} = \frac{\rho(a_j, \pi_{post}) \cdot \gamma_j}{1 - G(\gamma; a)} \cdot \left(\frac{\rho(a_i, \pi_{post}) V(a) - |\pi_{post}^{obj}|}{V(a)^2} + \frac{1}{1 - G(\gamma; a)} \cdot \frac{\partial G(\gamma; a)}{\partial |a_i|}\right).$$

As Fact 2 shows, the marginal effects of $|a_i|$ on the subjective posterior and the scaling factors decompose into two terms: a direct effect and a relative malleability effect. This is because strengthening argument $a_i$ both changes the objective profile of information as well as alters the magnitude-weighted average of malleability parameters $G(\gamma; a)$. For the subjective posterior, the
direct effect pushes in the direction of \( a_i \): strengthening \( a_i \) makes the objective posterior tend in \( a_i \)’s direction, and the agent reaches a positively scaled version of the objective posterior. The relative malleability effect either extends the existing subjective posterior, when \( \gamma_i > G(\gamma; a) \), or contracts it, when \( \gamma_i < G(\gamma; a) \).

With respect to the scaling of source \( j \)’s argument, the relative malleability effect likewise depends on whether the strengthened source \( i \)’s malleability parameter \( \gamma_i \) is above or below the existing average \( G(\gamma; a) \). In the former case, strengthening \( a_i \) increases average malleability, makes the subjective posterior more extreme, and hence leads to higher scaling of source \( j \) if and only if \( j \)’s argument has positive agreement with the subjective posterior. When \( \gamma_i \) is lower than \( G(\gamma; a) \), strengthening \( a_i \) dampens the subjective posterior, leading to higher scaling of \( s_j \) only when \( j \)’s argument disagrees with the subjective posterior. The direct effect of \( |a_i| \) on \( s_j \) depends whether the arguments \( a_i \) and \( a_j \) agree or disagree with each other. If \( a_i \) and \( a_j \) agree, then the effect is necessarily positive: either the subjective posterior agrees with both arguments, so an increase in \( |a_i| \) extends \( \pi_{\text{post}} \) and accordingly leads to an even more amplified \( s_j \), or the subjective posterior disagrees with both arguments, in which case an increase in \( |a_i| \) dampens \( \pi_{\text{post}} \) and leads to a less mitigated \( s_j \). When \( a_i \) and \( a_j \) disagree, the statements above flip in direction, leading to a negative effect.

The Effect of Noise. Another perspective from which to examine the marginal effects of the model is the following thought experiment. The agent begins with an existing profile of arguments \( a = (a_1, \ldots, a_N) \). Subsequently she is presented with two additional arguments \( a_{N+1} \) and \( a_{N+2} \) that push in opposite directions with equal magnitude: \( a_{N+1} = -a_{N+2} \). What happens as the magnitude \( |a_{N+1}| \) of additional ‘noise’ in the agent’s profile is increased? If the new arguments have the same malleability as the original profile, \( \gamma_{N+1} = \gamma_{N+2} = G(\gamma_1, \ldots, \gamma_N|a_1, \ldots, a_N) \), then there is no effect on the agent’s posterior, \( \pi_{\text{post}}(a_1, \ldots, a_{N+2}) = \pi_{\text{post}}(a_1, \ldots, a_N) \), while all of the original sources are scaled less extremely: if \( s_i \) is the original scaling factor and \( s_i' \) accounts for \( a_{N+1} \), then \( |s'_i - 1| < |s_i - 1| \). Thus, the effect of noise is ultimately neutral. While additional conflicting sources increase the total volume of arguments faced by the agent, increase the marginal cost to scaling any existing source, and thus dampen the distortion of the existing sources, the new sources provide an additional medium through which the agent can increase perceived agreement. These two countervailing pressures precisely cancel each other. If the malleability factors of the
new arguments are greater (less) than $G(\gamma; a)$, then the agents’ posterior is accordingly more (less) extreme.

**Sufficient Statistics.** An additional way to view the mechanics of the bias at work and to see the marginal effect of any single argument more clearly is through the key data that influences the agent’s decisions. For any profile of arguments $(a_1, \ldots, a_N)$, the agent’s scaling of $a_1$ depends on how $a_1$ relates to $a_2, \ldots, a_N$ through three sufficient statistics: the *information content* of sources 2 through $N$: $\sum_{i=2}^{N} a_i$, the *volume* of those sources: $\sum_{i=2}^{N} |a_i|$, and their (volume-weighted) *average malleability*: $\sum_{i=2}^{N} (|a_i|/\sum_{j=2}^{N} |a_j|)\gamma_i$. That is, if $a_2, \ldots, a_N$ are replaced by an alternative set of arguments $\tilde{a}_2, \ldots, \tilde{a}_M$ with the same information content, volume, and average malleability, then the agent’s scaling of $a_1$ and her subjective posterior will be unchanged.

### 2.5. Dissonance Reduction as Self Persuasion

Yet another perspective from which to view the agent’s bias is self-persuasion. The agent’s behavior is equivalent to two forms of self-persuasion, one in which her partisan identity is immediately determined by the set of arguments and one in which she reasons gradually to reach an ideal arrangement of scaling.

**Hidden Motive.** Consider an agent with the following *alternative* utility function. Instead of maximizing pairwise agreement between arguments, he seeks agreement between the arguments and an endogenously specified bias argument $a_{\text{bias}}$:

$$
\tilde{U}(s|a) = \sum_{i=1}^{N} \langle s_i a_i, a_{\text{bias}} \rangle - \sum_{i=1}^{N} \gamma_i^{-1} \cdot c_i(a) \cdot (1 - s_i)^2.
$$

It follows that his optimal scaling is

$$
s^*_i = 1 + \frac{1}{2} \frac{\gamma_i \cdot c_i(a)^{-1} \langle a_i, a_{\text{bias}} \rangle}{\gamma_i}.
$$

As with the dissonance reducer, this agent scales arguments up or down according to their degree of agreement with a reference argument. Moreover, in the case that his exogenous bias is $a_{\text{bias}} = 2\pi_{\text{obj}}$, twice the objective posterior, then he will scale all of the sources with the same factors as the dissonance-reducing agent uses. In this way, the dissonance-reducing agent always acts as though she she happened to have a particular exogenous bias, when in fact that ‘bias’ depends on the Bayesian posterior.
**Graded Adjustment.** To understand how the dissonance-reducing agent might reach her profile of optimal scaling values, consider the following dynamic process. She begins with a provisional posterior $π_0$, then adjusts

$$π_k = π^\text{obj}_\text{post} + \sum_{i=1}^{N} γ_i · c_i(a)^{-1} · ⟨a_i, π_{k-1}⟩$$

for $k = 1, 2, \ldots$, iteratively assessing whether her provisional posterior agrees or disagrees with each of the sources. To the extent that it agrees, she adjusts in that direction; to the extent it disagrees, she counters in the opposite direction. This sequential revision procedure always converges to $π^\text{post}_\text{post}$.

**Fact 3.** $\lim_{k→∞} π_k = π^\text{post}_\text{post}$ for all $π_0$.

Understanding the agent’s optimal scaling as the limit of gradual adjustment also provides intuition as to why it coincides with an inflated version of the Bayesian conclusion. Starting from her initial provisional posterior, the dissonance-reducing agent scales up complementary arguments and scales down opposing ones. Nonetheless, her subsequent provisional belief becomes marginally more aligned with the Bayesian posterior, and repetition of the process causes the agent’s posterior to converge on the inflated version of the Bayesian conclusion.

### 2.6. Non-Neutral Prior

Until now, it has been assumed that the agent starts with the neutral belief $π^\text{prior}_\text{prior} = 0$. Now consider how the agent deals with a non-neutral prior. Observe that any starting belief $π^\text{prior}_\text{prior}$ reflects the receipt of a unique argument which would shift the agent’s belief from 0 to $π^\text{prior}_\text{prior}$, and hence can be modeled as though it were yet another argument. Accordingly,

$$π^\text{post}_\text{post} = π^\text{prior}_\text{prior} + \left( \frac{G(γ; a)}{1 - G(γ; a)} π^\text{prior}_\text{prior} + \frac{1}{1 - G(γ; a)} \sum_{i=1}^{N} a_i \right).$$

The prior does not affect how new arguments are scaled other than through the average malleability factor. Instead, there is a positive-reinforcement effect, strictly weaker than the scaling on new arguments, which boosts the information content of the prior. This means that the agent with a non-neutral prior and no new information would still continue to strengthen her beliefs.\(^6\) The dynamic version in the subsequent section elaborates on this.

\(^6\)Thaler (2019) finds experimentally that people with non-neutral political opinions indeed strengthen those opinions when presented with objectively meaningless signals.
3. DYNAMIC ADJUSTMENT

Consider the following sequential version of the model. The agent begins with an initial belief $\pi_0 = 0$. Then, in each period $t = 1, 2, \ldots$, she receives a new argument $a_t$. At the start of period $t$, her current belief is her time $t - 1$ posterior $\pi_{t-1}$. The agent treats $a_t$ and $\pi_{t-1}$ as her only two pieces of information and scales them as in the static model to produce her time $t$ posterior $\pi_t$.

To concentrate on the dynamic elements of this environment, assume all of the arguments share a common malleability factor $\gamma$. Furthermore, assume the arguments are \textit{i.i.d.} conditional on the state and have a finite support.

**Main Effects.** In each period $t$, the agent remembers the \textit{previously-scaled} versions of former arguments, and hence she acts as though the objective balance of evidence were $\pi_{t-1} + a_t$. Per Fact 1, the agent scales up $a_t$ in exactly two conditions: first if the new argument agrees with her current belief, $\langle a_t, \pi_{t-1} \rangle > 0$, and second if the new argument disagrees with it but also has greater magnitude than her current belief: $\langle a_t, \pi_{t-1} \rangle < 0$ and $|a_t| > |\pi_{t-1}|$. In terms of her treatment of the new argument $a_t$, the agent therefore exhibits behavior similar to confirmation bias. Small marginal arguments are exaggerated when they conform to the agent’s existing belief and mitigated otherwise.

As discussed in greater detail below, what distinguishes the dissonance-reducing agent from one with confirmation bias is that, in addition to distorting the marginal argument $a_t$ in relation to the current belief $\pi_{t-1}$, the agent simultaneously \textit{re-evaluates} the current belief in relation to the marginal argument. This produces a reinforcement effect. Belief $\pi_{t-1}$ itself is scaled up provided that $a_t$ is not sufficiently strong in the opposite direction. As with the static agent who has a non-neutral prior, the two effects combine so that the agent’s belief in period $t$ as a function of her current belief and the marginal argument is:

$$\pi_t = \frac{1}{1 - \gamma} (\pi_{t-1} + a_t) = \pi_{t-1} + \left[ \frac{1}{1 - \gamma} a_t + \frac{\gamma}{1 - \gamma} \pi_{t-1} \right].$$

This formula shows that the new belief $\pi_t$ is equal to the current belief $\pi_{t-1}$ plus the marginal argument $a_t$, exaggerated by factor $1/(1 - \gamma)$, plus the current belief itself scaled by $\gamma/(1 - \gamma)$. Note

\footnote{These assumptions impose a small loss of generality. Allowing for variance in the arguments’ malleability factors does not qualitatively change how the agent aggregates a new argument with her current belief, but it significantly complicates the calculation of long-run behavior. Similarly, allowing for a non-finite support does not affect any single update, but it opens the door to a failure of convergence.}
that although the agent treats the marginal argument differentially depending on its relationship
with the current belief, the marginal argument’s net effect on the belief sequence is independent of
all past arguments. Instead, the re-evaluation of past beliefs leads to a positive reinforcement of
\( \pi_{t-1} \). Even when \( a_t = 0 \), this reinforcement means that \( \pi_t \) is a more extreme version than \( \pi_{t-1} \).

CONVERSIONS. At a local level, the agent’s treatment of marginal arguments depends on her current
belief, but her belief sequence is punctuated by occasional conversions in which her belief changes
direction, that is \( \langle \pi_t, \pi_{t-1} \rangle < 0 \). As illustrated by the formula above, a conversion occurs only
when the marginal argument not only disagrees with the agent’s current belief but is also stronger:
\(|a_t| > |\pi_{t-1}|\). When a relatively small marginal argument disagrees with the current belief, the
net effect on belief is to strengthen the current belief. Assuming without loss of generality that
\( \pi_{t-1} > 0 \), it follows \( \pi_t > \pi_{t-1} \) whenever \( a_t > 0 \) or \( |a_t| < \gamma |\pi_{t-1}| \). For conflicting marginal arguments
of intermediate strength, \( |a_t| \in (\gamma |\pi_{t-1}|, |\pi_{t-1}|) \), the current belief is weakened but not reversed:
\( \pi_t < \pi_{t-1} \) but \( \pi_t > 0 \).

ASYMPTOTIC BEHAVIOR. In terms of long-run learning, the agent’s reinforcement of prior beliefs is
a path-dependent force that causes the agent to err when early arguments initially guide her in the
wrong direction, while the possibility of conversions mitigates the prospect of ultimate mis-learning.
To understand how these forces balance each other, we can re-express the agent’s belief sequence
in a non-recursive form:

\[
\pi_t = \sum_{\tau=1}^{t} \left( \frac{1}{1 - \gamma} \right)^{t+1-\tau} a_{\tau} = \frac{1}{\gamma (1 - \gamma)^t} \left( \sum_{\tau=1}^{t} (1 - \gamma)^{t-1} a_{\tau} \right).
\]

This expression decomposes \( \pi_t \) as the product of two components. Factor \( m_t \) is a growth coefficient
that depends only on calendar time and not the information received by the agent. In contrast, \( \xi_t \)
is a geometrically-weighted average of the arguments received through period \( t \). As \( t \) grows, the
sum of the weights converges to 1, and \( \xi_t \) converges to the long-run average of

\[
\xi_\infty = \gamma \sum_{\tau=1}^{\infty} (1 - \gamma)^{t-1} a_{\tau}.
\]

Because the growth coefficient \( m_t \) increases without bound while the directional component con-
verges, it follows that either \( \pi_t \rightarrow \infty \) or \( \pi_t \rightarrow -\infty \) (so long as \( \xi_\infty \neq 0 \)). Thus, the agent becomes
fully convinced of one of the two state values. However, as established in the following proposition, there is always positive probability that the agent learns incorrectly.

**Proposition 2.** The agent always becomes certain about the state: \( P[\pi_t \rightarrow \infty | \omega] + P[\pi_t \rightarrow -\infty | \omega] = 1 \), and she settles on each of the two values with positive probability: \( P[\pi_t \rightarrow \infty | \omega] > 0 \) and \( P[\pi_t \rightarrow -\infty | \omega] > 0 \). *Proof in the appendix.*

The agent’s potentially incorrect learning stems from the reinforcement effect causing arguments received earlier in time to be weighted more heavily than those received later. As seen in the expression of \( \xi_\infty \), fraction \( \gamma \) of the total weight in the long-run average is placed on the very first argument \( a_1 \). Hence, the agent is susceptible to being misled by the initial arguments she receives.

Several additional points can be noted about the relative likelihood of the agent learning correctly versus incorrectly.

**Proposition 3.** Suppose \( \omega = R \). For all \( \varepsilon > 0 \), there exists a value of \( \gamma \) and a distribution of \( a_t \) such that \( P[\pi_t \rightarrow \infty] < \varepsilon \), and there exists a value of \( \gamma \) and a distribution of \( a_t \) such that \( P[\pi_t \rightarrow \infty] > 1 - \varepsilon \). *Proof in the appendix.*

To construct examples in which the agent either learns correctly or incorrectly with arbitrarily high probability, it suffices to restrict attention to distributions in which \( a_t \) only takes on one of two values, a right-leaning argument \( a_R > 0 \) or a left-leaning argument \( a_L < 0 \). Assuming without loss of generality that the true state is \( \omega = R \), and denoting the probability of the agent receiving either of the two arguments by \( q_R \) and \( q_L \), it must be that the average argument is right-leaning: \( q_R a_R + q_L a_L > 0 \). Nonetheless, if the agent is more likely to receive the left-leaning argument, and the common malleability factor \( \gamma \) is high, then the agent’s learning is dominated by successive reinforcement of early arguments that are likely to be \( a_L \). Hence, it is possible that the agent is more likely to incorrectly learn \( \omega = L \) than to correctly learn \( \omega = R \); as the probability of receiving \( a_L \) and the value of \( \gamma \) both approach 1, incorrect learning becomes arbitrarily certain.

**Comparison with the Static Model.** Let \( \pi_t^{\text{static}} \) denote the belief that the agent would reach were she to simultaneously receive the arguments \( a_1, \ldots, a_t \). By Fact 1,

\[
\pi_t^{\text{static}} = \frac{1}{1 - \gamma} \sum_{\tau=1}^{t} a_\tau = \pi_{t-1}^{\text{static}} + \frac{1}{1 - \gamma} a_t.
\]
The sequence of the agent’s beliefs as determined by iterated recalculation of the static model is governed by a difference equation that matches the dynamic version but without any reinforcement effect. Hence, the static beliefs feature less dramatic conversions as well as no probability of ultimate mis-learning. As illustrated in Figure 2, the static agent’s belief always changes in the direction of the marginal argument, even when that argument disagrees with the provisional conclusion. Viewed as functions of the marginal argument, the static and dynamic versions of the dissonance-reducing agent increase in parallel, with the dynamic belief consistently more in the direction of the agent’s current belief.

![Figure 2. Updating with Dissonance Reduction vs Alternatives. At time $t$, an agent has belief $\pi_t$ and receives argument $a_t$; her time $t + 1$ belief is shown for the case of objective learning ($\pi_{t+1}^{\text{obj}}$), dynamic dissonance reduction ($\pi_{t+1}^{\text{DR}}$), static dissonance reduction ($\pi_{t+1}^{\text{static}}$), and confirmation bias ($\pi_{t+1}^{\text{CB}}$).](image)

**Comparison with Confirmation Bias.** Adapting the approach in Rabin and Schrag (1999),\(^8\) say that a confirmation-biased agent’s belief sequence $\{\pi_t^{\text{CB}}\}_{t=1,2,...}$ follows $\pi_t^{\text{CB}} = f(a_t|\pi_{t-1}^{\text{CB}})$ where function $f$ is increasing in the first argument, $f(0|\pi) = \pi$ (no information produces no change in

---

\(^8\)In Rabin and Schrag’s model, an agent may with some probability misperceive one of two possible signals for its opposite value when the truth would push the agent’s beliefs against the agent’s current direction. Rather than introduce an additional random element here, I compare dissonance reduction to a deterministic version in which confirmation bias means the agent always departs from the Bayesian update according to her current direction.
beliefs), and

\[
\begin{align*}
    f(a_t|\pi_{t-1}^{CB}) & \geq \pi_{t-1}^{CB} + a_t \quad \text{if } \pi_{t-1}^{CB} > 0 \\
    f(a_t|\pi_{t-1}^{CB}) & \leq \pi_{t-1}^{CB} + a_t \quad \text{if } \pi_{t-1}^{CB} < 0.
\end{align*}
\]

That is, a confirmation-biased agent always exaggerates marginal arguments that support the current belief and mitigates those which conflict with the current belief.

As illustrated in Figure 2, the key difference between dissonance reduction and confirmation bias lies in the nature of conversions. Although it is possible for a confirmation-biased agent to transition from a left-leaning belief to a right-leaning one or visa-versa, changes are comparatively gradual. Assuming without loss of generality that \( \pi_t > 0 \), the confirmation-biased agent always has a time \( t + 1 \) belief weakly more right-leaning than the Bayesian baseline, even for very strong left-leaning arguments \( a_{t+1} \). By contrast, the dissonance-reducing agent arrives at a right-biased belief for weak left-leaning arguments \( a_{t+1} \) but a left-biased belief for strong left-leaning arguments. In this sense the dissonance-reducing agent resists small conflicting arguments more strongly but displays more dramatic conversions.

**Fact 4.** Suppose \( \pi > 0 \). For sufficiently small conflicting arguments, \( a_{t+1} \in (a_{t+1}, 0) \), the dissonance-reducing agent displays a stronger form of confirmation bias:

\[
\pi_{t+1}(a_{t+1}) > \pi_{t+1}^{CB}(a_{t+1}) > \pi_{t+1}^{obj}(a_{t+1}),
\]

but for sufficiently large conflicting arguments, \( a_{t+1} < a_{t+1} \), the agent’s belief diverges from the Bayesian benchmark in the opposite direction:

\[
\pi_{t+1}^{CB}(a_{t+1}) > \pi_{t+1}^{obj}(a_{t+1}) > \pi_{t+1}(a_{t+1}).
\]

4. **Multiple Issues**

As an extension of the model in Section 2, suppose an agent learns about an unknown state \( \omega \) by aggregating information from sources \( i = 1, \ldots, N \), where the state’s true value is a left or right assignment for \( D \geq 1 \) different issues: \( \omega \in \{L, R\}^D \). I assume the issues are mutually independent. Hence, the agent’s prior belief can be expressed as a \( D \)-dimensional vector of log odds-ratios \( \pi_{prior} \in \mathbb{R}^D \). Each source’s argument can similarly be expressed as a \( D \)-dimensional
vector of log likelihood-ratios, \( a_i \in \mathbb{R}^D \), and, as in Section 2, Bayes’ rule dictates

\[
\pi_{\text{post}}^{\text{obj}} = \pi_{\text{prior}} + a_1 + \cdots + a_N.
\]

The form of the agent’s utility function remains unchanged from the baseline environment,

\[
U(s|a) = \sum_{i,j=1}^{N} \langle s_i a_i, s_j a_j \rangle - \sum_{i=1}^{N} c_i(a) \cdot \gamma_i^{-1} \cdot (1 - s_i)^2,
\]

but in this expression, the agreement function \( \langle \cdot, \cdot \rangle \) is a generalized version of its one-dimensional counterpart. Specifically, the agreement between two arguments is a weighted sum of their one-dimensional agreements on the \( D \) different issues:

\[
\langle a_i, a_j \rangle = \sum_{d=1}^{D} w_d \cdot a_{i,d} \cdot a_{j,d}.
\]

The coefficients \( w_d > 0 \) are issue-specific weighting factors which identify some issues as more or less important for determining agreement and disagreement. See Appendix B for further discussion of the utility function specification, and note that concavity of \( U \) (Proposition 1) holds in the multidimensional case as well. Finally, as some consequences of dissonance reduction in the general model occur only when the level of bias (as quantified by the malleability factors) is sufficiently extreme, it is useful to introduce a parametrization for raising the level of bias. Define a \( k \)-strengthening of the agent’s dissonance-reduction by replacing \( \gamma_i \) with \( \tilde{\gamma}_i \equiv k \cdot \gamma_i \) for all sources \( i = 1, \ldots, N \) and \( k \geq 1 \).

4.1. Calculation of the Subjective Posterior

As in the unidimensional environment, the agent’s optimal scaling behavior follows from the first order conditions:

\[
s^*_i = 1 + \gamma_i \cdot c_i(a)^{-1} \cdot \langle a_i, \pi_{\text{post}} \rangle.
\]

The sources which agree with the subjective posterior are scaled up; those that disagree are scaled down. What differs in the multidimensional environment is the determination of the agent’s subjective posterior. In choosing each scaling factor, the agent is confronted with tradeoffs between issues. Scaling up a source may increase the agreement perceived on one dimension while simultaneously decreasing the agreement perceived on another.
To derive the subjective posterior, scale $a_i$ by each side of equation (1) above and sum across $i = 1, \ldots, N$ to produce

$$\pi_{\text{post}} = \pi_{\text{post}}^{\text{obj}} + \sum_{i=1}^{N} \gamma_i \cdot \frac{1}{|a_i|} \sum_{j=1}^{N} \langle a_i, \pi_{\text{post}} \rangle a_i.$$  \hspace{1cm} (2)

In the unidimensional case, this reduces to $\pi_{\text{post}} = \pi_{\text{post}}^{\text{obj}} + G(\gamma; a)\pi_{\text{post}}$, from which it follows $\pi_{\text{post}} = (1 - G(\gamma; a))^{-1}\pi_{\text{post}}^{\text{obj}}$. In the multidimensional case, a related formula follows from equation (2):

$$\pi_{\text{post}} = \pi_{\text{post}}^{\text{obj}} + \Gamma(\gamma|a)\pi_{\text{post}}, \quad \text{where} \quad \Gamma(\gamma|a) = \sum_{i=1}^{N} \left( \frac{|a_i|}{V(a)} \right) \gamma_i \cdot \frac{a_i a_i'}{|a_i|^2}.$$  

The expression $a_i a_i' / |a_i|^2$ is the $(N \times N)$ matrix which orthogonally projects vectors onto the span of $a_i$, so $\Gamma(\gamma|a)$ is a weighted sum of projection matrices. It functions as a direction-specific average malleability factor. For any unit-magnitude vector $v \in \mathbb{R}^D$, which we can interpret as a hypothetical reference argument, define the weighted alignment between $v$ and the profile of sources’ arguments by the function

$$G_v(\gamma|a) = |\Gamma(\gamma|a)v| = \sum_{i=1}^{N} \left( \frac{|a_i|}{V(a)} \right) \gamma_i \cdot \rho(v, a_i)^2,$$

where, as before, $\rho(v, a_i) \in [-1, 1]$ is the magnitude-adjusted level of agreement between $v$ and $a_i$. The value of $\rho(v, a_i)^2$ measures the extent to which $v$ and $a_i$ express information on the same topics. If $v$ and $a_i$ are both non-zero the same single issue (as is the case in the unidimensional environment), then $\rho(v, a_i)^2 = 1$ regardless of whether $v$ and $a_i$ agree or disagree. Likewise, if $v$ and $a_i$ are non-zero on disjoint subsets of issues, then $\rho(v, a_i)^2 = 0$. We can therefore interpret $G_v(\gamma|a)$ as a direction-specific average of the malleability parameters in which the weight given to each source’s parameter depends not only on the magnitude of the source’s argument, but also on how aligned the argument is with vector $v$.

The agent’s subjective posterior is a distorted version of the objective posterior which is stretched in each direction $v$ by a factor approximately equal to $1/(1 - G_v(\gamma|a))$. That approximation is precise for a specific set of directions $v^1, \ldots, v^D$ that are mutually orthogonal and constitute an alternative coordinate system for $\mathbb{R}^D$. Formally, define the principal axes of the set of arguments as follows.
Definition. An orthonormal basis \( \{v^d\}_{d=1}^D \) constitutes a principal axis system if

\[
v^d \in \underset{|v|=1; v \perp v^1, \ldots, v^{d-1}}{\text{argmax}} G_v(\gamma; a)
\]

for all \( d = 1, \ldots, D \).

As shown by Lemma 2 in Appendix B, the principal axis system is generically unique; I assume all argument profiles studied admit a unique system. Re-expressing the original set of arguments \( a_1, \ldots, a_n \) in the principal axis coordinates, i.e. denoting \( a^i = (a^i_1, \ldots, a^i_d) \) where \( a_i = a_i^1 v^1 + \cdots + a_i^D v^D \), allows for a clean generalization of the single-dimensional formula. To simplify notation, let \( G(\gamma; a)^d = G_{v^d}(\gamma|a) \).

Fact 5. The agent’s subjective posterior is

\[
\pi^d_{\text{post}} = \frac{1}{1 - G(\gamma; a)^d \pi_{\text{post}}} \pi^D_{\text{obj}}
\]

and her optimal scaling of source \( i \) is

\[
s^*_i = 1 + \gamma_i \cdot c_i(a)^{-1} \cdot \langle a_i, \pi_{\text{post}} \rangle.
\]

4.2. Issue Skewing

What most distinguishes the model with multiple issues from the core model is a phenomenon termed issue skewing, by which the agent chooses scaling factors in the interest of avoiding disagreement on one issue and consequently alters her posterior on other issues as a byproduct.

Examples. Figures 3 - 5 illustrate several examples involving two issues. First, first suppose the agent receives two arguments of equal magnitude and malleability, each of which concerns a different issue. As illustrated in Figure 3, the agent scales up both of the arguments equally, so that the subjective posterior is a positively scaled copy of the objective. However, if one of the arguments is stronger, then the principal axis is rotated towards it, and the subjective posterior aligns more with the strong argument than the objective posterior does. This case is depicted in Figure 4. If, alternatively, there are three arguments, \( a_1 = (1, 0), a_2 = (0, 2), \) and \( a_3 = (0, -1) \), such that the objective conclusion is the same as in Figure 3, but there is more debate over the second issue than the first, then the subjective conclusion is likewise skewed towards the second issue. See Figure 5.
Figure 3. Two Issues; Two Arguments. The agent receives one argument each for issues (i) and (ii). The arguments have equal magnitude and equal malleability, hence the agent scales each by the same factor. The principal axis lies on the diagonal between the two arguments.

Figure 4. Two Issues; One Strong, One Weak Argument. The agent receives one argument each for issues (i) and (ii). The arguments have equal malleability, but the argument for issue (ii) is stronger, hence the principal axis and the subjective posterior are both skewed towards issue (ii).

General Result. The formula in Fact 6 shows how the agent’s multidimensional opinion aligns more with the first principal axis $v^1$ and away from the last axis $v^D$. As in the unidimensional
case, the subjective posterior has a greater magnitude than the objective one, but some directions are stretched more than others, so $G(\gamma; a)^1 > G(\gamma; a)^D$; it follows from this that $\pi_{\text{post}}$ aligns more with $v^1$ and less with $v^D$ relative to $\pi_{\text{obj}}$.

**Fact 6.** For any argument profile,

$$\rho(v^1, \pi_{\text{post}})^2 > \rho(v^1, \pi_{\text{obj}})^2 \quad \text{and} \quad \rho(v^D, \pi_{\text{post}})^2 < \rho(v^D, \pi_{\text{obj}})^2.$$

Furthermore, for any $\varepsilon > 0$, there exists $\hat{k} \geq 1$ such that $\rho(v^1, \pi_{\text{post}})^2 > 1 - \varepsilon$ under any $k$–strengthening with $k \geq \hat{k}$. *Proof in the appendix.*

The last part of Fact 6 confirms that the relative alignment effect is made extreme as the level of bias increases. Note that the overall degree of stretching is bounded between $1/(1 - G_v)$ as calculated for $v = v^1$ and $v = v^D$:

$$\frac{1}{1 - G_v^1(\gamma|a)} < \frac{|\pi_{\text{post}}|}{|\pi_{\text{obj}}|} < \frac{1}{1 - G_v^D(\gamma|a)}.$$  

**Noise Issues.** As a consequence of issue skewing, seemingly unimportant issues can have outsized effects on the dissonance-reducing agent’s opinion when those issues are highly contested. For
example, let \((a_1, \ldots, a_N)\) denote any profile of arguments. For an arbitrarily chosen issue \(d^*\), add two new arguments \(a_{N+1}\) and \(a_{N+2}\) such that the marginal arguments only concern issue \(d^*\) (\(a_{N+k}^d = 0\) for \(d' \neq d^*, k = 1, 2\)) and the two new arguments are directly opposing: \(a_{N+2} = -a_{N+1}\). The introduction of the two marginal arguments serves only to add debate on issue \(d^*\) into the argument set; the Bayesian posterior is unchanged by their addition.

**Fact 7.** Fix \(a_1, \ldots, a_N\). As \(|a_{N+1}| \to \infty\), \(\rho^2(v, a_{N+1}) \to 1\). *Proof in the appendix.*

As the debate on issue \(d^*\) grows arbitrarily large, the agent’s posterior is increasingly determined by first deciding which issues to exaggerate according to the balance of evidence on issue \(d^*\), and then listening to these sources to set the agent’s posterior on all other issues.

4.3. Partisanship with Multiple Issues

The main characteristics of biased learning established in Section 2 have counterparts in the multidimensional environment. This subsection traces through the generalized versions, which – although they are not qualitatively different that the original results – serve as robustness checks. Recall that the agent’s partisanship in the base model consisted of choosing sides in a debate, scaling up the arguments on that side and scaling down arguments on the opposing side. The agent’s chosen side, however, aligned with the balance of evidence.

**Factional Arguments.** With multiple issues, the set of arguments may not divide into two opposing sides, and it is first necessary to formally distinguish those cases where there are two distinct ‘factions.’

**Definition.** The sources are factional if there exist subsets \(I^+, I^- \subset \{1, \ldots, N\}\) such that

\[
\begin{align*}
\langle a_i, a_j \rangle > 0 & \quad \text{if } i, j \in I^+ \text{ or } i, j \in I^- \\
\langle a_i, a_j \rangle < 0 & \quad \text{if } i \in I^+ \text{ and } j \in I^-.
\end{align*}
\]

The agent ‘chooses sides’ if there exist preferred and opposed factions \(I^{\text{pref}}, I^{\text{opp}} \in \{I^+, I^-\}, I^{\text{pref}} \neq I^{\text{opp}}, \) such that

\[
\begin{align*}
\langle \pi_\text{post}, a_i \rangle > 0 & \quad \text{if } i \in I^{\text{pref}} \\
\langle \pi_\text{post}, a_i \rangle < 0 & \quad \text{if } i \in I^{\text{opp}}.
\end{align*}
\]

That is, when the sources can be divided into two mutually agreeing camps which disagree with each other, the agent is said to choose sides if her posterior agrees with all arguments of one faction.
and disagrees with all arguments of the other. It follows that a partisan agent necessarily scales up all of the arguments from her preferred faction and scales down all from the opposing side.

The main result is that, when facing a factional set of sources, the agent is always partisan for sufficiently strong bias.

**Proposition 4.** Suppose the sources are factional. There exists $\bar{k} \geq 1$ such that the agent chooses sides for any $k-$strengthening with $k \geq \bar{k}$. If the Bayesian posterior $\pi^\text{obj}_{\text{post}}$ itself chooses sides, then $\bar{k} = 1$. *Proof in the appendix.*

At first glance, the qualification that partisanship only obtains for sufficiently strong bias may seem to be a weakening of what appeared in the single-topic model. In fact, it is simple consequence of the higher dimensional geometry. With a single issue, the Bayesian posterior itself is guaranteed to agree with one side and disagree with the other. In multiple dimensions this is not the case; even when the sources are themselves factional, the Bayesian posterior need not be. The agent’s subjective posterior is continuous in the malleability factors, so as all of them approach zero, her conclusion converges to the Bayesian. Hence, in cases where the Bayesian posterior is not partisan, the dissonance-reducing agent cannot choose sides for sufficiently low bias. As Proposition 4 notes, when the Bayesian posterior chooses sides, the dissonance-reducing agent is guaranteed to do so even without any strengthening of her bias.

Finally, note that the subjective posterior always agrees with the objective. The logic behind this is the same as in the unidimensional case: it is always cost efficient for the agent to scale up sources that already agree with each other. Due to skewing, the subjective posterior is not a positively scaled copy of the objective, but skewing cannot cause the two posteriors to disagree.

**Fact 8.** For any argument profile, $\langle \pi^\text{post}, \pi^\text{obj}_{\text{post}} \rangle > 0$.

5. Applications

This section discusses the consequences of dissonance reduction in three additional areas of application. The first two return to the single-issue framework of the main model; the third, regarding persuasion, includes multiple issues to identify a meaningful spillover effect across issues.
5.1. Network Learning

What happens when multiple dissonance-reducing agents acquire information from each other? On one hand, the core model with a single agent leads to some sources treated as enemies and scaled down; on the other hand, networks of communicating agents converge to consensus in much prior literature. The exercise below demonstrates that, with minor caveats, the consensus forces dominate. However, dissonance reduction serves to speed up the drive to mutual agreement.

Formally, suppose that $N$ agents $i = 1, \ldots, N$ all begin with the common prior belief $\pi_0 = 0$, then each receives a single private argument $a_i$, producing an initial interim belief $\pi_i^1$. In periods $t = 2, 3, \ldots$, agent $i$ observes the beliefs of agents in the previous period and treats them as communicating new arguments. Agent $i$ computes $\pi_i^t$ by taking the prior beliefs $\{\pi_j^{t-1}\}_{j \in \mathcal{N}_i}$ as the set of arguments. Subset $\mathcal{N}_i \subset \{1, \ldots, N\}$ is $i$’s neighborhood; it is assumed that $i$ is always in her own neighborhood, $i \in \mathcal{N}_i$. For simplicity, each agent $i$ treats the other agents in her neighborhood with the same malleability factor $\gamma_i$. Finally, it is assumed the network is fully connected: for every two agents $i, j$ there is a chain of neighborhood connections between them. The principal question is what happens in the limit after successive updating with dissonance reduction, and the following result establishes that there is a common limit direction of beliefs.

**Proposition 5.** There exists a consensus limit direction $v^* = \pm 1$ such that (1) $\lim_{t \to \infty} \pi_i^t / |\pi_i^t| = v^*$ and (2) $|\pi_i^t| \to \infty$. Proof in the appendix.

Consensus obtains because, with each time step, the agents aggregate the perceived arguments of their neighbors. This has the effect of averaging then extending current beliefs. The iterated extending leads beliefs to have higher and higher magnitudes. The successive averaging operates much as in Markov chains and DeGroot learning, eventually converging to a single limit direction. However, unlike these earlier precedents, the effective weights which agents place on each others’ beliefs do not sum to 1 (nor another constant). Although all agents converge in the same direction, some proceed strictly faster than others. The differential rates of convergence emerge from the same factors that produce eigenvector centrality in the network literature.

The overall effect of dissonance reduction on network learning is evident in $k$-strengthenings. For technical simplicity, assume all of the $\gamma$’s are the same. Two things happen as $k$ increases. First is that the overall rate of convergence increases: in each time step, agents extend their perceptions of
neighbor beliefs by greater amounts. The second effect is that the (normalized) transition matrix
which governs the angular convergence has smaller second, third, and lower eigenvalues, which,
following DeMarzo, Vayanos and Zwiebel (2003), means that directional convergence also happens
quicker.

5.2. Information Acquisition

Consider a dissonance-reducing agent who has received arguments $a_1, \ldots, a_N$ and is considering
whether to accept an additional argument $a_{N+1}$. Assume the agent is aware of her bias, so she
anticipates scaling $a_{N+1}$ as well as rescaling $a_1, \ldots, a_N$. The agent’s indirect utility function shows
how she reasons about the tradeoffs of additional information:

$$V(a|\gamma) = \frac{1}{1 - G(\gamma; a)} \left( \sum_{i=1}^{N} a_i \right)^2.$$ 

In the case that the $N+1$’th argument has the same malleability as the first $N$ arguments, the
agent’s utility would increase after receiving $a_{N+1}$ in two cases: if it supported the existing evidence,
$\langle a_{N+1}, \sum_{i=1}^{N} a_i \rangle > 0$, or if it completely outweighed the existing evidence, $|a_{N+1}| > |\sum_{i=1}^{N} a_i|$. From an ex-ante perspective, the agent would gain in expected utility from accepting the $N+1$’th
argument, because $V(\cdot)$ is strictly convex.\(^9\) If the malleability of the marginal arguments exceeded
that of the first $N$ arguments, the agent’s utility would be even greater. Only in the case that the
marginal arguments’ malleability is substantially lower than the that of the original arguments can
exceptions to the agent’s general pattern of curiosity be found.

5.3. Persuasion

How should a strategic principal leverage the bias of a dissonance-reducing agent to push the
agent’s posterior to agree more with a particular ‘goal’ argument $a_{\text{goal}}$? Suppose the agent has also
received arguments $a_1, \ldots, a_N$. The agent’s subjective posterior is a function of these arguments
as well as $a_{N+1}$, which is endogenously chosen by the principal, whose utility is measured as the
agreement between the agent’s posterior and the goal argument.

If the principal were unconstrained in her choice of $a_{N+1}$, she would necessarily benefit from
sending more extremely scaled versions of $a_{\text{goal}}$. Suppose instead that the principal is constrained to

\(^9\)In probability terms, denoting the agent’s belief by $p$, the indirect utility function is of the form $k(\log p/(1 - p))^2$, which is also strictly convex.
offer only an argument of small fixed magnitude $\varepsilon > 0$. When the magnitude of $a_{N+1}$ is negligible, the principal axes and the exaggeration factors are determined all but exclusively by $a_1, \ldots, a_N$, and therefore the agent’s belief is approximately

$$\pi_{post}(a_{N+1}|a_1, \ldots, a_N)^d \approx \frac{1}{1 - G^d} \left( a_{N+1} + \sum_{i=1}^{N} a_i \right),$$

where the system of axes and the exaggeration factors $G(\gamma; a)^d$ depend only on the original $N$ arguments. The principal’s objective function is accordingly

$$U^{prin.}(a_{N+1}|a_1, \ldots, a_N) \approx \sum_{d=1}^{D} a^{goal,d} \cdot \left( \frac{1}{1 - G^d} \sum_{i=1}^{N+1} a_i \right),$$

and solving for the optimal $a_{N+1}$ under the constraint $|a_{N+1}| = \varepsilon$ yields

$$a_{N+1}^d \propto \frac{a^{goal,d}}{1 - G^d}.$$

This constitutes a compromise between the principal’s goal and the agent’s exaggeration patterns. Effectively ‘flattering’ the agent, the principal scales each dimension of her goal argument by the agent’s exaggeration factor. Note that there is no effect in a single dimension: the principal simply sends the agent a positively scaled version of her goal argument. Only with multiple issues does the principal wish to skew her goal argument to leverage the side-effects of the agent’s dissonance reduction.

6. Discussion

This paper has shown how a dissonance-reducing learner manipulates information she receives from multiple sources in order to increase the agreement she perceives among them. The agent’s distortion leads to three main effects. First, she treats information in a partisan manner by exaggerating the information content from one side of a debate and dampening that from the other. Second, she over-learns, reaching a posterior belief strictly more extreme than an unbiased agent would. Finally, the agent’s chosen side coincides with the objective balance of the evidence.

These effects cause the agent to appear as though she had other biases, however subtly disciplined by her ex-ante neutral drive to perceive consonance. The path dependent nature of a
dissonance-reducing agent receiving new information over time, for instance, resembles confirmation bias. Nonetheless, the direction in which she exaggerates is endogenous, changing with the objective circumstances. With multiple issues, the nature of the agent’s scaling is disproportionately influenced by relatively contentious issues, yielding systematic skewing by topic. Thinking prospectively, a dissonance-reducing agent prefers to receive more information. A network of similarly biased agents eventually converge to a consensus conclusion. A strategic principal warps her goal argument so as to be more amenable to a naive agent’s exaggeration.

**The Model Through Alternative Lenses.** It is worth noting an equivalence between the agent’s motive to increase perceived agreement and two other interpretations. Let \( \pi_{\text{post}}(s|a) = \sum_{i=1}^{N} s_ia_i \) be the agent’s subjective posterior as a function of her scaling factors. Then

\[
\sum_{i,j=1}^{N} \langle s_ia_i, s_ja_j \rangle = \sum_{i=1}^{N} \langle s_ia_i, \pi_{\text{post}}(s|a) \rangle = \langle \pi_{\text{post}}(s|a), \pi_{\text{post}}(s, a) \rangle = |\pi_{\text{post}}(s|a)|^2.
\]

This equation shows that the agent’s objective, the total perceived agreement between her sources, is mathematically equivalent both to the total agreement between the sources and her subjective posterior (‘conformity’) as well as to the agreement of her subjective posterior with itself (‘extremity’). Although prior models of conformity, notably Bernheim (1994), have concentrated on one’s desire to appear virtuous to others, there is also evidence (see Golman et al. (2016)) that people seek to agree with others. Through this lens, the dissonance-reducing agent bolsters evidence from her favored side in an attempt to agree more with those sources. As for the extremity interpretation, Chater and Loewenstein (2016) summarize evidence that people seek out firm conclusions about the world around them, and the agent’s behavior can be understood as strategic manipulation of information in the service of stronger opinions. Finally, the agent’s distortion of information by choosing source-specific scaling factors can be understood as a form of identity. In contrast to theories built on the desire to take group-appropriate actions (Akerlof and Kranton, 2000, 2002, 2005) or relative altruism towards in-group members (Basu, 2006; Chen and Li, 2009), scaling represents an endogenously determined level of affinity.

**Future Applications.** There are various ways to build upon the core model which, although beyond the scope of the present paper, could occupy future work on the subject. For instance,
several simplifying features of the model could be relaxed, such as the agent’s full trust in the veracity of information from the sources. A skeptical agent would presumably react differently in the presence of strategically minded sources. Similarly, an agent who received utility both from perceived agreement among sources and material consequences of actions taken would be constrained to distort information in more complex ways. There are also various additional applications to consider. How might a population of informed but dissonance-reducing voters behave in the context of different policy alternatives? The streamlined case of persuasion studied here could be augmented to consider competition among different strategic principals or a dynamic time frame.
References


Gottlieb, Daniel, “Imperfect Memory and Choice Under Risk,” Games and Economic Behavior,


Appendix A. Proofs

Proof of Proposition 1. Note that this proof is for the general, multi-issue version of the model in Section 4; the statements hold for the model in Section 2 as a special case. The concavity of $U$ given $\gamma_i < 1$ for all $i$ is most clearly demonstrated by expressing $U$ in vector form. Let $s$ denote the profile of scaling factors, let $\Delta$ be an $N \times N$ matrix of all the pairwise agreement between the individual arguments, and let $G$ denote a diagonal matrix with all of the individual malleability factors. That is,

$$
\Delta \equiv \begin{pmatrix}
    \langle a_1, a_1 \rangle & \cdots & \langle a_1, a_N \rangle \\
    \vdots & \ddots & \vdots \\
    \langle a_N, a_1 \rangle & \cdots & \langle a_N, a_N \rangle 
\end{pmatrix}
$$

and

$$
G^{-1} \equiv \begin{pmatrix}
    \frac{\gamma_1}{V(a|a_1)} & \cdots & 0 \\
    \vdots & \ddots & \vdots \\
    0 & \cdots & \frac{\gamma_N}{V(a|a_N)}
\end{pmatrix}.
$$

With this notation, the agent’s utility function can be expressed

$$
U(s|a) = s^T \Delta s - (1 - s)^T G(1 - s) = s^T (\Delta - G)s + 2 \cdot 1^T Gs - 1^T G1.
$$

Denote the matrix of quadratic coefficients by $-W = \Delta - G$. Completing the square,

$$
U(s|a) = -s^T Ws + 2 \cdot 1^T Gs - 1^T G1 = -(s - s^*)^T W(s - s^*) + U^*,
$$

where $s^* = W^{-1} G1$ and $U^* = 1^T GW^{-1} G1 - 1^T G1$. Thus $U$ is strictly concave if and only if $-W$ is negative definite, that is if $-W$ admits only negative eigenvalues. The Gershgorin Circle Theorem guarantees that for any real symmetric matrix $X = (X_{ij})_{i,j}$, any eigenvalue $\lambda$ of $X$ satisfies $|\lambda - X_{ii}| \leq \sum_{j \neq i} |X_{ij}|$. It follows that an upper bound for all eigenvalues is

$$
\max_i \left\{ X_{ii} + \sum_{j \neq i} |X_{ij}| \right\}.
$$
In the context of $-W$, for the column corresponding to argument $i$, the above value is the maximum across $i = 1, \ldots, N$ of
\[
\sum_{j=1}^{N} |\langle m_i, m_j \rangle| - \left( \frac{\gamma_i}{V(a)|a_i|} \right)^{-1} \leq \sum_{j=1}^{N} |a_i| |a_j| - \left( \frac{\gamma_i}{V(a)|a_i|} \right)^{-1} = (1 - \gamma_i^{-1}) |a_i| V(a) < 0,
\]
where the first inequality follows from the Cauchy-Schwartz theorem. This completes the proof of $U$ strictly concave. Next we establish the partial converse. If there exists $i^*$ such that $\gamma_{i^*} > 1$, let $m^*$ be any argument profile such that $a_j^* = 0$ for $j \neq i^*$ and $a_{i^*}^* \neq 0$. Then the utility function simplifies to
\[
U_{m^*}(s) = |a_{i^*}^*|^2 \left( s_{i^*}^2 - \gamma_{i^*}^{-1}(1 - s_{i^*})^2 \right),
\]
which is nowhere concave and admits no optimum.

\begin{proof}[Proof of Fact 3]
Define
\[
f(\pi) = \pi^{\text{obj}}_{\text{post}} + G(\gamma; a) \pi,
\]
so that the dynamic adjustment process can be expressed $\pi_k = f(\pi_{k-1})$, and note
\[
|f(\pi_1) - f(\pi_2)| = G(\gamma; a) |\pi_1 - \pi_2| < |\pi_1 - \pi_2|,
\]
so by the contraction mapping theorem, $f$ admits a unique fixed point, evidently equal to $\pi_{\text{post}}$, to which the gradual adjustment sequence converges.
\end{proof}

\begin{proof}[Proof of Proposition 2]
First we establish convergence to $\xi_{\infty}$. Let $a_-$ and $a_+$ denote the smallest and largest values of $a$, then define the normalized version
\[
\tilde{a}_\tau = \frac{a_\tau - a_-}{a_+ - a_-},
\]
so that $\tilde{a}_\tau \in [0, 1]$. Let $\tilde{\xi}_{\infty} = \gamma \sum_{\tau=1}^{\infty} (1 - \gamma)^{\tau-1} \tilde{a}_\tau$ be the corresponding normalized version of $\xi_{\infty}$. The sequence of random variables
\[
\left\{ \gamma \sum_{\tau=1}^{t} (1 - \gamma)^{\tau-1} \tilde{a}_\tau \right\}_{t=1}^{\infty}
\]

38
is monotonically increasing (as $\tilde{a}_\tau \geq 0$) and uniformly bounded (by 1), and therefore converges to $\tilde{\xi}_\infty$. Hence $\xi_t$ converges to $\xi_\infty$. Next we establish that the distribution of $\tilde{\xi}_\infty$ has no mass points. Towards a contradiction, suppose that $\tilde{\xi}_\infty$ did admit mass points. In that case, there would have to be a largest mass point size $M$. As $P[\tilde{\xi}_\infty = 0] = P[\tilde{\xi}_\infty = 1] = 0$, there must be some $x \in (0, 1)$ such that $P[\tilde{\xi}_\infty = x] = M$. Noting that

$$\tilde{\xi}_\infty \overset{d}{=} \gamma a_1 + (1 - \gamma)\tilde{\xi}_\infty,$$

it follows

$$P[\tilde{\xi}_\infty = x] = \sum_{\tilde{a} \in \text{supp}\{\tilde{a}_1\}} P[\tilde{a}_1 = \tilde{a}] \cdot P[\tilde{\xi}_\infty = \frac{x - \gamma \tilde{a}}{1 - \gamma}].$$

As $M$ is the largest mass point size, it follows

$$P[\tilde{\xi}_\infty = \frac{x - \gamma \tilde{a}}{1 - \gamma}] = M$$

for all $\tilde{a} \in \text{supp}\{\tilde{a}_1\}$. In particular, setting $\tilde{a} = 0$ and repeating this argument implies

$$P[\tilde{\xi}_\infty = x] = P[\tilde{\xi}_\infty = \frac{x}{1 - \gamma}] = P[\tilde{\xi}_\infty = \frac{x}{(1 - \gamma)^2}] = \cdots = M,$$

an obvious impossibility. Conclude that $\tilde{\xi}_\infty$ has no mass points, so

$$P[\pi_t \to -\infty] + P[\pi_t \to \infty] = P[\tilde{\xi}_\infty < \frac{-a_-}{a_+ - a_-}] + P[\tilde{\xi}_\infty > \frac{-a_-}{a_+ - a_-}] = 1.$$

Lastly we establish $\tilde{\xi}_\infty$ is arbitrarily close both to 0 and 1 with positive probability: if $\tilde{a}_1 = \cdots = \tilde{a}_t = 0$, the maximum conditional value of $\tilde{\xi}_\infty$ is $\gamma \sum_{\tau=t+1}^{\infty} (1 - \gamma)^{\tau-1}$, so

$$P[\tilde{\xi}_\infty \leq \gamma \sum_{\tau=t+1}^{\infty} (1 - \gamma)^{\tau-1}] \geq P[\tilde{a}_1 = 0]^t,$$

and, similarly,

$$P[\tilde{\xi}_\infty \geq \gamma \sum_{\tau=1}^{t} (1 - \gamma)^{\tau-1}] \geq P[\tilde{a}_1 = 1]^t.$$

This implies

$$P[\pi_t \to -\infty] = P[\tilde{\xi}_\infty < \frac{-a_-}{a_+ - a_-}] > 0 \quad \text{and} \quad P[\pi_t \to \infty] = P[\tilde{\xi}_\infty > \frac{-a_-}{a_+ - a_-}] > 0.$$
Proof of Proposition 3. Let \( a_t \) take value \( a_R \) with probability \( q_R \) and \( a_L \) with probability \( q_L \). To construct a distribution with arbitrarily certain mis-learning, fix \( a_L \equiv -1 \), and set
\[
a_R = \frac{q_L}{q_R} + 1
\]
as to satisfy the requirement that \( q_R a_R + q_L a_L > 0 \). Then, borrowing notation from the previous proof,
\[
P[\pi_t \rightarrow -\infty] = P\left[\xi_\infty < \frac{-a_L}{a_R - a_L}\right] = P\left[\xi_\infty < \frac{1}{2 + q_L/q_R}\right].
\]
Setting \( q_L > 1 - \varepsilon \) and
\[
\gamma > 1 - \frac{1}{2 + q_L/q_R},
\]
Equation (3) implies \( P[\pi_t \rightarrow -\infty] > 1 - \varepsilon \). To construct a distribution with arbitrarily certain correct learning, set \( q_R > 1 - \varepsilon \) and
\[
\gamma > \frac{1}{2 + q_L/q_R},
\]
so by Equation (4), \( P[\pi_t \rightarrow \infty] > 1 - \varepsilon \).

\[\Box\]

Proof of Fact 6. For the first statement,

\[
\rho(\pi_{post},v^1)^2 = \frac{1}{|\pi_{post}|^2} \left( \frac{1}{1 - G(\gamma; a)^1} \right)^2 \left( \pi_{obj,1} \right)^2 \\
= \left( \sum_{d=1}^{D} \left( \frac{1}{1 - G(\gamma; a)^D} \right)^2 \left( \pi_{obj,D} \right)^2 \right)^{-1} \left( \frac{1}{1 - G(\gamma; a)^1} \right)^2 \left( \pi_{obj,1} \right)^2 \\
= \left( \sum_{d=1}^{D} \left( \frac{1 - G(\gamma; a)^1}{1 - G(\gamma; a)^D} \right)^2 \left( \pi_{obj,D} \right)^2 \right)^{-1} \left( \pi_{obj,1} \right)^2 \\
> \left( \sum_{d=1}^{D} \left( \pi_{obj,D} \right)^2 \right)^{-1} \left( \pi_{obj,1} \right)^2 = \frac{1}{|\pi_{post}|^2} \left( \pi_{obj,1} \right)^2 = \rho(\pi_{post},v^1)^2.
\]

An identical proof shows \( \rho(\pi_{post},v^D)^2 < \rho(\pi_{obj},v^D)^2 \). As for guaranteeing \( \rho(\pi_{post},v^1)^2 > 1 - \varepsilon \) under a sufficiently strong strengthening, let \( \tilde{\gamma}_i = k \cdot \gamma_i \) and \( \tilde{G}(\gamma; a)^d \) denote the malleability.
parameters and dimension exaggeration factors under the $k$–strengthening. Then
\[
\rho(\pi_{\text{post}}, v^1)^2 = \left( \sum_{d=1}^{D} \frac{1 - \tilde{G}(\gamma; a_1)}{1 - G(\gamma; a_1^d)} \right)^2 \left( \pi_{\text{obj,}D} \right)^2 \left( \frac{\pi_{\text{obj,}1}}{\pi_{\text{post}}} \right)^2
\]
\[
= \left( \frac{\pi_{\text{obj,}1}}{\pi_{\text{post}}} \right)^2 + \sum_{d>1} \left( \frac{1 - k \cdot G(\gamma; a_1^d)}{1 - k \cdot G(\gamma; a_1^d)} \right)^2 \left( \frac{\pi_{\text{obj,}D}}{\pi_{\text{post}}} \right)^2 \left( \frac{\pi_{\text{obj,}1}}{\pi_{\text{post}}} \right)^2
\]
\[
> \left( \frac{\pi_{\text{obj,}1}}{\pi_{\text{post}}} \right)^2 + \sum_{d>1} \left( \frac{1 - k \cdot G(\gamma; a_1^d)}{1 - k \cdot G(\gamma; a_1^d)} \right)^2 \left( \frac{\pi_{\text{obj,}D}}{\pi_{\text{post}}} \right)^2 \left( \frac{\pi_{\text{obj,}1}}{\pi_{\text{post}}} \right)^2.
\]
As
\[
\lim_{k \to 1/G(\gamma; a_1)^1} \frac{1 - k \cdot G(\gamma; a_1)}{1 - k \cdot G(\gamma; a_1)^d} = 0,
\]
it follows \(\lim_{k \to 1/G(\gamma; a_1)^1} \rho(\pi_{\text{post}}, v^1)^2 = 1\), proving the statement.

\[\Box\]

**Proof of Fact 7.** The addition of \(a_{N+1}\) and \(a_{N+2}\) makes the directional malleability factor function
\[
G_v(\gamma|a) = \frac{2|a_{N+1}|}{V(a)} \gamma_{N+1} \cdot \rho(v, a_{N+1})^2 + \sum_{i=1}^{N} \left( \frac{|a_i|}{V(a)} \gamma_i \cdot \rho(v, a_i)^2 \right).
\]
As \(|a_{N+1}| \to \infty\),
\[
\frac{2|a_{N+1}|}{V(a)} \to 1 \quad \text{and} \quad \frac{|a_i|}{V(a)} \to 0 \quad \text{for} \ i = 1, \ldots, N,
\]
so it follows that the maximum value of \(G_v\) is attained for \(v\) arbitrarily aligned with \(a_{N+1}\).

\[\Box\]

**Proof of Proposition 4.** Let
\[
\mathcal{X} \equiv \{a_i : i \in \mathcal{I}^+\} \cup \{-a_i : i \in \mathcal{I}^-\}
\]
be the set of arguments from \(\mathcal{I}^+\) and ‘anti-arguments’ from \(\mathcal{I}^-\). Note that \(\langle a_1, a_2 \rangle > 0\) for any \(a_1, a_2 \in \mathcal{X}\), so \(\mathcal{X}\) is contained in a single half-space of \(\mathbb{R}^D\). Let \(\mathcal{C}(\mathcal{X})\) be the convex cone generated by \(\mathcal{X}\). We can show that \(v^1 \in \mathcal{C}(\mathcal{X})\). By contradiction: if \(v^1 \notin \mathcal{C}(\mathcal{X})\), where without loss of generality \(v^1\) is contained in the same half-space as \(\mathcal{X}\), then let \(\hat{v}^1\) be the orthogonal projection of \(v^1\) onto \(\mathcal{C}(\mathcal{X})\). Note that for any two arguments \(a_1, a_2\) on the unit sphere intersected with a
common half-space, \( \rho(a_1, a_2)^2 \) is strictly decreasing in \( |a_1 - a_2| \). Thus

\[
\frac{|\hat{v}^1| - a_i}{|a_i|} \leq \frac{|v^1| - a_i}{|a_i|}
\]

implies

\[
\rho(\hat{v}^1, a_i)^2 = \rho\left(\frac{\hat{v}^1}{|\hat{v}^1|}, \frac{a_i}{|a_i|}\right)^2 \geq \rho\left(\frac{v^1}{|v^1|}, \frac{a_i}{|a_i|}\right)^2 = \rho(v^1, a_i)^2,
\]

so \( G_{\hat{v}^1}(\gamma; a) \geq G_{v^1}(\gamma; a) \). By the definition of \( v^1 \), it must be that \( v^1 = \hat{v}^1 \in C(\mathcal{X}) \), a contradiction.

Since \( v^1 \in C(\mathcal{X}) \), there exist coefficients \( \alpha_i \geq 0 \) for \( i \in I^+ \) and \( \beta_j \leq 0 \) for \( j \in I^- \) such that

\[
v^1 = \sum_{i \in I^+} \alpha_i a_i + \sum_{j \in I^-} \beta_j a_j.
\]

Because

\[
\langle a_l, v^1 \rangle = \sum_{i \in I^+} \alpha_i \langle a_i, a_l \rangle + \sum_{j \in I^-} \beta_j \langle a_j, a_l \rangle \begin{cases} > 0 \text{ if } l \in I^+ \\ < 0 \text{ if } l \in I^- \end{cases},
\]

it follows that there exists some \( \hat{\rho}^2 < 1 \) such that for all \( v \) with \( \rho(a, v^1)^2 > \hat{\rho}^2 \) then either (depending on the half-space in which \( v \) sits)

\[
\langle a_l, v \rangle \begin{cases} > 0 \text{ if } l \in I^+ \\ < 0 \text{ if } l \in I^- \end{cases} \text{ or } \langle a_l, v \rangle \begin{cases} < 0 \text{ if } l \in I^+ \\ > 0 \text{ if } l \in I^- \end{cases}.
\]

By Fact 6, \( \rho(v^1, \pi_{\text{post}})^2 \) is arbitrarily close to 1 for a sufficiently high \( k \)–strengthening. Thus there exists some \( \bar{k} \) such that for \( k \)–strengthenings with \( k \geq \bar{k} \) the agent chooses sides.

\( \diamond \)

**Proof of Proposition 5.** Let \( \mathcal{M} \) denote the ‘transition’ matrix such that \( \pi_{t+1} = \mathcal{M} \pi_t \). We note that \( \mathcal{M} \) must satisfy several basic properties which govern the long-term behavior of the agents’ beliefs. Because the network is fully connected, \( \mathcal{M} \) is irreducible; as \( \mathcal{M}_{ii} > 0 \) for all agents \( i \), \( \mathcal{M} \) is aperiodic. Being both irreducible and aperiodic, \( \mathcal{M} \) is also primitive: \( \mathcal{M}^t \) contains only strictly positive entries for some finite power \( t \). The Perron-Frobenius theorem thus provides that the largest eigenvalue \( \lambda_1 \), denoted, is real-valued, positive, and has a single-dimensional associated eigenspace. Furthermore, denoting \( v \) to be the associated right eigenvector and \( u \) to be the associated left eigenvector, scaled such that \( u \cdot v = 1 \), successive powers of the normalized matrix \( (\mathcal{M}/\lambda_1)^t \) converge to \( \mathcal{M}^\infty = uv' \). Moreover, \( u_i > 0 \) and \( v_i > 0 \) for all \( i = 1, \ldots, N \). Hence,
any one agent $i$'s beliefs converge as:

$$\lim_{t \to \infty} \frac{\pi_i^t}{\lambda_i^t} \rightarrow u_i \bar{\pi}_\infty,$$

where $\bar{\pi}_\infty = \sum_{j=1}^N v_j \pi_j^0$ is the common limit direction.

---

**Appendix B. Reasons for Key Assumptions**

**B.1. Functional Form of Agreement**

The linear functional form adopted in this paper is rooted in the following axiom. Suppose two arguments $a_1$ and $a_2$ have a given level of agreement $\langle a_1, a_2 \rangle \in \mathbb{R}$ and that argument $a_1$ is then ‘decomposed’ into $a_1'$ and $a_1''$, such that $a_1 = a_1' + a_1''$. Because the components $a_1'$ and $a_1''$ combine to produce the original argument $a_1$, it ought to be that the sum of the agreements $\langle a_1', a_2 \rangle$ and $\langle a_1'', a_2 \rangle$ should likewise combine to produce $\langle a_1, a_2 \rangle$. As the following Lemma 1 shows, the only functional form that meets this criterion and satisfies several regularity conditions is an inner product. This parallels a similar result in Chauvin (2020) concerning agreement functions for more general forms of arguments; the proof here is more direct.

**Lemma 1.** Symmetric function $f : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ is an inner product if and only if

1. for all arguments $a_1' + a_1'' = a_1$ and $a_2$, $f(a_1, a_2) = f(a_1', a_2) + f(a_1'', a_2)$.
2. $f$ is continuous.
3. $f(a, a) > 0$ for any argument $a \neq 0$.

*Proof in the appendix.*

**Proof of Lemma 1.** An inner product is symmetric and satisfies the three properties by inspection. It remains to show any symmetric $f : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ satisfying (1)-(3) is itself an inner product, and for this all is needed is to show $f$ is bilinear, that $f(ka_1' + a_1'', a_2) = kf(a_1', a_2) + f(a_1'', a_2)$ for all $k \in \mathbb{R}$ and arguments $a_1', a_1'', a_2$. Note that $f(ka_1' + a_1'', a_2) = f(ka_1', a_2) + f(a_1'', a_2)$ by the decomposition principle, so it suffices to show $f(ka_1, a_2) = kf(a_1, a_2)$ for all $k \in \mathbb{R}$.
If \( k = 1 \), the proposition is trivial; we proceed by induction for \( k = 2, 3, \ldots \):

\[
f(ka_1, a_2) = f((k-1)a_1 + a_1, a_2) = f((k-1)a_1, a_2) + f(a_1, a_2)
\]

\[
= (k-1)f(a_1, a_2) + f(a_1, a_2) = kf(a_1, a_2).
\]

For \( k = 0 \), 

\[
f(a_1, a_2) = f(a_1 + 0a_1, a_2) = f(a_1, a_2) + f(0a_1, a_2),
\]

implying 

\[
f(0a_1, a_2) = 0 = 0f(a_1, a_2).
\]

For \( k = -1, -2, \ldots \),

\[
f(ka_1, a_2) + f(-ka_1, a_2) = f(ka_1, a_2) - kf(a_1, a_2) = f(0a_1, a_2) = 0,
\]

completing the proof for \( k \in \mathbb{Z} \). If \( k \in \mathbb{Q} \), then \( k = p/q \) for \( p, q \in \mathbb{Z} \), \( q \neq 0 \), so

\[
qf(ka_1, a_2) = f((qk)a_1, a_2) = f(pa_1, a_2) = pf(a_1, a_2)
\]

\[
\implies f(ka_1, a_2) = (p/q)f(a_1, a_2) = kf(a_1, a_2).
\]

For \( k \in \mathbb{R} \), let \( k_n \in \mathbb{Q} \) be any sequence for which \( k_n \rightarrow k \). Then, by continuity,

\[
f(ka_1, a_2) = \lim_{n \to \infty} f(k_na_1, a_2) = \lim_{n \to \infty} k_nf(a_1, a_2) = \left( \lim_{n \to \infty} k_n \right) f(a_1, a_2) = kf(a_1, a_2).
\]

With multiple issues, the specific form of the agreement inner product is assumed to be

\[
\langle m_1, m_2 \rangle = \sum_{d=1}^{D} w_d \cdot m_{1d} \cdot m_{2d},
\]

which guarantees that the individual issues are treated as orthogonal.

\textbf{B.2. Functional Form of Scaling Costs}

The cost functions are calibrated in a way that advances three interests. First, the functional form isolates for each \( i \) a source-specific and argument-independent malleability parameter \( \gamma_i > 0 \) identifying the ease with which the agent can adjust the source’s arguments. Second, it links adjustment costs with the magnitudes of the sources’ arguments so that \( \gamma_i = 1 \) is a uniform upper bound which guarantees the concavity of \( U \). Third, it guarantees the utility function is homogenous of degree 2, making the agent insensitive to the absolute magnitudes of the arguments.

As the following proposition shows, the only functional forms that satisfy these three interests are
quadratic cost functions with a normalization factor of a form closely related to the specification used in the main text, \( c_i(a) = |a_i| \sum_{j=1}^{n} |a_j| \).

**Proposition 6.** Let \( U(s|a) = \sum_{i,j=1}^{N} (s_i a_i s_j a_j) - \sum_{i=1}^{N} \gamma_i^{-1} \cdot c_i(a) \cdot (1 - s_i)^2 \). Then

1. \( c_i \) is smooth, direction invariant, and semi-symmetric: \( c_i(a) = c(|a_i|, f(|a_1|, \ldots, |a_N|)) \) for some symmetric function \( f \), and
2. \( U \) is homogenous of degree 2 in \( a \) if and only if \( c_i \) is of the form
   \[
   c_i(a) = \beta_{ii} a_i^2 + \beta_{ij} a_i a_j \quad \text{or} \quad c_i(a) = \beta_{ii} a_i^2 + \beta_{jj} a_j^2 + \beta_{kl} a_k a_l.
   \]

**Proof of Proposition 6.** If conditions (1) and (2) are satisfied, the first order conditions for \( U \) show that homogeneity of degree 2 requires \( c_i(ka) = k^2 c_i(a) \). Now, let \( \alpha = (\alpha_1, \ldots, \alpha_N) \) denote a profile of exponents, with \( a^\alpha = a_1^{\alpha_1} \cdots a_N^{\alpha_N} \), and let \( \#\alpha = \sum_{i=1}^{n} \alpha_i N \). Writing \( c_i \) as a Taylor expansion around 0 produces

\[
c_i(a) = \sum_{\alpha} \beta_\alpha a^\alpha,
\]
where the \( \beta_\alpha \in \mathbb{R} \) coefficients are determined by Taylor’s theorem. The fact that \( c_i(ka) = k^2 c_i(a) \) means

\[
\left[ \sum_{j=0}^{\infty} \left( \sum_{\#\alpha = j} \beta_\alpha a^\alpha \right) \right] k^2 = \sum_{j=0}^{\infty} \left( \sum_{\#\alpha = j} \beta_\alpha a^\alpha \right) k^j.
\]
Recall that any polynomial function can only be expressed with a unique profile of coefficients.\(^{10}\)

Viewing both sides of the equation as polynomial functions of \( k \) implies

\[
\sum_{\#\alpha = j} \beta_\alpha a^\alpha = 0
\]
for all \( j \neq 2 \), and viewing both sides of this equation as polynomial functions of \( a \) implies \( \beta_\alpha = 0 \) for all \( \alpha \) with \( \#\alpha \neq 2 \). This shows \( c_i \) must be of the form

\[
c_i(a) = \sum_{j=1}^{n} \beta_{jj} a_j^2 + \sum_{j,k=1}^{n} \beta_{jk} a_j a_k.
\]

\(^{10}\) Suppose \( f \) is a polynomial function of one variable, with \( f(x) = \sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n \). Then \( g(x) = \sum_{n=0}^{\infty} (a_n - b_n) x^n = 0 \). As \( g(x) \) is identically equal to 0, the \( n \)’th derivative of \( g \) at \( x = 0 \) must also be 0. Hence, \( g^{(n)}(0) = n!(a_n - b_n) = 0 \), so \( a_n = b_n \) for all \( n \). An analogous proof establishes the statement when \( f \) is a polynomial function of multiple variables.
The imposition of the constraint \( c_i(a) = c(|a_i|, f(|a_1|, \ldots, |a_N|)) \) narrows the possible functional forms to those in the proposition statement. Finally, simple algebra confirms that any cost functions of the above form satisfy conditions (1) and (2).

B.3. Uniqueness of Principal Axes

**Lemma 2.** For any argument profile \( a \) and any \( \varepsilon > 0 \) there exists a profile \( \tilde{a} \) with at most \( N + D \) sources and
\[
\sum_{i=1}^{N} |\tilde{a}_i - a_i| + \sum_{j=N+1}^{N+D} |\tilde{a}_j| < \varepsilon
\]
such that \( \tilde{a} \) admits a unique system of ordered axes \( \{v^d\}_{d=1}^{D} \).

**Proof of Lemma 2.** Given argument profile \( a \) with a system of ordered axes \( \{\bar{v}^d\}_{d=1}^{D} \), and given a value of \( \varepsilon > 0 \), let \( \tilde{a} \) be defined by \( \tilde{a}_i = a_i \) for \( i = 1, \ldots, N \), and \( \tilde{a}_{N+d} = \varepsilon \cdot \left( \frac{1}{d} \right) \cdot 1_{(d=\varepsilon)} \) for \( d, e = 1, \ldots, D \). Sources \( N + 1, \ldots, N + D \) are given common malleability factor \( \gamma \). As the individual orthogonal projections \( a_{N+d} a'_{N+d} / |a_{N+d}|^2 \) all share a common eigendecomposition with the original profile’s \( \Gamma(\gamma|a) \), it follows that \( \tilde{\Gamma} \) also shares the same eigendecomposition, with
\[
\tilde{G}(\gamma; a)^d = \left( \frac{V(a)}{V(a) + \sum_{e=1}^{D} \varepsilon \cdot (1/e)} \right) \cdot G(\gamma; a)^d + \left( \frac{\sum_{e=1}^{D} \varepsilon \cdot (1/e)}{V(a) + \sum_{e=1}^{D} \varepsilon \cdot (1/e)} \right) \cdot \varepsilon \cdot (1/d).
\]

Thus for any \( \varepsilon \in (0, \varepsilon) \), where \( \varepsilon > 0 \) is set sufficiently small, it follows that for any pair of dimensions \( d < e \),
\[
\circ \text{ If } G(\gamma; a)^d = G(\gamma; a)^e, \text{ then } \tilde{G}(\gamma; a)^d > \tilde{G}(\gamma; a)^e;
\]
\[
\circ \text{ If } G(\gamma; a)^d > G(\gamma; a)^e, \text{ then } \tilde{G}(\gamma; a)^d > \tilde{G}(\gamma; a)^e, \text{ and}
\]
\[
\circ \text{ If } G(\gamma; a)^d < G(\gamma; a)^e, \text{ then } \tilde{G}(\gamma; a)^d < \tilde{G}(\gamma; a)^e.
\]

This verifies that \( \tilde{a} \) has a unique system of ordered axes.