ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 411

Robustly Non-Harmful Information for Biased Learners

Malte Kornemann

May 2026

www.econtribute.de

UNIVERSITÄT BONN

UNIVERSITÄT ZU KÖLN

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866 – is gratefully acknowledged.

Robustly Non-Harmful Information for Biased Learners *

Malte Kornemann

University of Bonn

May 26, 2026

Abstract

I examine when robustly beneficial information can be provided to a receiver who also learns from misspecified background sources that are outside the provider's control. In contrast to information provision for rational receivers, any source can be harmful under certain misspecifications of background sources. I show that the key aspect of the background environment enabling robustly beneficial design is the receiver's perception rather than the true structure. For any background source structure and design of the provided source, there exists a misspecification under which harm occurs. Consequently, even complete knowledge of the true structure is insufficient and knowledge of the receiver's perception is necessary. Under complete knowledge of the perception, I demonstrate how to design an information source that is robustly non-harmful and often strictly beneficial, regardless of the true background sources.

Keywords: misspecified learning, value of information, robust information design

JEL Codes: D80, D83, D90

*I am deeply grateful to Botond Kőszegi, whose generous help and time have been invaluable at all stages of this paper. I also thank Sarah Auster, Roberto Corrao, Daniel Hauser, Elliot Lipnowski, Larry Samuelson, Philipp Strack, and workshop participants at Bonn and Yale for their valuable comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866.

1 Introduction

Real-world information providers rarely control or fully know all sources receivers consult when making decisions. Yet for a rational receiver who correctly understands and integrates external sources, any additional information is robustly non-harmful. In practice, however, receivers may exhibit learning biases about external sources that providers neither control nor fully understand. How does an additional source interact with misinterpreted external sources? How is the value of provided information affected? And when can one design information that is robustly non-harmful?

These challenges arise in various settings. A policymaker may aim to convey the importance of economic interventions to voters, but voters may consume information from media outlets and interpret it through a biased lens. Public health agencies seek to promote vaccination benefits, while social media and personal anecdotes may over-amplify selective narratives, an effect users may fail to recognize. An investment advisor might highlight key financial market statistics to an investor who relies on other sophisticated sources but interprets them using oversimplified heuristics. Even an expert advising a high-stakes decision maker, such as a CEO or judge, is rarely the only information source and cannot ensure that the receiver is unaffected by biases[1].

I study these questions in a setting where the receiver is a misspecified learner with two types of information sources: (i) background sources, outside the provider's control and only partially within their knowledge, and (ii) an additional controlled source supplied by the provider. I primarily consider the case where the provider ensures correct understanding of their own source, such that only background sources are misperceived. Even in this favorable case, additional information can worsen inference and I establish three main results on when a source can be robustly non-harmful. First, in a multidimensional state space, any additional source can harm the receiver. This sharply contrasts with the benchmark of information provision to correctly specified receivers, where any additional information robustly has weakly positive value. Second, even full knowledge of the true data-generating

[1]For evidence on biases of CEOs see for example Malmendier and Tate (2005). For judges see Guthrie et al. (2007) and Danziger et al. (2011). While these studies do not examine learning directly, the biases they document plausibly affect learning processes and, more generally, indicate that such high-stakes decision makers are susceptible to biases.

process is insufficient to restore robustly non-harmful design. Third, I characterize which knowledge of the receiver's perception allows for robustly non-harmful design and show that full knowledge is sufficient.

Formally, I model a single agent who learns about fundamentals, represented by a real vector f, through multiple information sources. Each source generates periodic noisy observations about one fixed linear combination of fundamentals over an infinite discrete-time horizon. The agent is misspecified, believing that these sources yield signals about different linear combinations. Starting from a full-support prior, the agent updates their beliefs according to Bayes' rule. The true structure of background sources is then characterized by a matrix M, where each row models the linear combination associated with one source and the agent's (mis)perception is given by a possibly different matrix N. The controlled information source has the same structure and produces periodic signals about a fixed linear combination alongside background sources. I prove that the agent's beliefs converge to a single point, denoted $\tilde{f}$, which I refer to as the inferred fundamentals. To illustrate this abstract setting and the key mechanism of misinference substitution I provide an example of a receiver learning about asset returns in an investment decision setting.

The effect of a controlled source on inference is captured by the difference between inferred fundamentals with and without it, for which I derive a tractable expression. The direct effect of an additional source is positive: it improves inference along the dimension it targets. However, through the interaction with misspecified background sources it can simultaneously distort inference in other dimensions. To maintain maximal consistency between observations and their misspecified model, the receiver may compensate for the reduced error in one dimension by worsening errors in another. I call this mechanism *misinference substitution*. When misinference substitution occurs, it generates a trade-off with the improving direct effect. To assess the overall value of a controlled source, I consider the distance, primarily using the Euclidean norm, between inferred fundamentals and true ones.

I analyze when an additional source is robustly non-harmful with respect to the true structure and perception of background sources, i.e., considering the value of a fixed controlled source across all possible M and N. Misinference substitution operates across dimensions

and therefore the existence of robustly non-harmful information crucially depends on the dimensionality. When f is one-dimensional, all information is robustly non-harmful. With multiple fundamentals, however, for any controlled source there exist a true structure and perception of background sources under which misinference substitution occurs and the receiver is better off without the additional information. Moreover, the magnitude of harm can be unbounded. Hence, when background sources may be misperceived, designing robustly beneficial information becomes impossible without further knowledge of those sources.

Nevertheless, non-harmful information always exists and providing strictly positive value is possible whenever initial inference is imperfect. Achieving such robust design requires additional knowledge of background sources and their perception. Specifically, such knowledge must enable a design that limits the occurrence of misinference substitution. I show that misinference substitution is primarily governed by the interaction of the controlled source and the receiver's perception. The true structure has no substantive effect on the mechanism and knowledge of it alone is insufficient: fixing a true structure of background sources and considering any controlled source, there exists a perception under which harm occurs. Knowledge of the receiver's perception can allow the construction of a source for which misinference substitution does not occur. Such a source has a (weakly) positive direct effect without any potentially harmful indirect effects and is therefore robustly non-harmful. I fully characterize which knowledge is necessary and sufficient absent any knowledge about the true structure.

Finally, I discuss the robustness of the established impossibility of robustly non-harmful information. By their nature, these results continue to hold when the set to which robustness is required expands; for example, when the decision environment is not known and therefore the value of an additional source is not restricted to be the difference of Euclidean norms. I show that they also hold under further restrictions of the misspecification. Most notably, they extend when the misspecification is arbitrarily close to the truth, highlighting the fragility of the classical result. I also consider alternative structures of additional information and show that the robustness is not restored when allowing for incorrectly perceived additional sources or by providing multiple sources. In particular, matching the dimensionality of the controlled source and the fundamentals does not eliminate the difference between one- and

multidimensional settings. Finally, I consider different model assumptions and discuss extensions to other, fixed measures of value, alternative structures of the true background sources, and sufficiently precise finite samples.

Section 2 provides the investment decision example. Section 3 introduces the misspecified learning model and solves for the learning outcome. Section 4 analyzes the effect of additional information on inference. Section 5 discusses main results on the harm of information and robustness, followed by extensions in Section 6. Section 7 concludes by discussing avenues for further research. All proofs are provided in the Appendix.

Related Literature. This paper contributes to a broad literature on information provision, the value of information, and robustness considerations. It also relates to the growing body of research on misspecified learning and non-Bayesian updating[2]. However, no prior work has analyzed the central question addressed here: under what conditions can additional information be robustly beneficial with respect to the receiver's learning biases, such as misspecified beliefs about background sources as in this paper. Moreover, most existing studies treat the receiver's perception jointly with the true data-generating process. Hence, the distinct role of perception, decoupled from the true structure, remains largely unexplored.

Robustness with respect to priors and decision environments in information provision for rational receivers has been widely analyzed (e.g., Blackwell, 1951, 1953, Radner and Stiglitz, 1984, Morris, 1991, Moscarini and Smith, 2002, Azrieli, 2014, Mu et al., 2021). In that setting, which includes receivers consulting correctly specified background sources, it is well established that all information has weakly positive value (Blackwell, 1951) and that partial orders can robustly rank experiments' welfare effects (Blackwell, 1951, 1953, Mu et al., 2021).

Research on robustness in non-Bayesian updating and misspecified learning typically fixes the learning bias and studies robustness with respect to the decision problem. For instance, Frick et al. (2024) adopt a robustness approach to compare the welfare effects of different learning biases across decision environments. Whitmeyer (2023) and Bordoli (2024) characterize updating rules, belonging to a certain class of distortions, for which all Blackwell

[2]For an in-depth comparison of these two approaches and their formal connection, see Bohren and Hauser (2023).

experiments are robustly preferred across decision problems[3]. Morris and Shin (1997) analyze experiments where agents misperceive message likelihoods given states, while Braghieri (2023) examines a closely related model where beliefs about both priors and conditional signal distributions deviate from the truth. Both papers focus on robustness across decision environments for a fixed misperception.

On the technical side, I contribute to the literature on misspecified learning by providing new tractable results in a concrete multidimensional setting. Such results are relatively scarce and may therefore be of independent interest for other applications. The most closely related papers are Heidhues et al. (2026) and He et al. (2023), both of which also solve multidimensional linear Gaussian learning models. Although the learning setups are closely related, these models study different types of misspecification. Heidhues et al. (2026) focus on a direct misspecification about a single fundamental and the effects on inferring other variables. He et al. (2023) consider a similar misspecification extended to multiple fundamentals and apply it to an agent predicting a time-dependent variable using inferences about what this paper refers to as fundamentals. Both papers also address fundamentally different questions from those explored here. Beyond this, other papers exploring multidimensional settings (e.g., Hestermann and Le Yaouanq, 2021, He, 2022, Chauvin, 2023) feature very different learning environments and, again, address unrelated questions.[4]

2 An Investment Decision Example

To illustrate the key mechanism in this paper, I consider a portfolio choice problem with two risky assets. Let $f = (f_1, f_2)^\top$ denote the unknown vector of mean returns, where f_1 and f_2 are the mean returns above the risk-free rate of assets 1 and 2, respectively. The

[3]Whitmeyer (2023) evaluates this under a conventionalist view based on realized utility, while Bordoli (2024) also considers a prospective view based on anticipatory utility.

[4]Beyond the papers mentioned, a broader literature examines general solution concepts and convergence properties of beliefs under model misspecification (e.g., Esponda and Pouzo, 2016, 2021, Bohren and Hauser, 2021, Esponda et al., 2021, Fudenberg et al., 2021, Frick et al., 2023) that can be applied to multidimensional settings. Yet, these provide primarily abstract results rather than closed-form solutions as in this paper. An even more distantly related literature examines the effects of specific learning biases in one-dimensional or binary state environments, both in single-agent settings (e.g., Fudenberg et al., 2017, Heidhues et al., 2018, 2021, 2023, Gagnon-Bartsch and Bushong, 2022) and in social learning contexts (e.g., Eyster and Rabin, 2010, Bohren, 2016, Frick et al., 2020).

covariance matrix of returns is, for simplicity, known and equal to the identity matrix. The agent has mean-variance preferences and chooses an allocation $a \in \mathbb{R}^2$ based on their belief $\tilde{f}$ to maximize

$$U(a, \tilde{f}) = \tilde{f}^\top a - \frac{1}{2} a^\top a.$$

Standard first-order conditions yield the optimal allocation $a^*(\tilde{f}) = \tilde{f}$ and the corresponding realized utility, evaluated at the true returns f, is

$$U(a^*(\tilde{f}), f) = \tilde{f}^\top f - \frac{1}{2}\tilde{f}^\top \tilde{f} = \frac{1}{2}||f||_2^2 - \frac{1}{2}||\tilde{f} - f||_2^2.$$

In particular, the induced value from a belief $\tilde{f}$ is a strictly decreasing function of the Euclidean distance between beliefs and the truth.

Information sources and beliefs. I assume that the agent forms beliefs based on multiple information sources, each revealing a linear combination of f. For this example, I take each source to be perfectly revealing. Concretely, consider one background source that reports the return of asset 1 and a second that reports the average market return. In matrix form, these sources are given by

$$M = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}, \quad \text{so that observed signals are } Mf.$$

Crucially, the agent may be misspecified about the sources' structure. Let N denote the matrix capturing the agent's perception. The agent dogmatically interprets observations through N and their inference exactly rationalizes the observed signals:

$$Mf = N\tilde{f} \quad \Longrightarrow \quad \tilde{f} = N^{-1}Mf. \tag{1}$$

Suppose now that a controlled source revealing a linear combination x of the returns is provided and that the agent correctly understands this source. The combined system of sources is

$$M^\frown x = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \\ x_1 & x_2 \end{bmatrix} \quad \text{and} \quad N^\frown x = \begin{bmatrix} N_{1,1} & N_{1,2} \\ N_{2,1} & N_{2,2} \\ x_1 & x_2 \end{bmatrix}.$$

Since there are three sources but only two unknowns, beliefs generally cannot rationalize all signals exactly. Consistent with the misspecified learning framework developed later, the

best possible rationalization is given by the least-squares projection

$$\tilde{f} = ((N^\frown x)^\top (N^\frown x))^{-1} (N^\frown x)^\top (M^\frown x) f. \tag{2}$$

Misinference substitution. To highlight the core mechanism, consider the misspecification where the agent believes that reported returns exceed those they can achieve, for instance due to perceived fees or limited trading ability; specifically $N = aM$ with $a > 1$. Without additional information, (1) implies $\tilde{f} = f/a$, so the agent is uniformly underconfident across both assets' returns.

Now consider providing an additional source about the first asset's return, given by $x = (1, 0)^\top$. Solving (2) yields

$$\tilde{f}_1 = \frac{f_1}{a} + f_1 \frac{a-1}{a(a^2+1)}, \quad \tilde{f}_2 = \frac{f_2}{a} - f_1 \frac{a-1}{a(a^2+1)}.$$

The additional correctly understood source about the first asset's return indeed improves inference about that return. However, the agent must jointly rationalize all signals through the misspecified model. Here, the return of asset 2 is inferred solely through the signal on the average market return and therefore $\tilde{f}_2$ will be such that it fully rationalizes this signal. This means that any increase in the inferred return of asset 1 is offset one-for-one by a decrease in the inferred return of asset 2. I call such indirect effects misinference substitution.

Unambiguous harm. In the above case, whether the additional source is harmful or beneficial depends on f. I now show that misinference substitution can strictly reduce utility. To do so, let

$$N = \begin{bmatrix} a & 0 \\ a/2 & 1/2 \end{bmatrix} \quad \text{where} \quad a > 1.$$

As before, the agent is misspecified about the reported first asset's return but has a correct perception about the second asset's return. Without an additional source, the agent underestimates the return of asset 1, $\tilde{f}_1 = f_1/a$, but correctly infers the return of asset 2.

Now consider providing additional information on an equally weighted portfolio, $x = c(1, 1)^\top$. As in the previous case, this source improves inference about the linear combination it provides; specifically, it improves inference about the sum of returns. However, this improvement is achieved through an unambiguously harmful change in both dimensions: the

return of asset 1 is underestimated even further, while the return of asset 2 is overestimated. Hence, the agent is strictly worse off when the additional source is provided[5]. Importantly, this harm arises even for arbitrarily small misspecifications (a arbitrarily close to 1) and persists regardless of how strong or precise the additional source is, which in this example corresponds to large values of c.

Harm for any additional source. As I formalize in Proposition 4, harm can occur for any structure of the additional source. Table 1 provides a set of background sources and corresponding perceptions that suffice to generate this result in this example[6]. In each case, there are two background sources among the return of a single asset, the average market return, and the return differential. The misspecification follows the structure introduced above, where the agent believes that reported returns of one asset exceed those they can achieve. For additional sources that convey information about both assets, the effect on

Table 1: Sources and background environments under which they cause harm

Add. Source	**True Structure**	**Agent's Perception**	**Effect on Inference**
$\text{sgn}(x_1) = \text{sgn}(x_2)$ $\lvert x_1\rvert \geq \lvert x_2\rvert > 0$	$\begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}$	$\begin{bmatrix} a & 0 \\ a/2 & 1/2 \end{bmatrix}$	Inference of all f_i worsens
$\text{sgn}(x_1) = \text{sgn}(x_2)$ $\lvert x_2\rvert \geq \lvert x_1\rvert > 0$	$\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}$	$\begin{bmatrix} 0 & a \\ 1/2 & a/2 \end{bmatrix}$	Inference of all f_i worsens
$\text{sgn}(x_1) \neq \text{sgn}(x_2)$ $\lvert x_1\rvert \geq \lvert x_2\rvert > 0$	$\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$	$\begin{bmatrix} a & 0 \\ a & -1 \end{bmatrix}$	Inference of all f_i worsens
$\text{sgn}(x_1) \neq \text{sgn}(x_2)$ $\lvert x_2\rvert \geq \lvert x_1\rvert > 0$	$\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & a \\ -1 & a \end{bmatrix}$	Inference of all f_i worsens
$x_1 \neq 0$ $x_2 = 0$	$\begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}$	$\begin{bmatrix} a & 0 \\ b/2 & 1/2 \end{bmatrix}$	Inference of f_1 improves but of f_2 worsens
$x_2 \neq 0$ $x_1 = 0$	$\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}$	$\begin{bmatrix} 0 & a \\ 1/2 & b/2 \end{bmatrix}$	Inference of f_2 improves but of f_1 worsens

[5]More precisely, the agent's utility is strictly lower with the additional information if and only if $f_1 \neq 0$. In the degenerate case $f_1 = 0$, there is no misinference and utility is unchanged.

[6]The table partitions the set $\mathbb{R}^2 \setminus \{0\}$ into six cases and specifies, for each case, a pair (M, N) such that the agent is strictly worse off for almost all f (and weakly worse off for all f) when provided with any additional source from the corresponding partition. The case $x = 0$ is excluded, as such a source conveys no information (its signal is identically zero independently of f) and therefore has no effect on inference.

inference is as described above. While the additional source improves inference about the targeted linear combination, beliefs about both individual returns move further away from the truth, so utility is strictly lower. For sources targeting a single asset, inference about that asset improves. However, misinference substitution can be sufficiently strong that the induced distortion in the untargeted asset dominates, leading to an overall welfare loss. This arises, for example, when the agent's misperception differs across background sources: reported returns in the single-asset source are scaled by a, while those in the market-return source are scaled by b. When the distortion in the aggregate source is sufficiently larger, $\frac{b-a}{b} > a - 1$, harm is unambiguous.

These cases illustrate that robustly non-harmful information does not exist even under a very limited ambiguity set to which robustness is required. The background sources are simple and economically natural and the misspecification arises from a single reasonable bias that can be arbitrarily small.

3 A Multidimensional Misspecified Learning Model

I now present the general framework in which I then conceptualize background and controlled sources. I analyze the resulting learning outcomes and conclude the section by showing how this general setup encompasses several concrete learning biases.

3.1 Framework

A single Bayesian agent aims to learn simultaneously about multiple variables of interest, represented by a vector of fundamentals $f \in \mathbb{R}^F$. Starting from a full-support prior, the agent observes a data-generating process yielding signals $s_t \in \mathbb{R}^S$ in discrete time $t \in \{1, 2, \dots\}$ over an infinite horizon. Each signal consists of linear combinations of the fundamentals, with added Gaussian noise that may be correlated across dimensions. Formally,

$$s_t = Mf + \epsilon_t \text{ where } M \in \text{Mat}_{S,F}(\mathbb{R}) \text{ and } \epsilon_t \sim \mathcal{N}(0, \Omega) \text{ i.i.d. over time.}$$

Crucially, the agent's perception of the data-generating process may be flawed. They believe that

$$s_t = Nf + \epsilon_t \text{ where } N \in \text{Mat}_{S,F}(\mathbb{R}) \text{ and } \epsilon_t \sim \mathcal{N}(0, \Sigma) \text{ i.i.d. over time.}$$

This misspecification is dogmatic; the perception, given by N and Σ, is persistent and does not get updated while learning. I assume $\text{rank}(M) = \text{rank}(N) = F$ so that both the true and perceived models are sufficiently rich to be fully identified. The agent is correctly specified if $M = N$ and $\Omega = \Sigma$, and misspecified otherwise. Given these assumptions, a correctly specified agent learns the correct fundamentals.

Modeling background and controlled sources. I model the background sources as described above and assume that the controlled information source has an analogous structure. It provides additional, periodic signals about a fixed linear combination with added i.i.d. Gaussian noise that are observed alongside background signals. The receiver holds a dogmatic (mis)perception of both the linear combination the source reflects and its noise level. I assume that the controlled source's signals are uncorrelated with background signals and normalize both the true and perceived noise variance to one.[7] Using this normalization, the precision of additional information is governed by the absolute values of the weights and scaling a source by a constant with absolute value greater (smaller) one means increasing (decreasing) the precision of the source. Overall, the additional controlled information source is the pair (x, y), where $x \in \mathbb{R}^F$ denotes a vector of true weights and $y \in \mathbb{R}^F$ of perceived weights. If $x = y$, I say that the additional information is correctly perceived and, with slight abuse of terminology, refer to x as the controlled source. Fixing background sources determined by matrices M, Ω, N, and Σ, the learning environment including an additional source retains the same structure, with extended matrices:

$$M^\frown x \coloneqq \begin{bmatrix} M \\ x^\top \end{bmatrix}, \ \Omega' \coloneqq \begin{bmatrix} \Omega & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}, \ N^\frown y \coloneqq \begin{bmatrix} N \\ y^\top \end{bmatrix}, \text{and } \Sigma' \coloneqq \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}.$$

Unless otherwise stated, I assume $y \neq 0$. If $y = 0$, the receiver treats the additional information as pure noise, resulting in a null effect.

[7] Fixing the variance is without loss of generality, since a different variance is equivalent to multiplying the weights by an appropriate scalar. Assuming that the additional source is uncorrelated is with loss of generality but necessary for this paper's tools.

3.2 Learning Outcome

First, I show that the agent's beliefs converge and characterize the form of the limiting belief. The agent's beliefs $\pi_0, \pi_1, \ldots$ are said to *concentrate* on a set A if, for every open set U containing A the subjective probability that the fundamentals lie in U, $P_{\pi_t}(f \in U)$, almost surely converges to 1; that is,

$$\mathbb{P}[\lim_{t\to\infty} P_{\pi_t}(f \in U) = 1] = 1.$$

Utilizing a classical result by Berk (1966), I show that beliefs concentrate on the set of points minimizing the Kullback-Leibler divergence. Then, using perceived linearity and normality, I prove that there is a unique minimizer. As a result, the beliefs concentrate on a single point, which I refer to as the *inferred fundamentals*. These inferred fundamentals are the object of interest in the remainder of the analysis.

Proposition 1. *Given any true fundamentals f and matrices M, Ω, N, and Σ, the beliefs concentrate on*

$$\tilde{f} = \left(N^\top \Sigma^{-1} N\right)^{-1} \left(N^\top \Sigma^{-1} M\right) f. \tag{3}$$

The inferred fundamentals are those that best rationalize the true signal mean given the agent's specification. When M and N are square matrices, a perfect replication is possible and (3) reduces to $N\tilde{f} = Mf$. Generally, when the signals' dimension exceeds the number of fundamentals, a perfect explanation may not exist. Inferred fundamentals instead minimize the distance between the predicted signal mean and the true one, where the distance is taken with respect to the agent's model. Analogous to the geometry of an OLS estimator, the remaining difference is orthogonal to the agent's model[8].

Because inferred fundamentals are determined by the signals' mean and the agent's model, the true signals' covariance Ω has no effect on long-run inference. Hence, moving forward I omit it when referring to background sources.

In the Appendix, I provide general properties of the inferred fundamentals that serve as technical tools for the proofs of later results.

[8]Indeed, the inferred fundamentals are equal to the GLS estimator with infinite data. Under the perception of the agent the GLS estimator is the maximum likelihood estimator which converges precisely to the Kullback-Leibler minimizer. Therefore, the outcome of such a GLS estimator coincides with the inferred fundamentals.

3.3 Misperceptions Captured by the Model

Since the main question concerns robustness to misspecification in general, the framework deliberately does not impose structure on fundamentals, information sources, or specific biases. Nevertheless, I illustrate how familiar biases can be represented within this framework.

Over- and underconfidence. Misperceptions about one's own ability can be incorporated in two ways. First, ability may directly affect how fundamentals map into signals. For instance, the impact of a fundamental such as a production technology may depend on the agent's skill. Over- or underconfidence is then captured by the corresponding entries being higher (or lower) in N than in M. Alternatively, ability can enter as a fundamental that is dogmatically misperceived, as in Heidhues et al. (2026). Suppose there is one signal dimension equal to c times the unit vector corresponding to ability, while its perception is ac times that unit vector. As $c \to \infty$, the agent becomes arbitrarily confident that their ability is exactly a^{-1} times its true value. Dogmatic misperceptions about the agent's ability can thus be arbitrarily well approximated within the model[9].

Attribution bias. Attribution bias refers to systematic errors in assessing the relative importance of underlying causes of observed outcomes (see, e.g., Ross, 1977, Gilbert and Malone, 1995, for a review). Consider an agent observing project outputs, where outcomes depend on worker ability and project merit. Attribution bias arises when the agent's perception assigns incorrect relative weights to the impact of these causes. Formally, N is obtained from M by scaling the columns corresponding to ability and project merit by the agent's perceived relative weight.

Projection bias. Projection bias describes the tendency to project features of one's own environment onto less familiar ones (see, e.g., Engelmann and Strobel, 2000, Van Boven et al., 2003, Engelmann and Strobel, 2012, Ambuehl et al., 2021, Bushong and Gagnon-Bartsch, 2023). Suppose one signal dimension corresponds to a familiar environment while others reflect unfamiliar settings. For the familiar dimension, the agent's perception is accurate

[9]Indeed, the general model of Heidhues et al. (2026) can be approximated and nested in this way. If this is the only source of misspecification (i.e., $M = N$ otherwise), their approach is of course more direct.

and the corresponding row of N coincides with that of M. For unfamiliar dimensions, projection arises: the rows of N are convex combinations of the true rows of M and the row associated with the familiar environment.

Sparsity and correlation neglect. Disentangling which fundamentals drive which signals is cognitively demanding and agents may simplify by ignoring sufficiently weak relationships. This is closely related to sparsity (Gabaix, 2014)[10] and aligns with experimental evidence on correlation neglect (Enke and Zimmermann, 2019)[11]. Concretely, an agent may behave as if each signal dimension depends only on a subset of the most influential fundamentals. In the model, this corresponds to N being a sparse version of M, obtained by setting the smallest entries of rows to zero.

4 The Effect of an Additional Source

Proposition 1 applies to the learning environment both with and without the additional source and thus it yields closed-form solutions for the inferred fundamentals in both cases. In this section, I derive a tractable expression for the difference between these inference outcomes and thus for the effect of an additional source. Moreover, I formalize the observation from the leading example that an additional source always improves inference along the dimension it targets. All proofs are purely algebraic.

Consider a controlled source (x, y) and background sources given by M, N, and Σ. I denote the inferred fundamentals incorporating the controlled source by $\tilde{f}(x, y)$, writing $\tilde{f}(x)$ when the source is correctly perceived ($x = y$). The inferred fundamentals under only the background sources are denoted $\tilde{f}(0)$. It follows from the analysis below that this notation is consistent.

[10] Gabaix (2014) explores sparsity in decision-making processes by assuming that agents have a restricted number of components they can incorporate into the maximization. Adapted to this learning framework, instead of presuming that relevant variables for decision-making have to be sparse, one could assume the set of relevant fundamentals behind each observation is sparse.

[11] Enke and Zimmermann (2019) run an experiment in which participants received a multidimensional signal influenced by multiple fundamentals but behaved as though each component was influenced by only one. While their objective was to learn about the average of the fundamentals rather than the individual components, their findings plausibly extend to this framework.

Proposition 2. *Let f be a vector of fundamentals, M, N, and Σ define background sources, and (x, y) an additional source. Then,*

$$\tilde{f}(x,y) - \tilde{f}(0) = \frac{1}{1+g}\left[(N^\top\Sigma^{-1}N)^{-1}yy^\top(f - \tilde{f}(0)) + (N^\top\Sigma^{-1}N)^{-1}y(x^\top - y^\top)f\right] \quad (4)$$

with $g = y^\top(N^\top\Sigma^{-1}N)^{-1}y > 0$.

For correctly perceived additional information (4) simplifies to

$$\tilde{f}(x) - \tilde{f}(0) = E(f - \tilde{f}(0)) \text{ where } E = \frac{1}{1+g}(N^\top\Sigma^{-1}N)^{-1}xx^\top.$$

The effect of the controlled source decomposes into the matrix E and the misinference absent the source, $f - \tilde{f}(0)$, upon which E acts. To gain intuition, consider information about the first fundamental, $x = e_1$. Then E has nonzero entries only in its first column, meaning the effect is driven by $f_1 - \tilde{f}(0)_1$. If there is no misinference of the first fundamental absent the additional source, the same belief fully rationalizes the additional information and there is no effect. When misinference is present, then the additional source reduces it; $E_{1,1}$ lies in $(0, 1)$. Crucially, the remaining entries in the first column of E translate the misinference of the first fundamental onto other dimensions, exactly as highlighted in the leading example — this can result in misinference substitution[12]. Importantly, E is determined entirely by the additional source and the receiver's model. The true structure of background sources affects only the misinference absent the additional source, $f - \tilde{f}(0)$.

While the overall effect generally depends on the entire structure, a robust feature of additional information is that it always improves inference along the dimension it targets. Moreover, a more precise source will improve that dimension strictly more and an infinitely precise source completely eliminates any misinference in the targeted dimension.

Proposition 3. *Given any additional source x and background sources M, N, and Σ there exists some $\alpha \in [0, 1]$ such that $x^\top\tilde{f}(x) = \alpha x^\top\tilde{f}(0) + (1-\alpha)x^\top f$ for all f. If $x^\top\tilde{f}(0) \neq x^\top f$, then $\alpha \in (0, 1)$ is unique and scaling x by a constant c strictly decreases (increases) α if $|c| > 1$ ($|c| < 1$). Moreover, when scaling x by $c \to \infty$ α tends to 0, i.e. $x^\top\tilde{f}(x) \to x^\top f$.*

[12]Heidhues et al. (2026) also observe that changes in a multidimensional learning environment that reduce the difference between inference and truth in one dimension can worsen it in other dimensions; they call this bias substitution in their application. Importantly, Heidhues et al. (2026) focus on particular changes of the environment driven by their application and bias substitution occur in them as one interesting force. Here, on the other hand, misinference substitution is at the core of the results and I specifically focus on its prevalence and magnitude.

5 Robustness with Respect to Background Sources

I now turn to the main results of the paper and study whether an additional information source can be guaranteed to be non-harmful, irrespective of background sources and their misspecification.

In this section, I focus on the key case of a correctly perceived additional source and define the value of a controlled source x as the reduction in the Euclidean distance between inferred fundamentals and the truth:

$$V(x \mid M, N, \Sigma, f) := ||\tilde{f}(0) - f||_2 - ||\tilde{f}(x) - f||_2. \tag{5}$$

Thus, an additional source has positive value if and only if it generates beliefs closer to the truth. This is consistent with the agent's objective in the leading example, where utility is decreasing in the Euclidean distance between beliefs and fundamentals. I consider misperceived additional sources and discuss alternative value functions in Section 6.

Note that any additional information has zero value for a correctly specified receiver in the limit of infinite data. Such a receiver perfectly infers the fundamentals and this remains unchanged when a correctly perceived information source is added. However, this is a property of the limit: for any fixed finite number of observations, additional information would have strictly positive value, since the information structure defined by (M, Ω) is strictly Blackwell dominated by the information structure $(M^\frown x, \Omega')$ resulting from adding the source x. Taking seriously that the receiver is fully convinced that their perception is correct, they therefore weakly prefer any additional information and strictly prefer it in finite settings. While information can in fact be harmful, as shown below, the receiver does not anticipate that this applies to them.

5.1 The Harm of Information

The following result characterizes when information is robustly beneficial with respect to learning from background sources in a potentially misspecified manner.

Proposition 4. ***I*** *If $F = 1$, then any additional source x has non-negative value for all background sources M, N, and Σ and fundamentals f. Moreover, it has positive value*

if and only if $f \neq \tilde{f}(0)$*, in which case the value increases strictly in* $|x|$*.*

II.A *If* $F \geq 2$*, then for any additional source* x *there exist background sources given by* M*,* N*, and* Σ *such that* $V(x \mid M, N, \Sigma, f) < 0$ *for almost all* f*.*

II.B *Assume* $F \geq 2$ *and fix any constant* $K \in \mathbb{R}_+$ *and any* $\delta > 0$*. For any additional source* x *there exist background sources* M*,* N*, and* Σ *such that* $V(x \mid M, N, \Sigma, f) < -K$ *for all* f *where* $|x^\top f| > \delta$*.*

Part I shows that in one-dimensional settings the classical result, that information is always weakly beneficial, extends to receivers that also learn from misspecified background sources. Furthermore, if there is misinference absent an additional source, then information is strictly beneficial and more information is strictly better. Part II.A establishes an impossibility when there are multiple dimensions: for any additional source, there exist background sources and a misperception under which the value is negative. Part II.B says that the magnitude of harm can be arbitrarily large[13].

Comparing part I with part II highlights that multidimensionality is essential for information to be harmful. As shown in Proposition 3, a controlled source always improves inference in the dimension it targets and if there is only a single one, additional information thus always (weakly) improves inference. When there are multiple dimensions, misinference substitution can occur for any structure of the controlled source. The rate of misinference substitution can be sufficiently large to outweigh the beneficial direct effect and induce even arbitrarily large harm. Importantly, this difference between a multi- and one-dimensional environment does not depend on restricting the controlled source to a single dimension. As a preview to Section 6, even an additional information source of the same dimension as the fundamentals is not necessarily robustly non-harmful.

An information provider's problem is straightforward in a one-dimensional environment. Providing the most precise information possible is always optimal, as it is never harmful

[13]The condition $|x^\top f| > \delta$ is necessary, since the effect of the additional source is scaled by its direct effect $x^\top f$. While the mass of f where this condition does not hold is positive, the x in part II.B has strictly negative value for almost all f. Moreover, in the measure-zero set where the value is not strictly negative, the value of the additional source is 0.

and offers the greatest potential benefit. In a multidimensional environment there is also a clear, albeit negative result: no information source can be designed that is guaranteed to be (weakly) beneficial across misperceived background sources.

The contrast to the classical benchmark is sharp: under correct perception, all information is robustly non-harmful, whereas when misperception is possible, none is robustly non-harmful. This is substantially stronger than the minimal failure of the classical benchmark which, within the setting considered here, would only require the existence of a single configuration of x, M, N, Σ, f that generates harm. Moreover, harm for any source does occur under a fixed and standard value function, and can arise for almost all fundamentals at once.

5.2 Knowledge and Robustly Beneficial Information

The previous subsection establishes an impossibility for unrestricted background sources and their misspecification. Additional knowledge about the environment allows an information provider to restrict the set to which robustness is required and may therefore permit robustly beneficial design. I now characterize what knowledge is sufficient and what is not.

Knowledge of the true structure. Possessing full knowledge of the true structure M of background sources is insufficient to design information that is robustly beneficial: fixing a true structure, there is no information that is robustly beneficial with respect to all possible misperceptions.

Proposition 5. *Let $F \geq 2$ and fix the true structure of background sources M. For any controlled source x there exists a perception N and a Σ such that $V(x \mid M, N, \Sigma, f) < 0$ for almost all f.*

The intuition follows from Proposition 2. Recall that $\tilde{f}(x) - \tilde{f}(0) = E(f - \tilde{f}(0))$, where the true structure M affects only the misinference absent the additional source, $f - \tilde{f}(0)$. In particular, knowledge about M places no restrictions on the possible matrices E. Since E governs how the additional source acts on that misinference — and in particular whether and how misinference substitution occurs — knowledge of M alone provides no control over

the mechanism that can cause harm. The role an additional source plays in the receiver's learning process is fully determined by its interaction with the receiver's perception, and complete knowledge of the true data-generating process therefore provides no opportunity for robust information design.

Knowledge of the receiver's perception. Since full knowledge of the true structure is not sufficient, some restriction on the receiver's perception is necessary for designing robustly non-harmful information. Moreover, perfect knowledge of the receiver's perception alone is also sufficient:

Proposition 6. *Given any perception of background sources N and Σ, there exists an additional source x such that $V(x \mid M, N, \Sigma, f) \geq 0$ for all true structures M and f.*

The key observations again stem from Proposition 2. For a fixed perception, no restrictions on M and f are equivalent to no restriction on the misinference absent the additional source, $f - \tilde{f}(0)$. However, knowledge of N and Σ allows the provider to control E by designing x. Concretely, one can choose x to be a basis vector in the orthonormal basis that diagonalizes $(N^\top \Sigma^{-1} N)^{-1}$. This ensures that E does not affect any other dimension — there is no misinference substitution in this orthonormal basis. As in Proposition 3, inference along the targeted dimension weakly improves, guaranteeing weakly positive value. The resulting source is generically strictly beneficial: its value is zero only if inference is already correct along the targeted dimension, which requires M to satisfy a knife-edge condition (e.g., $M = N$). Hence, for most M the value is strictly positive for almost all f. Moreover, as in Proposition 3, the value strictly increases with precision whenever improvement is possible.

Intuitively, knowledge of the receiver's perception reveals the dimensions along which the receiver learns independently. Providing information that targets exactly one such dimension preserves this independence and eliminates any spillover — misinference substitution — while strictly improving inference along the targeted dimension.

This logic extends beyond full knowledge and yields a full characterization of when partial knowledge of the receiver's perception suffices.

Proposition 7. *Let $F \geq 2$ and fix $\Sigma = I_S$ without loss of generality. For any subset $\mathcal{A}$ of rank-F matrices N, there exists an additional source x such that $V(x \mid M, N, \Sigma, f) \geq 0$ for all $N \in \mathcal{A}$, all M, and all f if and only if there exists a vector v that is a common eigenvector of $(N^\top N)^{-1}$ for all $N \in \mathcal{A}$. If no such vector exists, then for any x there exist $N \in \mathcal{A}$ and M such that $V(x \mid M, N, \Sigma, f) < 0$ for almost all f.*

If such a vector v exists, choosing $x = cv$ for some $c \neq 0$ yields a source that affects only that common eigendirection and the argument of Proposition 6 applies. If no such direction exists, then for any x one can find a perception under which the induced matrix E affects dimensions besides the targeted one: there is a perception where this source can generate misinference substitution. I show by construction that a corresponding M exists such that this yields negative value.

The condition in Proposition 7 can for example be fulfilled in the case of sparse perceptions discussed in Section 3.3. When perception is maximally sparse and each background source is perceived to depend on a single fundamental, the corresponding matrices share a common eigenbasis. Even without knowing which fundamental corresponds to which background source or the true structure of sources, providing information about any single fundamental is then robustly non-harmful.

Full knowledge. When both the true structure and the perception of background sources are known, a robustly non-harmful source can be constructed as above. In addition, knowledge of M allows the provider to identify dimensions along which misinference is generically present absent any additional source. Providing information along such a dimension then guarantees strictly positive value for almost all f.

Combined partial knowledge. What remains is combined partial knowledge of the true structure and the perception. Analyzing such restrictions in generality appears to be intractable[14]. While such structures may be natural in specific applications, a case-by-case

[14]When the set of perceptions does not feature the necessary property of Proposition 7, there is no general property of the matrix E that would prevent harm. At the same time, combined knowledge of M and the perception limits $f - \tilde{f}(0)$, so there is no full freedom here either, and the construction yielding the negative part of Proposition 7 does not apply. The combined restrictions on E and $f - \tilde{f}(0)$ can be sufficient to restore robustness in some cases. However, there seem to be no tractable, general conditions that do so.

analysis is outside the scope of this paper.

6 Robustness of the Non-Robustness

Unlike positive results on robustness, the impossibilities established in Section 5 continue to hold for any larger set of environments to which robustness is required. Conversely, imposing additional structure restricts this set and makes the results more demanding to establish. For example, imposing a linear-Gaussian structure on background sources and their misspecification is a restriction that strengthens the results. It limits the class of environments under consideration and the impossibility holds even within this structured setting. In any setting where background sources can be more complex but the environments considered here lie within the set a provider seeks robustness to, the impossibility extends immediately.

In the analysis below I consider three classes of variations. First, I examine additional restrictions on background sources and their misspecification, which shrink the set of environments and thereby strengthen the impossibility. Second, I consider alternative structures of the additional information itself, which expand the provider's options and thereby strengthen the impossibility results. Third, I explore alternative specifications of the value function and the learning environment, which neither shrink nor enlarge the set but change its character and therefore also explore when the main results extend.

6.1 Additional Restrictions on Misspecification.

Proposition 5 shows that even the strongest possible restriction on the true structure is insufficient to obtain robustly non-harmful information. It is therefore natural to ask whether restricting the set of admissible misspecifications can restore robustness. Such restrictions are equivalent to assuming the provider has additional knowledge and the discussion below thus complements that in Subsection 5.2.

Arbitrarily small misperception. Consider first misspecifications that are arbitrarily close to the truth. Even when combined with full knowledge of the true structure, this

restriction is insufficient to guarantee non-harmfulness.

The intuition is that proximity of N to M limits the magnitude of baseline misinference $f - \tilde{f}(0)$, but does not restrict the structure of the matrix E governing how the additional source acts on that misinference. In particular, misinference substitution can still occur and overturn the direct beneficial effect[15].

Formally, part II.A of Proposition 4 and Proposition 5 can be strengthened as follows:

Proposition 8. *Let $F \geq 2$ and fix any true structure M. For any x there exists a fixed Σ and a sequence of $N_k \to M$ such that $V(x \mid M, N_k, \Sigma, f) < 0$ for almost all f.*

This result not only shows the robustness of the impossibility for misspecified receivers but also highlights the fragility of the classical benchmark. While any information is robustly non-harmful under exactly correct perception, allowing for even arbitrarily small misspecification reverses this.

Restrictions on covariance misperception. I have allowed the receiver's perception of the covariance matrix, Σ, to vary freely. This is natural, as misspecification of the mean structure is naturally accompanied by misspecification of the covariance structure. However, the above results do not rely on variation in Σ.

In particular, the proofs in Section 5 and Proposition 8 specify $\Sigma = I_S$. Thus, restricting Σ to a fixed matrix, which is arguably the most natural choice, does not affect results.

More generally, when M is unrestricted, the assumption $\Sigma = I_S$ is without loss. As seen from (3), inference is invariant to the transformation that multiplies M and N by $\Sigma^{-1/2}$ and replaces Σ with the identity. Moreover, since the true covariance Ω plays no role in inference, one can impose $\Sigma = \Omega$ and assume correct perception of the covariance[16].

[15] It is worth noting that when the misspecification is small, inference is nearly correct. Since the effect of an additional source operates on the baseline misinference, the resulting distortion is also small. Formally, fixing M, x, and Σ, as $N \to M$ we have $\tilde{f}(0) \to f$ and $V(x \mid M, N, \Sigma, f) \to 0$, though the limit may be approached from below.

[16] While this normalization is without loss for the formal statements, it affects interpretation. If M is known, then the mean of the true distribution is known and since Ω does not affect inference, the entire distribution is effectively known. Imposing $\Sigma = \Omega$ then links knowledge of the true distribution to knowledge of the agent's perception, which complicates the separation between the two.

6.2 Alternative Structures of Additional Information

The analysis so far has focused on a single, correctly perceived additional source. I now show that allowing the additional source to be misperceived or multidimensional does not necessarily restore robustness.

Misspecified additional sources. Suppose the provider cannot control how the receiver perceives the additional source. In this case, the potential for harm is weakly greater, since correct perception is a special case. The impossibility of robustly non-harmful information therefore extends immediately.

A different possibility is that the provider can choose the perceived structure of the additional source independently of its true structure. Even in this case, robustness fails. If the receiver's perception of background sources is correct, then inference absent the additional source is already perfect. Any misperceived additional source renders the receiver's overall model misspecified and thereby generically induces misinference. Consequently, such a source cannot be robustly non-harmful.

Multidimensional additional sources. Allowing to provide multiple sources does not in general restore robustness, because the negative effects induced by one component of a multidimensional source need not be offset by other components. For instance, even an additional source corresponding to the matrix I_F can be harmful under suitable background sources and their misspecification. Note that this reflects that the effect of additional sources depends on their perceived precision relative to the background sources. For any fixed background sources, sufficiently precise signals of this form lead inference to approach the truth. However, what constitutes "sufficiently precise" depends on the receiver's perception, and without such knowledge a robust design remains impossible.[17]

On the positive side, when the receiver's perception is known, multidimensional sources can be used to improve inference in all relevant directions. In particular, applying the construction underlying Proposition 6 to each independent dimension of the perceived model

[17]While Proposition 2 could in principle be applied iteratively to characterize such settings, doing so leads to substantial tractability costs without altering the core insight: expanding the structure of the signal does not eliminate the possibility of harm.

yields a collection of sources that eliminate distortions across dimensions. As their precision increases, inferred fundamentals converge to the truth.

6.3 Different Specifications

I now examine the robustness of the main results under alternative modeling assumptions. Concretely, I consider alternative fixed and thus implicitly known value functions, different data-generating processes, and approximations with finite signals.

Alternative measures of value. Additional sources affect the receiver's utility only through their influence on actions, which in turn depend on beliefs. The leading example illustrates that the effect on utility can reduce to comparing the distance of beliefs from the truth. Generally, it seems intuitive that an agent with beliefs closer to the truth will make better decisions and thus achieve higher utility. The impossibility results do not depend on the concrete notion of closeness of beliefs, i.e., which norm on $\mathbb{R}^F$ is used:

Proposition 9. *Let $F \geq 2$ and fix any norm $||\cdot||$. For any true structure M and any controlled source x, there exist a perception N and Σ such that*

$$V_{||\cdot||}(x \mid M, N, \Sigma, f) \coloneqq ||\tilde{f}(0) - f|| - ||\tilde{f}(x) - f|| < 0 \quad \textit{for almost all } f.$$

There are, however, decision environments in which robustly beneficial information can be provided. For instance, when the value of beliefs depends only on a one-dimensional summary statistic, i.e., when the value function takes the form $|v^\top(\cdot)|$ for some fixed $v \neq 0$. In this case, the objective is effectively one-dimensional and even if the learning environment is not, the harmful effects of misinference substitution can be avoided. It follows directly from Proposition 3 that for any $c \neq 0$ the source $x = cv$ is robustly non-harmful: whenever inference along v can be improved, the source has strictly positive value, increasing in $|c|$.

There are, of course, several decision problems where the value function does not fall under either of the above cases. For example, decision problems with discrete actions induce discontinuities and thus do not reduce to any case discussed. Generally, it is necessary that actions depend in a multidimensional sense on the beliefs and if they do so misinference substitution can result in several trade-offs that seem intuitively harmful. Formally, however, a

more general analysis becomes less tractable and is beyond the scope of this paper.

Finally, Proposition 6 exploits that the Euclidean norm is invariant under orthonormal transformations. The construction operates in a basis that diagonalizes $(N^\top \Sigma^{-1} N)^{-1}$ and thus isolates independent learning dimensions. While providing information that targets such a dimension still preserves this independence, the value function might not treat beliefs in these dimensions as independently valuable. In specific instances, for example when an independent dimension is an individual fundamental, it might still be the case that knowledge about only the receiver's perception reveals a robustly beneficial source but the general sufficiency result may fail. Nevertheless, since Proposition 9 extends to all norms, some knowledge of the receiver's perception remains necessary regardless of the evaluation criterion. Understanding the receiver's perceived model therefore remains the more important input for robust information design.

Background source following other distributions. Throughout, I have assumed that background sources follow a linear-Gaussian structure. The linear mean is necessary for stating results tractably, and the Gaussian noise assumption ensures that implicit knowledge about the true and perceived structures coincides, simplifying the exposition. Both assumptions can be relaxed. I show in the Appendix that the convergence result of Proposition 1 extends to arbitrary data-generating processes with mean Mf, and consequently so do all main results. When the mean μ of the true background sources depends non-linearly on the fundamentals, the convergence result still holds though the relationship between signals and fundamentals becomes less transparent. Since there is no longer a matrix M that can be analyzed independently of f, the results extend but in appropriately adapted form.

Finite signals. The model concerns long-run inference from infinitely many signals. This modeling choice simplifies both the analysis and exposition, enabling a clearer description of the learning outcome. This limit is, by definition, approximated by settings with a finite but large number of signals. Since sources generate normal i.i.d. signals and the agent is Bayesian, their order is irrelevant and multiple signals of a given precision are ex-ante equivalent

to a single signal of proportionally higher precision.[18] The insights therefore extend to environments with finitely many but sufficiently informative signals, or to settings in which the controlled source is observed separately from background sources. The key requirement is that learning is sufficiently strong relative to the prior for the long-run dynamics to dominate.

7 Conclusion

This paper studies the challenges and possibilities of providing robustly non-harmful information to potentially biased learners. It does so in a multidimensional yet otherwise simple setting: receivers learn from misspecified background sources where their true and perceived structure and the provided information all take a particular, tractable form. As a first step toward analyzing robustness with respect to learning biases, focusing on such a structured environment is natural. At the same time, these restrictions strengthen the impossibility results by limiting the set of environments over which robustness is required. There remain, however, naturally interesting restricted environments and learning biases that exclude those analyzed here, for which conclusions may differ.

The model considers environments in which the receiver's observed signals are exogenous, that is, unaffected by their actions. The results therefore capture fundamental features of misspecified learning, isolated from added complexity arising from endogenous feedback. In many applications, however, actions influence the signals receivers observe and thereby alter the learning process. From a robustness perspective, the exogenous benchmark studied here may still be relevant as one of the environments a provider must account for, in which case the negative results apply. When the interaction between actions and signals is known or sufficiently constrained, outcomes could differ substantially. Relatedly, the provision of additional information may itself alter how the receiver uses background sources. This can occur, for instance, if a controlled source replaces an existing background source rather than simply adding to it.

Even when background sources are fixed, providers may endogenously adapt the con-

[18] Formally, fixing M and N and scaling Ω and Σ down by an appropriate factor is ex-ante equivalent to modeling repeated signals.

trolled source in response to observed beliefs or actions. Intuitively, such dynamic information provision could allow the source to depend indirectly on the receiver's perception. This raises the possibility that dynamic information provision could be designed to be, in the long run, robustly non-harmful even without additional knowledge of the receiver's mental model.

Finally, relaxing the assumption of dogmatic misperception presents a further avenue for research. Receivers may revise their perceived model in response to discrepancies between predictions and observed outcomes. The interaction between such model updating and additional information can crucially alter the value of information. Information that is harmful under a fixed misperception may exacerbate inconsistencies and induce the receiver to adopt a more accurate model, thereby becoming beneficial in the long run. Conversely, information that is beneficial under a fixed perception may reinforce an incorrect model that would otherwise be abandoned. More broadly, when receivers entertain multiple plausible models, it becomes natural to study the value of providing information about their perception itself, rather than only about underlying fundamentals.

References

Ambuehl, Sandro, B Douglas Bernheim, and Axel Ockenfels, "What motivates paternalism? An experimental study," *American Economic Review*, 2021, *111* (3), 787–830.

Azrieli, Yaron, "Comment on "the law of large demand for information"," *Econometrica*, 2014, *82* (1), 415–423.

Berk, Robert H, "Limiting behavior of posterior distributions when the model is incorrect," *The Annals of Mathematical Statistics*, 1966, *37* (1), 51–58.

Blackwell, David, "Comparison of experiments," in "Proceedings of the second Berkeley symposium on mathematical statistics and probability," Vol. 2 University of California Press 1951, pp. 93–103.

_ , "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, 1953, pp. 265–272.

Bohren, J Aislinn, "Informational herding with model misspecification," *Journal of Economic Theory*, 2016, *163*, 222–247.

_ **and Daniel N Hauser**, "Learning with heterogeneous misspecified models: Characterization and robustness," *Econometrica*, 2021, *89* (6), 3025–3077.

_ **and** _ , *The Behavioral Foundations of Model Misspecification: A Decomposition*, Penn Institute for Economic Research, Department of Economics, University of . . . , 2023.

Bordoli, Davide, "Non-Bayesian updating and value of information," *Available at SSRN 5280727*, 2024.

Boven, Leaf Van, George Loewenstein, and David Dunning, "Mispredicting the endowment effect:: Underestimation of owners' selling prices by buyer's agents," *Journal of Economic Behavior & Organization*, 2003, *51* (3), 351–365.

Braghieri, Luca, "Biased Decoding and the Foundations of Communication," *Available at SSRN 4366492*, 2023.

Bushong, Benjamin and Tristan Gagnon-Bartsch, "Failures in Forecasting: An Experiment on Interpersonal Projection Bias," *Management Science, Forthcoming*, 2023.

Chauvin, Kyle, "A misattribution theory of discrimination," *Working Paper*, 2023.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, 2011, *108* (17), 6889–6892.

Engelmann, Dirk and Martin Strobel, "The false consensus effect disappears if representative information and monetary incentives are given," *Experimental Economics*, 2000, *3*, 241–260.

_ **and** _ , "Deconstruction and reconstruction of an anomaly," *Games and Economic Behavior*, 2012, *76* (2), 678–689.

Enke, Benjamin and Florian Zimmermann, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

Esponda, Ignacio and Demian Pouzo, "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 2016, *84* (3), 1093–1130.

_ **and** _ , "Equilibrium in misspecified Markov decision processes," *Theoretical Economics*, 2021, *16* (2), 717–757.

_ **,** _ **, and Yuichi Yamamoto**, "Asymptotic behavior of Bayesian learners with misspecified models," *Journal of Economic Theory*, 2021, *195*, 105260.

Eyster, Erik and Matthew Rabin, "Naive herding in rich-information settings," *American Economic Journal: Microeconomics*, 2010, *2* (4), 221–243.

Frick, Mira, Ryota Iijima, and Yuhta Ishii, "Misinterpreting others and the fragility of social learning," *Econometrica*, 2020, *88* (6), 2281–2328.

_ , _ , **and** _ , "Belief convergence under misspecified learning: A martingale approach," *The Review of Economic Studies*, 2023, *90* (2), 781–814.

_ , _ , **and** _ , "Welfare comparisons for biased learning," *American Economic Review*, 2024, *114* (6), 1612–1649.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack, "Limit points of endogenous misspecified learning," *Econometrica*, 2021, *89* (3), 1065–1098.

_ **, Gleb Romanyuk, and Philipp Strack**, "Active learning with a misspecified prior," *Theoretical Economics*, 2017, *12* (3), 1155–1189.

Gabaix, Xavier, "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

Gagnon-Bartsch, Tristan and Benjamin Bushong, "Learning with misattribution of reference dependence," *Journal of Economic Theory*, 2022, *203*, 105473.

Gilbert, Daniel T and Patrick S Malone, "The correspondence bias.," *Psychological bulletin*, 1995, *117* (1), 21.

Guthrie, Chris, Jeffrey J Rachlinski, and Andrew J Wistrich, "Blinking on the bench: How judges decide cases," *Cornell L. Rev.*, 2007, *93*, 1.

He, Junnan, Lin Hu, Matthew Kovach, and Anqi Li, "Learning Source Biases: Multisource Misspecifications and Their Impact on Predictions," 2023.

He, Kevin, "Mislearning from censored data: The gambler's fallacy and other correlational mistakes in optimal-stopping problems," *Theoretical Economics*, 2022, *17* (3), 1269–1312.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic expectations and misguided learning," *Econometrica*, 2018, *86* (4), 1159–1214.

_ , _ , **and** _ , "Convergence in models of misspecified learning," *Theoretical Economics*, 2021, *16* (1), 73–99.

_ , _ , **and** _ , "Misinterpreting yourself," *Available at SSRN 4325160*, 2023.

_ , _ , **and** _ , "Overconfidence and prejudice," *Review of economic studies*, 2026, *93* (2), 968–1000.

Hestermann, Nina and Yves Le Yaouanq, "Experimentation with self-serving attribution biases," *American Economic Journal: Microeconomics*, 2021, *13* (3), 198–237.

Malmendier, Ulrike and Geoffrey Tate, "CEO overconfidence and corporate investment," *The Journal of Finance*, 2005, *60* (6), 2661–2700.

Morris, Stephen and Hyun Song Shin, "The rationality and efficacy of decisions under uncertainty and the value of an experiment," *Economic Theory*, 1997, *9*, 309–324.

Morris, Stephen Edward, *The role of beliefs in economic theory*, Yale University, 1991.

Moscarini, Giuseppe and Lones Smith, "The law of large demand for information," *Econometrica*, 2002, *70* (6), 2351–2366.

Mu, Xiaosheng, Luciano Pomatto, Philipp Strack, and Omer Tamuz, "From Blackwell dominance in large samples to Rényi divergences and back again," *Econometrica*, 2021, *89* (1), 475–506.

Radner, Roy and Joseph Stiglitz, "A Nonconcavity in the Value of Information," *Bayesian models in economic theory*, 1984, *5*, 33–52.

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

Sherman, Jack, "Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix," *Annals of mathematical statistics*, 1949, *20* (4), 621.

Whitmeyer, Mark, "Bayes= Blackwell, Almost," *arXiv preprint arXiv:2302.13956*, 2023.

Appendix

A Additional Results

A.1 Additional General Properties of Inference

I record several properties of long-run inference that may be useful in other applications of the multidimensional misspecified learning model and that are used throughout the proofs.

First, it follows directly from Proposition 1 that inference does not depend on the labeling of signal components or fundamentals:

Observation 1. *1. Inference is invariant to permutations of signal components.*

2. Inference is equivariant under permutations of the fundamental dimensions.

Next, I consider properties that simplify the analysis of inference. If signal components are perceived as pure noise, they are irrelevant for inference. Formally, a signal component is treated as pure noise if the following condition holds.

Definition 1. The i-th signal component is *perceived irrelevant* if the i-th row of N is zero and $\Sigma_{ij} = \Sigma_{ji} = 0$ for all $j \neq i$.

Second, I characterize when inference about a subset of fundamentals can be analyzed independently. A sufficient condition is the following.

Definition 2. A subset $A \subseteq \{1, \ldots, F\}$ *does not interact with its complement* if, after a permutation of fundamentals that places A first and a reordering of signal components, there exist matrices M_1, M_2, N_1, N_2 and positive definite Σ_1, Σ_2 such that

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}, \quad N = \begin{bmatrix} N_1 & 0 \\ 0 & N_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}.$$

Third and finally, I show that if the impact of a fundamental is correctly specified, it does not generate misinference. However, misinference may still arise through interaction with misspecified components. The formal statements are collected below.

Lemma 1. *1. If the i-th signal component is perceived irrelevant, then $\left(N^\top \Sigma^{-1} N\right)^{-1} \left(N^\top \Sigma^{-1} M\right)$ and thus inference is invariant to its exclusion.*

2. If $A \subseteq \{1, \ldots, F\}$ does not interact with its complement, then $\left(N^\top \Sigma^{-1} N\right)^{-1} \left(N^\top \Sigma^{-1} M\right)$ has, up to permutation, a block-diagonal structure with blocks $\left(N_i^\top \Sigma_i^{-1} N_i\right)^{-1} \left(N_i^\top \Sigma_i^{-1} M_i\right)$ for $i \in \{1, 2\}$ and matrices as in the definition.

3. If $N_{*k} = M_{*k}$ for some $k \in \{1, \ldots, F\}$, then the k-th column of $\left(N^\top \Sigma^{-1} N\right)^{-1} \left(N^\top \Sigma^{-1} M\right)$ equals the k-th canonical basis vector.

A.2 General Result on Convergence

I now generalize the convergence result stated in Proposition 1 to any true signal-generating processes ρ with mean μ and finite covariance. Importantly, the agent's perception remains linear Gaussian as in the main text as this assumption is key for the proof.

Theorem 1. *Given any process ρ with mean μ and matrices N and Σ, the beliefs concentrate on*

$$\tilde{f} = \left(N^\top \Sigma^{-1} N\right)^{-1} \left(N^\top \Sigma^{-1} \mu\right).$$

Beyond relaxing assumptions on the true data-generating process, this result highlights that convergence is driven by the agent's perceived model, not the true one. In extreme cases, the true distribution may contain no information about the fundamentals. Yet, if the agent believes that signals are informative, they update accordingly.

A.3 Supporting Results and Notation

Finally, I introduce some useful notation and supporting results specific to the application of the model to the analysis of additional sources' value.

Notation. Let $S_1 \coloneqq \left(N^\top \Sigma^{-1} N\right)^{-1} \left(N^\top \Sigma^{-1} M\right)$ denote the matrix governing inference without an additional source, such that $\tilde{f}(0) = S_1 f$. For inference with a correctly perceived additional source x, let S_2 denote the corresponding matrix. Then, by Proposition 2,

$I - S_2 = (I - E)(I - S_1)$ where $E = \frac{1}{1+x^\top (N^\top \Sigma^{-1} N)^{-1} x}(N^\top \Sigma^{-1} N)^{-1} x x^\top$. The value of an additional source is $||(I - S_1)f||_2 - ||(I - S_2)f||_2$ of which I make use throughout the proofs.

A useful normalization follows from the fact that only relative precision matters. A uniform rescaling of background precision leaves inference without an additional source unchanged and is equivalent to rescaling the precision of the additional source.

Lemma 2. *For any $c > 0$, inference with versus without an additional source x under background precision Σ coincides with inference under background precision $c\Sigma$ and additional source $\sqrt{c}\, x$.*

Finally, when evaluating the value of an additional source under the Euclidean norm, inference can be analyzed in any orthogonal basis. That is, the value of a source is invariant under joint unitary transformations of the signal and fundamental spaces.

Lemma 3. *Let U and V be unitary matrices and define $M' \coloneqq UMV^\top$, $N' \coloneqq UNV^\top$, $\Sigma' \coloneqq U\Sigma U^\top$, and $z \coloneqq V^\top x$. Then*

$$V(x \mid M, N, \Sigma, f) = V(z \mid M', N', \Sigma', V^\top f)$$

for all f.

B Proofs

Proof of Proposition 1. This follows directly from Theorem 1. □

Proof of Proposition 2. I prove the equivalent statement in matrix notation. Let $A = N^\top \Sigma^{-1} N$ and $B = N^\top \Sigma^{-1} M$. One observes that $(N \frown y)^\top (\Sigma')^{-1} (N \frown y) = A + yy^\top$ and $(N \frown y)^\top (\Sigma')^{-1} (M \frown x) = B + yx^\top$. Since $\operatorname{rank}(N) = F$, both A and $A + yy^\top$ are invertible, so the Sherman–Morrison formula (Sherman, 1949) yields $(A + yy^\top)^{-1} = A^{-1} - \frac{1}{1+g} A^{-1} yy^\top A^{-1}$ where $g = y^\top A^{-1} y > 0$ by positive definiteness of A^{-1}. Hence,

$$(A + yy^\top)^{-1}(B + yx^\top) = A^{-1}B + A^{-1}yx^\top - \frac{1}{1+g}\left[A^{-1}yy^\top A^{-1}B + A^{-1}yy^\top A^{-1}yx^\top\right].$$

Since $A^{-1}y(y^\top A^{-1}y)x^\top = A^{-1}ygx^\top = g\,A^{-1}yx^\top$, one obtains

$$A^{-1}B + A^{-1}yx^\top - \frac{1}{1+g}[A^{-1}yy^\top A^{-1}B + gA^{-1}yx^\top] = A^{-1}B + \frac{1}{1+g}A^{-1}yx^\top$$
$$-\frac{1}{1+g}A^{-1}yy^\top(A^{-1}B) = A^{-1}B + \frac{1}{1+g}[A^{-1}y(x^\top - y^\top) + A^{-1}yy^\top(I - A^{-1}B)].$$

□

Proof of Proposition 3. By Proposition 2 with $y = x$, $\tilde{f}(x) - \tilde{f}(0) = E(f - \tilde{f}(0))$ where $E = \frac{1}{1+g}(N^\top\Sigma^{-1}N)^{-1}xx^\top$ and $g = x^\top(N^\top\Sigma^{-1}N)^{-1}x > 0$. Setting $(1-\alpha) \coloneqq \frac{g}{1+g} \in (0,1)$ so that $x^\top E = (1-\alpha)x^\top$, one obtains

$$x^\top\tilde{f}(x) = x^\top(\tilde{f}(0) + (\tilde{f}(x) - \tilde{f}(0)) = x^\top\tilde{f}(0) + x^\top E(f - \tilde{f}(0)) = \alpha x^\top\tilde{f}(0) + (1-\alpha)x^\top f.$$

Uniqueness of α when $\tilde{f}(0) \neq f$ is immediate. Since $x^\top(N^\top\Sigma^{-1}N)^{-1}x$ is increasing in $||x||_2$, scaling x by a constant greater (less) than 1 decreases (increases) α. □

Proof of Proposition 4. I. Follows directly from Proposition 3.

II.A. Follows from the proof of II.B.

II.B. Consider first $x = (x_1, 0, \dots)^\top$ and set

$$\Sigma = I_F, \quad M = I_F, \quad N = \begin{bmatrix} a & 0 & 0 \\ Ka & 1 & 0 \\ 0 & 0 & I_{F-2} \end{bmatrix}.$$

By Lemma 1 one may restrict to the first two dimensions; abusing notation, let N and M denote the corresponding 2×2 submatrices. Then

$$S_1 = N^{-1}M = \begin{bmatrix} \frac{1}{a} & 0 \\ -K & 1 \end{bmatrix}, \qquad I_2 - S_1 = \begin{bmatrix} \frac{a-1}{a} & 0 \\ K & 0 \end{bmatrix}.$$

Computing E and $E(I_2 - S_1)$ and applying Proposition 2 yields

$$E(I_2 - S_2) = c\begin{bmatrix} \frac{a-1}{a} & 0 \\ -(a-1)K & 0 \end{bmatrix} \quad \text{and} \quad I_2 - S_2 = \begin{bmatrix} (1-c)\frac{a-1}{a} & 0 \\ (1+(a-1)c)K & 0 \end{bmatrix} \quad \text{for} \quad c = \frac{x_1^2}{a^2 + x_1^2} \in (0,1).$$

Hence, the value of x equals

$$|f_1| \cdot \left[||\left(\frac{a-1}{a}, K\right)^\top||_2 - ||\left(\frac{a-1}{a} - c\frac{a-1}{a}, K + (a-1)cK\right)^\top||_2 \right].$$

Since $c(a-1) > c\frac{a-1}{a}$, for $K \geq 1$, the increase in the second component strictly dominates the decrease in the first. If additionally $K > |\frac{a-1}{a}|$, the second dimension is strictly larger than

the first with and without x. Hence, for a sufficiently large K the value is strictly negative for all f with $f_1 \neq 0$ (and zero when $f_1 = 0$, establishing II.A). As $K \to \infty$, the error reduction in dimension 1 remains bounded while the error increase in dimension 2 is unbounded, so for any $\delta > 0$ there exists K large enough that the value is less than $-K$ for all f with $|f_1| > \delta$.

For arbitrary x, let V be a unitary matrix such that $x = ||x||_2 V^\top e_1$ and apply the above construction to $||x||_2 e_1$ with threshold $\delta' = ||x||_2^{-1}\delta$. Setting $M' = MV^\top$ and $N' = NV^\top$, Lemma 3 gives $V(x \mid M', N', \Sigma, f) = V(||x||_2 e_1 \mid M, N, \Sigma, V^\top f)$. Since $(V^\top f)_1 = ||x||_2^{-1} x^\top f$, the condition $|(V^\top f)_1| > \delta'$ is equivalent to $|x^\top f| > \delta$, completing the proof. □

Proof of Proposition 5. This follows from the strictly stronger Proposition 8. □

Proof of Proposition 6. Let $A = (N^\top \Sigma^{-1} N)^{-1}$ and let $A = Q^\top \Lambda Q$ be its eigendecomposition with Q unitary and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_F)$. Set $x = \alpha Q^\top e_i$ for any $\alpha > 0$ and any i. Then $Axx^\top = \alpha^2 \lambda_i Q^\top \delta_{ii} Q$, so $g = x^\top A x = \alpha^2 \lambda_i > 0$ and $E = \frac{\alpha^2 \lambda_i}{1+\alpha^2 \lambda_i} Q^\top \delta_{ii} Q$. Defining $B \coloneqq Q S_1 Q^\top$, one has $I - S_1 = Q^\top (I - B) Q$ and $I - S_2 = Q^\top (I - E)(I - B) Q$. The value of the additional source for fundamentals $Q^\top f$ is

$$
\begin{aligned}
&||Q^\top (B - I) Q Q^\top f||_2 - ||Q^\top \left[I - \frac{\alpha^2 \lambda_i}{1 + \alpha^2 \lambda_i} \delta_{i,i} \right] (B - I) Q Q^\top f||_2 \\
&= ||(B - I) f||_2 - || \left[I - \frac{\alpha^2 \lambda_i}{1 + \alpha^2 \lambda_i} \delta_{i,i} \right] (B - I) f||_2
\end{aligned}
$$

where the equality holds because the Euclidean norm is invariant under multiplication with a unitary matrix. Since $\left(I - \frac{\alpha^2 \lambda_i}{1+\alpha^2 \lambda_i} \delta_{ii} \right)$ scales only the i-th row by $\frac{1}{1+\alpha^2 \lambda_i} \in (0, 1)$ and leaves all other rows unchanged, the second norm is no larger than the first. Hence, the value is non-negative for all $Q^\top f$ and thus all f. If $\tilde{f}(0) \neq f$ for some f, then $S_1, B \neq I$. Choosing an i where the i-th row of B differs from $e_i^\top$ ensures the value is strictly positive whenever the i-th entry of $(I - B)^\top Q f$ is nonzero, which is generically the case. □

Proof of Proposition 7. *(⇐)* If v is a common eigenvector of $(N^\top N)^{-1}$ for all $N \in \mathcal{A}$, then there exists one unitary Q such that $A = (N^\top N)^{-1} = Q^\top \Lambda_N Q$ and $v = Q^\top e_i$ for some i simultaneously for all N. Choosing $x = \alpha v$, non-negative value follows by the same argument as in the proof of Proposition 6, simultaneously for all $N \in \mathcal{A}$, which proves the

direction.

(⇒) Suppose no common eigenvector exists. Fix any x; then there exists $N \in \mathcal{A}$ such that x is not an eigenvector of $A^{-1} = (N^\top N)^{-1}$. Fix this N and let $c \coloneqq \frac{1}{1+x^\top A^{-1}x} > 0$ and $v \coloneqq A^{-1}x$, such that $E = cvx^\top$ and therefore $I - S_2 = (I - cvx^\top)(I - S_1)$. Since x is not an eigenvector of A^{-1}, the vectors x and v are linearly independent, so there exists $y \in \mathbb{R}^F$ with $v^\top y = 0$ and $x^\top y \neq 0$. Let $B = N^\top M$. Since N has full rank, instead of setting M one can directly set B. Set $B = A(I - yz^\top)$ for an arbitrary nonzero z for which $(I - yz^\top)$ is invertible. Then $S_1 = A^{-1}B = I - yz^\top$ so that $I - S_1 = yz^\top$. Then for any f, $(I - S_1)f = (z^\top f)y$ and $(I - S_2)f$ equals

$$(I - cvx^\top)(I - S_1)f = (I - cvx^\top)(z^\top f)y = (z^\top f)(y - cvx^\top y) = (z^\top f)(y - c(x^\top y)v).$$

Since $v^\top y = 0$, the expanding $||(I - S_2)f||_2^2$ gives $(z^\top f)^2$ times

$$||y - c(x^\top y)v||_2^2 = ||y||_2^2 + ||-c(x^\top y)v||_2^2 = ||y||_2^2 + c^2(x^\top y)^2||v||_2^2.$$

Now $||I - S_1||_2^2 = (z^\top f)^2||y||_2^2$ and hence $V(x \mid M, N, \Sigma, f) = (z^\top f)^2 c^2 (x^\top y)^2 ||v||_2^2$ where all terms except $(z^\top f)^2 c$ are strictly greater 0. Thus, for all f with $z^\top f \neq 0$, which is generically the case, the value is strictly negative and it is weakly negative for all f. □

Proof of Proposition 8. *Step 1 — Reduction to diagonal M.* It suffices to prove the result for diagonal M. Since $\text{rank}(M) = F$, one may select F linearly independent rows and assume the remaining rows are perceived irrelevant; by Lemma 1 the learning outcome is identical if all remaining rows are deleted, so M may be taken square. By the singular value decomposition, write $M = UDV^\top$ with U, V unitary and D diagonal. Suppose the claim holds for D and $V^\top x$ with sequence N_k' and $\Sigma' = I$; set $N_k = UN_k'V^\top$ and $\Sigma = U\Sigma'U^\top = I$. By Lemma 3, the value of x under (M, N_k, Σ) equals the value of $V^\top x$ under (D, N_k', Σ'), which is negative for almost all f. Note that since U, V are independent of k, indeed $N_k \to M$.

Step 2 — Diagonal M. Fix $M = \text{diag}(m_1, \ldots, m_F)$ and x. Define $N(\epsilon) = M + u(\epsilon)x^\top$ where $u(\epsilon) = \epsilon u_1 + \epsilon^2 u_2$ for $u_1, u_2 \in \mathbb{R}^F$ and $\epsilon > 0$. Then $N(\epsilon) \to M$ and $N(\epsilon) \in GL_F(\mathbb{R})$ for ϵ small, since $GL_F(\mathbb{R})$ is open. Write $\mu(\cdot) \coloneqq \langle M^{-1}x, \cdot\rangle$, $d(\cdot) \coloneqq \langle M^{-3}x, \cdot\rangle$, and $c_0 \coloneqq x^\top M^{-2}x > 0$.

By the Sherman–Morrison formula,

$$N(\epsilon)^{-1} = M^{-1} - \frac{M^{-1}u(\epsilon)x^\top M^{-1}}{1 + x^\top M^{-1}u(\epsilon)}, \quad \text{so} \quad (I - S_1)f = \frac{x^\top f}{1 + \epsilon\mu(u_1) + \epsilon^2\mu(u_2)}\, M^{-1}u(\epsilon).$$

In particular $(I - S_1)f = 0 \iff x^\top f = 0$. Now $N(\epsilon)^\top N(\epsilon) = M^2 + \epsilon(Mu_1x^\top + xu_1^\top M) + O(\epsilon^2)$ and one can apply $(A+\epsilon B)^{-1} = A^{-1} - \epsilon A^{-1}BA^{-1} + O(\epsilon^2)$, which follows for example from the Neumann series. One obtains

$$\begin{aligned}(N(\epsilon)^\top N(\epsilon))^{-1} &= M^{-2} - \epsilon M^{-2}(Mu_1x^\top + xu_1^\top M)M^{-2} + O(\epsilon^2)\\ &= M^{-2} - \epsilon\left(M^{-1}u_1x^\top M^{-2} + M^{-2}xu_1^\top M^{-1}\right) + O(\epsilon^2)\end{aligned}$$

and thus

$$\begin{aligned}(N(\epsilon)^\top N(\epsilon))^{-1}x &= M^{-2}x - \epsilon(c_0M^{-1}u_1 + \mu(u_1)M^{-2}x) + O(\epsilon^2),\\ g = c_0 - 2\epsilon c_0\mu(u_1) + O(\epsilon^2), &\qquad \frac{1}{1+g} = \frac{1}{1+c_0} + O(\epsilon).\end{aligned}$$

Set $v \coloneqq (I - S_1)f$ and $w \coloneqq \frac{x^\top v}{1+g}(N^\top N)^{-1}x$, so $(I - S_2)f = v - w$. The value of x is negative iff

$$\Delta \coloneqq ||(I - S_2)f||_2^2 - ||(I - S_1)f||_2^2 = -2\langle v, w\rangle + ||w||_2^2 > 0.$$

The final constructions depends on whether $\mu(\cdot)$ and $d(\cdot)$ define the same or a different hyperplane (i.e. whether $M^{-1}x$ and $M^{-3}x$ are collinear).

Case 1 — Hyperplanes differ. There exists u_1 with $\mu(u_1) > 0$ and $d(u_1) < 0$; set $u_2 = 0$. Then $v = \frac{\epsilon x^\top f}{1+\epsilon\mu(u_1)}M^{-1}u_1$, $x^\top v = \frac{\epsilon\mu(u_1)}{1+\epsilon\mu(u_1)}x^\top f$, and $w = \frac{\epsilon\mu(u_1)x^\top f}{(1+c_0)(1+\epsilon\mu(u_1))}(M^{-2}x + O(\epsilon))$, giving

$$\langle v, w\rangle = \frac{\epsilon^2\mu(u_1)d(u_1)}{1+c_0}(x^\top f)^2 + O(\epsilon^3), \qquad ||w||_2^2 = \frac{\epsilon^2\mu(u_1)^2x^\top M^{-4}x}{(1+c_0)^2}(x^\top f)^2 + O(\epsilon^3).$$

Note that $x^\top M^{-4}x$. Hence $\Delta = \epsilon^2(x^\top f)^2\left[-\frac{2\mu(u_1)d(u_1)}{1+c_0} + \frac{\mu(u_1)^2c_0}{(1+c_0)^2}\right] + O(\epsilon^3)$, which is strictly positive for sufficiently small ϵ whenever $x^\top f \neq 0$, which is true for almost all f.

Case 2 — Identical hyperplanes. Then $d(\cdot) = \lambda\mu(\cdot)$ for some λ. Choose $u_1 \neq 0$ with $\mu(u_1) = 0$ (hence $d(u_1) = 0$) and u_2 with $\mu(u_2) > 0$. Then $v = \frac{x^\top f}{1+\epsilon^2\mu(u_2)}M^{-1}(\epsilon u_1 + \epsilon^2 u_2)$, $x^\top v = \frac{\epsilon^2\mu(u_2)}{1+\epsilon^2\mu(u_2)}x^\top f$, and $w = \frac{\epsilon^2\mu(u_2)x^\top f}{(1+c_0)(1+\epsilon^2\mu(u_2))}(M^{-2}x - \epsilon c_0M^{-1}u_1 + O(\epsilon^2))$. Using $d(u_1) = 0$,

$$\langle v, w\rangle = \frac{\epsilon^4\mu(u_2)}{1+c_0}(d(u_2) - c_0||M^{-1}u_1||_2^2)(x^\top f)^2 + O(\epsilon^5), \quad ||w||_2^2 = \frac{\epsilon^4\mu(u_2)^2x^\top M^{-4}x}{(1+c_0)^2}(x^\top f)^2 + O(\epsilon^5).$$

Hence $\Delta = \epsilon^4(x^\top f)^2\left[\frac{2\mu(u_2)(c_0||M^{-1}u_1||_2^2 - d(u_2))}{1+c_0} + \frac{\mu(u_2)^2x^\top M^{-4}x}{(1+c_0)^2}\right] + O(\epsilon^5)$. For a sufficiently scaled down u_2 the bracket is positive, so $\Delta > 0$ for small ϵ whenever $x^\top f \neq 0$. □

Proof of Proposition 9. *Step 1 — Preparations.* As in the proof of Proposition 8, it suffices to consider diagonal M. By Lemma 2, one may fix $\Sigma = I$ and scale x by a positive constant. By norm equivalence on $\mathbb{R}^F$, it suffices to find for any x and $C > 1$ a N and a scaling of x such that $||(I - S_2)f||_2/||(I - S_1)f||_2 > C$ for almost all f.

Step 2 — Diagonal M. Fix M, x, and set $c_0 \coloneqq x^\top M^{-2}x > 0$. For $u \in \mathbb{R}^F$, define $N = M + ux^\top$ and $\mu \coloneqq x^\top M^{-1}u$, assuming $1 + \mu \neq 0$. By the Sherman–Morrison formula,

$$N^{-1} = M^{-1} - \frac{M^{-1}ux^\top M^{-1}}{1+\mu}, \text{ and thus } (I - S_1)f = \frac{x^\top f}{1+\mu}M^{-1}u.$$

Set $v \coloneqq (I - S_1)f$ and thus $x^\top v = \frac{\mu}{1+\mu}x^\top f$. Since N is square, $(N^\top N)^{-1} = N^{-1}(N^{-1})^\top$. Moreover, because M is diagonal, $M^\top = M$ and $(M^{-1})^\top = M^{-1}$. Therefore one gets

$$\begin{aligned} N^{-1}(N^\top)^{-1} &= \left(M^{-1} - \frac{M^{-1}ux^\top M^{-1}}{1+\mu}\right)\left(M^{-1} - \frac{M^{-1}xu^\top M^{-1}}{1+\mu}\right) \\ &= M^{-2} - \frac{M^{-2}xu^\top M^{-1} + M^{-1}ux^\top M^{-2}}{1+\mu} + \frac{\mu\, M^{-1}ux^\top M^{-2}}{(1+\mu)^2}. \end{aligned}$$

Note that scaling the additional source x does not mean scaling the x in the construction of N, and therefore the above calculations are independent of scaling x. Now, one computes $(N^\top N)^{-1}(ax) = a\left(\frac{1}{1+\mu}M^{-2}x - \frac{c_0}{(1+\mu)^2}M^{-1}u\right)$ and $(ax)^\top(N^\top N)^{-1}(ax) = a^2\left(\frac{c_0}{(1+\mu)^2}\right)$. It follows that $g = \frac{a^2c_0}{(1+\mu)^2}$. Substituting into $(I - S_2)f = v - \frac{1}{1+g}(N^\top N)^{-1}(ax)\cdot(ax)^\top v$:

$$(I - S_2)f = \frac{x^\top f}{1+\mu}M^{-1}u - \frac{a^2\mu\, x^\top f}{(1+\mu)^2(1+g)}\left(\frac{M^{-2}x}{1+\mu} - \frac{c_0\, M^{-1}u}{(1+\mu)^2}\right).$$

Note that $1 + g = \frac{(1+\mu)^2 + a^2c_0}{(1+\mu)^2}$, so that $\frac{a^2}{(1+\mu)^2(1+g)} = \frac{a^2}{(1+\mu)^2+a^2c_0}$. Collecting terms in $M^{-1}u$ and $M^{-2}x$ yields

$$(I - S_2)f = \frac{x^\top f}{(1+\mu)^2 + a^2c_0}\Big[(1 + \mu + a^2c_0)M^{-1}u - a^2\mu\, M^{-2}x\Big].$$

Hence for $x^\top f \neq 0$, the ratio $||(I - S_2)f||_2/||(I - S_1)f||_2$ is independent of f and equals

$$\frac{|1+\mu|}{(1+\mu)^2 + a^2c_0} \cdot \frac{||(1 + \mu + a^2c_0)M^{-1}u - a^2\mu M^{-2}x||_2}{||M^{-1}u||_2}.$$

Letting $a \to \infty$,

$$\lim_{a\to\infty} \frac{||(I - S_2)f||_2}{||(I - S_1)f||_2} = \frac{|1+\mu|}{c_0} \cdot \frac{||c_0M^{-1}u - \mu M^{-2}x||_2}{||M^{-1}u||_2}.$$

Now choose w such that $M^{-1}w$ is not parallel to $M^{-2}x$ and $x^\top M^{-1}w \neq 0$, and set $u = kw$

for $k > 0$. Then $\mu = k(x^\top M^{-1}w)$ and

$$Q := \frac{||c_0 M^{-1}u - \mu M^{-2}x||_2}{||M^{-1}u||_2} = ||\frac{M^{-1}w}{||M^{-1}w||_2} - \frac{x^\top M^{-1}w}{c_0||M^{-1}w||_2}M^{-2}x||_2 > 0,$$

is independent of k, and strictly positive since $M^{-1}w$ and $M^{-2}x$ are not parallel. Thus

$$\lim_{a\to\infty} \frac{||(I - S_2)f||_2}{||(I - S_1)f||_2} = |1 + k(x^\top M^{-1}w)|\, Q \to \infty \quad (k \to \infty).$$

Choose k large enough that the limit exceeds $2C$, fix $u = kw$, then choose a sufficiently large so that the ratio exceeds C. □

Proof of Theorem 1. The proof proceeds in two steps. First, the main theorem of Berk (1966) is used to show that the beliefs are asymptotically carried on the set of points that minimize the Kullback-Leibler divergence. Then it is shown that there is a unique point that minimizes the divergence.

In order to use Berk's theorem there are four assumptions that need to be checked. First, the subjective densities have to be continuous in the parameters. The perceived density of signals given a fundamental g is

$$\frac{1}{\sqrt{(2\pi)^S \det(\Sigma)}} \exp\left(-\frac{1}{2}(s - Ng)^\top \Sigma^{-1}(s - Ng)\right)$$

which is obviously continuous in g.

Second, it has to hold that the subjective densities are only 0 on a set of measure 0 w.r.t. the true density. Since the subjective densities are nowhere 0, this condition is satisfied.

For the third assumption to hold, showing the following is sufficient. For every parameter g there exists some open neighborhood U where the expected supremum of the absolute value of the log-likelihood is finite, i.e.

$$\mathbb{E}\left[\sup_{g'\in U} |\log l(s|g')|\right] < \infty.$$

Now the absolute value of the log-likelihood is given by

$$|\log l(s|g')| = |-\frac{1}{2}\left(\log((2\pi)^S \det(\Sigma)) + (s - Ng')^\top \Sigma^{-1}(s - Ng')\right)| \leq$$
$$|\log((2\pi)^S \det(\Sigma))| + c||(s - Ng')||^2.$$

for a constant c[19]. The first part, $|\log((2\pi)^S \det(\Sigma))|$, is constant and $||(s - Ng')||^2$ is con-

[19]This holds because Σ and hence Σ^{-1} is a positive definite matrix. The constant depends on the eigen-

tinuous in g'. Therefore, the entire expression is bounded in $||s||$ for a bounded set U. Since $\mathbb{E}[||s||]$ exists, $\mathbb{E}\left[\sup_{g' \in U} |\log l(s|g')|\right] < \infty$.

The fourth and final condition is satisfied if one shows that for every $c \in \mathbb{R}$ there exists a co-compact set D, i.e., D^C is compact, such that

$$\mathbb{E}[\sup_{g' \in D} \log l(s|g')] \leq c.$$

Notice that

$$\log l(s|g') = -\frac{1}{2}\left(\log((2\pi)^S \det(\Sigma)) + (s - Ng')^\top \Sigma^{-1}(s - Ng')\right).$$

The first part is constant and $(s - Ng')^\top \Sigma^{-1}(s - Ng')$ is bounded from below by a positive constant λ[20] times $||s - Ng'||^2$. Now $||s - Ng'|| \geq ||(\mu - Ng')|| - ||\epsilon||$ for $\epsilon = s - \mu$. Consider the co-compact set given by $||(\mu - Ng')|| > k$ and assume that $||\epsilon|| \leq \frac{k}{4}$. Then

$$\log l(s|g') \leq -\frac{1}{2}\left(\log((2\pi)^S \det(\Sigma)) + \lambda \frac{k^2}{2}\right)$$

goes to $-\infty$ as $k \to \infty$. Moreover, since $0 \geq \log l(s|g')$,

$$\begin{aligned}\mathbb{E}\left[\sup_{g' \in D} \log l(s|g')\right] &\leq \mathbb{E}\left[\mathbb{1}_{||\epsilon|| \leq \frac{k_1}{4}} \sup_{g' \in D} \log l(s|g')\right] \\ &= \mathbb{P}\left[||\epsilon|| \leq \frac{k_1}{4}\right] \times \mathbb{E}\left[\sup_{g' \in D} \log l(s|g') \mid ||\epsilon|| \leq \frac{k_1}{4}\right].\end{aligned}$$

Since $\mathbb{P}[||\epsilon|| \leq \frac{k}{4}]$ goes to 1 as k goes to infinity, it follows that the claim has to hold for some large enough k.

Having shown these four assumptions, Berk's theorem yields that in the limit the beliefs concentrate on the set of points that minimize the Kullback-Leibler divergence. Since the perceived distribution is always normal, the Kullback-Leibler divergence given a belief g equals

$$\int \rho(x) \log(\rho(x)) dx - \int \rho(x) \frac{1}{2}[-\log(\det \Sigma) - k \log(2\pi) - (x - Ng)^\top \Sigma^{-1}(x - Ng)] dx.$$

Since all but g is constant, the only relevant part for minimization is

$$\int \rho(x)(x - Ng)^\top \Sigma^{-1}(x - Ng) dx = \mathbb{E}_\rho[(x - Ng)^\top \Sigma^{-1}(x - Ng)].$$

Multiplying out and using the linearity of the expected value, one sees that this is equivalent

values of Σ^{-1}.

[20] The constant again depends on the eigenvalues of Σ^{-1}.

to minimizing $(\mu - Ng)^\top \Sigma^{-1}(\mu - Ng)$. Taking first order conditions and using basic matrix calculus, the minimization has a unique solution satisfying $\tilde{f} = (N^\top \Sigma^{-1} N)^{-1}(N^\top \Sigma^{-1}\mu)$. □

Proof of Lemma 2. From Proposition 1, $\tilde{f}(0)$ is invariant to scaling Σ. From Proposition 2, scaling Σ by $c > 0$ and x by $\sqrt{c}^{-1}$ leaves the effect of the additional source unchanged, as the two scalings cancel. Hence scaling only Σ by c is equivalent to scaling x by $\sqrt{c}$. □

Proof of Lemma 3. Straightforward calculation gives $V(N^\top \Sigma^{-1} N)^{-1} N^\top \Sigma^{-1} M V^\top = [(N')^\top (\Sigma')^{-1} N']^{-1}(N')^\top (\Sigma')^{-1} M'$, so $V S_1 V^\top = S_1'$. Similarly, $(N')^\top (\Sigma')^{-1} N' + zz^\top = V[N^\top \Sigma^{-1} N + xx^\top]V^\top$ and $(N')^\top (\Sigma')^{-1} M' + zx^\top = V(N^\top \Sigma^{-1} M + xx^\top)V^\top$, so $V S_2 V^\top = S_2'$. Hence $V(I - S_2)V^\top f = (I - S_2')V^\top f$ and likewise for S_1. Since V is unitary, $||V(I - S_i)f||_2 = ||(I - S_i)f||_2$, and therefore $V(x \mid M, N, \Sigma, f) = V(z \mid M', N', \Sigma', V^\top f)$. □

C Calculations of the Investment Example

I now verify the claims in Table 1 for all six cases. Throughout, let $a > 1$. The agent's belief is determined by $S_1 = N^{-1}M$ and $S_2 = ((N^\frown x)^\top (N^\frown x))^{-1}(N^\frown x)^\top (M^\frown x)$ without and with the additional information, respectively. Calculating these matrices is straightforward (the inverses are taken of 2-by-2 matrices) and I will abstain from doing these step by step.

Below I demonstrate the calculations for the first third and fifth row of Table 1 and the other rows are the respective symmetric case and thus follow immediately.

Case 1: $\mathrm{sgn}(x_1) = \mathrm{sgn}(x_2)$**,** $|x_2| \geq |x_1| > 0$**.** Recall that in this case

$$M = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}, \qquad N = \begin{bmatrix} a & 0 \\ a/2 & 1/2 \end{bmatrix}$$

yielding

$$S_1 = N^{-1}M = \begin{bmatrix} 1/a & 0 \\ 0 & 1 \end{bmatrix}, \qquad I - S_1 = \begin{bmatrix} \dfrac{a-1}{a} & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence the misinference in dimension 1 and 2 is $\frac{a-1}{a}|f_1|$ and 0, respectively.

Let $\Delta = a^2 + 5a^2x_2^2 + x_1^2 - 2ax_1x_2 > 0$. Inference with the additional source is given by

$$S_2 = \begin{bmatrix} \dfrac{a + 5ax_2^2 + x_1^2 - (a+1)x_1x_2}{\Delta} & 0 \\ \dfrac{(a-1)x_1(5ax_2 - x_1)}{\Delta} & 1 \end{bmatrix}, \qquad I - S_2 = \begin{bmatrix} \dfrac{(a-1)(a + 5ax_2^2 - x_1x_2)}{\Delta} & 0 \\ -\dfrac{(a-1)x_1(5ax_2 - x_1)}{\Delta} & 0 \end{bmatrix}.$$

Thus, misinference is 0 in both cases if $f_1 = 0$. Assume now $f_1 \neq 0$. Then in dimension 2 misinference is non-zero, since

$$|x_2| > |x_1| \Longrightarrow 5a|x_2| > |x_1| \Longrightarrow (5ax_2 - x_1) \neq 0$$

and $x_1 \neq 0$. Therefore, misinference in dimension 2 strictly worsens. Misinference in the first dimensions worsens iff

$$\frac{(a-1)(a + 5ax_2^2 - x_1x_2)}{\Delta} > \frac{a-1}{a} \Longleftrightarrow a(a + 5ax_2^2 - x_1x_2) > \Delta$$
$$\Longleftrightarrow a^2 + 5a^2x_2^2 - ax_1x_2 > a^2 + 5a^2x_2^2 + x_1^2 - 2ax_1x_2 \Longleftrightarrow ax_1x_2 > x_1^2$$

which holds under the assumptions in this case.

Case 2: $\text{sgn}(x_1) \neq \text{sgn}(x_2)$**,** $|x_2| \geq |x_1| > 0$**.** Recall that in this case

$$M = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \qquad N = \begin{bmatrix} a & 0 \\ a & -1 \end{bmatrix}$$

yielding

$$S_1 = \begin{bmatrix} 1/a & 0 \\ 0 & 1 \end{bmatrix}, \qquad I - S_1 = \begin{bmatrix} \dfrac{a-1}{a} & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence the misinference in dimension 1 and 2 is $\frac{a-1}{a}|f_1|$ and 0, respectively.

Let $\Delta = a^2 + 2a^2x_2^2 + x_1^2 + 2ax_1x_2 > 0$. Inference with the additional source is given by

$$S_2 = \begin{bmatrix} \dfrac{a + 2ax_2^2 + x_1^2 + (a+1)x_1x_2}{\Delta} & 0 \\ \dfrac{(a-1)x_1(x_1 + 2ax_2)}{\Delta} & 1 \end{bmatrix}, \qquad I - S_2 = \begin{bmatrix} \dfrac{(a-1)(a + 2ax_2^2 + x_1x_2)}{\Delta} & 0 \\ -\dfrac{(a-1)x_1(x_1 + 2ax_2)}{\Delta} & 0 \end{bmatrix}.$$

Thus, misinference is 0 in both cases if $f_1 = 0$. Assume now $f_1 \neq 0$. Then in dimension 2 misinference is non-zero, since

$$|x_2| > |x_1| \Longrightarrow 2a|x_2| > |x_1| \Longrightarrow (2ax_2 + x_1) \neq 0$$

and $x_1 \neq 0$. Therefore, misinference in dimension 2 strictly worsens. Note $a + 2ax_2^2 + x_1x_2 =$

$a + 2a|x_2|^2 - |x_1||x_2| \geq a + x_2^2(2a-1) > 0$. Misinference in the first dimensions worsens iff

$$\frac{(a-1)(a+2ax_2^2+x_1x_2)}{\Delta} > \frac{a-1}{a} \iff a(a+2ax_2^2+x_1x_2) > \Delta$$
$$\iff a^2 + 2a^2x_2^2 + ax_1x_2 > a^2 + 2a^2x_2^2 + x_1^2 + 2ax_1x_2 \iff -ax_1x_2 > x_1^2$$

which holds under the assumptions in this case.

Case 5: $x_1 \neq 0$**,** $x_2 = 0$**.** Recall that in this case

$$M = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}, \qquad N = \begin{bmatrix} a & 0 \\ b/2 & 1/2 \end{bmatrix},$$

where $b > a$ and $\frac{b-a}{b} > a - 1$. Inference without the additional source is given by

$$S_1 = \begin{bmatrix} 1/a & 0 \\ 1 - b/a & 1 \end{bmatrix}, \qquad I - S_1 = \frac{1}{a}\begin{bmatrix} a-1 & 0 \\ b-a & 0 \end{bmatrix}.$$

Hence the misinference in dimension 1 and 2 is $\frac{a-1}{a}|f_1|$ and $\frac{b-a}{a}|f_1|$, respectively. Note that $\frac{b-a}{a} > \frac{b-a}{ba} > \frac{a-1}{a}$.

Inference with the additional source is given by

$$S_2 = \begin{bmatrix} \dfrac{a+x_1^2}{a^2+x_1^2} & 0 \\ \dfrac{a^2-ab+x_1^2(1-b)}{a^2+x_1^2} & 1 \end{bmatrix}, \qquad I - S_2 = \begin{bmatrix} \dfrac{a(a-1)}{a^2+x_1^2} & 0 \\ \dfrac{a(b-a)+x_1^2(b-1)}{a^2+x_1^2} & 0 \end{bmatrix}.$$

Thus, misinference is 0 in both cases if $f_1 = 0$. If $f_1 \neq 0$, then misinference in dimension 1 strictly reduces, since $\frac{a(a-1)}{a^2+x_1^2} < \frac{a-1}{a} \iff a^2 < a^2 + x_1^2$. Misinference in dimension 2 worsens iff

$$\frac{a(b-a)+x_1^2(b-1)}{a^2+x_1^2} > \frac{b-a}{a} \iff ax_1^2(b-1) > x_1^2(b-a)$$
$$\iff a(b-1) > b-a \iff ab > b \iff a > 1.$$

Since the misinference in dimension 2 is higher than the one in dimension 1 without the additional source, showing that the decrease in dimension 1 is smaller than the increase in

dimension 2 is sufficient to show that the overall change is harmful. This is the case iff

$$\frac{a(b-a)+x_1^2(b-1)}{a^2+x_1^2}-\frac{b-a}{a}>\frac{a-1}{a}-\frac{a(a-1)}{a^2+x_1^2}$$

$$\iff a(a(b-a)+x_1^2(b-1))-(a^2+x_1^2)(b-a)>(a-1)(a^2+x_1^2)-a^2(a-1)$$

$$\iff (a(b-1)-(b-a))x_1^2>(a-1)x_1^2 \iff b(a-1)>(a-1)$$

which holds under the assumptions in this case.