The benefits of being misinformed *

Marcus Roel
mcs.roel@gmail.com
Beijing Normal University

Manuel Staab
manuel.staab@univ-amu.fr
Aix-Marseille University, CNRS, AMSE

November 9, 2022

The most recent draft is available [here].

Abstract

We explore how two fundamental mistakes in information processing - incorrect beliefs about the world and misperception of information - can be mitigated by a benevolent information moderator who has no superior access to information but is more skilled at interpreting it. We introduce a simple sender-receiver model in which a moderator (i.e., sender) can intercept and manipulate signals that contain information about a payoff-relevant state. We characterize when such manipulation can be beneficial, both for a decision maker unaware of any interference (naïve), and one who takes it into account (sophisticated). Contrasting the optimal moderation policies for both types, we find that sophistication allows the moderator to beneficially intervene in more cases but can render moderation less effective. A particularly interesting case arises when moderator and decision maker completely disagree about which action should follow which signal. It is shown that if there are at least three states, such complete disagreement can be caused by only small differences in how information is interpreted. We provide necessary and sufficient conditions for the possibility of complete disagreement and examine the consequences for moderation and welfare. What might look to an outside observer like malicious misinformation can make the decision maker strictly better-off, yet completely misinformed.

Keywords: misperception, overconfidence, persuasion
JEL Codes: D03, D81, D83

*We would like to thank Yann Bramoullé, Andrew Ellis, Erik Eyster, Gilat Levy, Matthew Levy, Francesco Nava, Ronny Razin, Balázs Szentes, and seminar participants at the Aix-Marseille School of Economics and the London School of Economics for helpful comments and suggestions. Manuel Staab gladly acknowledges support from the French government under the "France 2030" investment plan managed by the French National Research Agency (ANR-17-EURE-0020) and from the Excellence Initiative of Aix-Marseille University - A*MIDEX.

1 Introduction

We commonly encounter situations in which information is difficult to evaluate or interpret. In such circumstances, we can observe people taking actions in line with opposing hypothesis, despite having access to similar information. For example, a significant number of people refuse even essential vaccinations despite the very strong evidence of their benefit and despite the measurable increase in outbreaks of the related disease as a consequence of this refusal.[1] Experts, such a physicians, then often try to guide their patients' views on how to interpret and act on information.

In light of these observations, we revisit Blackwell's (1951) comparison of experiments under the assumption that information processing is not always flawless but impeded by inaccuracies or potentially even systematic mistakes. More specifically, we focus on when and how decision makers can be better-off with access to less, less accurate, or even misleading information. We do this by examining the potential role of a benevolent expert or gatekeeper - which we refer to as *moderator* - who has no superior access to information but who is possibly more skilled at interpreting it. This moderator can manipulate or destroy information before it reaches the decision maker. We see this as an approximation of situations where an expert can influence which and how information is seen by an individual. For instance, physicians often 'interpret' diagnostic tests for their patients. While they are unaware of the true state of a patient's health, they might have a better understanding of the accuracy of tests as well as the ex-ante likelihood of a condition. In the same spirit, a CEO might be decisive in whether a new product is launched, but managers responsible for market research and testing can affect what and how the results are presented to the CEO. While it is well understood that differences in preferences can generate incentives to transmit noisy and misleading information (e.g., Crawford and Sobel (1982), Brocas and Carrillo (2007), Kamenica and Gentzkow (2011), etc.), we are interested in examining to what extent this can arise simply from a different understanding of the information environment.

We analyze a simple sender-receiver model that captures the fundamentals of information acquisition and processing: a decision maker (DM), who takes the role of the receiver, chooses an action profile conditional on the results of an information experiment. The DM then observes a signal from the experiment that provides information about the payoff relevant state of nature, and subsequently implements the corresponding action. Before the signal is perceived by the decision maker, however, it reaches a moderator, whose preferences are aligned with the DM. The moderator,

[1] See, for instance, Poland and Jacobson (2001) and Larson et al. (2011) for an overview of factors shaping public (dis-)trust in vaccine safety and efficacy, and their consequences for public health. Motta et al. (2018) provides evidence for widespread misinformation and overconfidence regarding medical knowledge in the general population in the U.S..

acting as a sender, can decide whether to forward the signal truthfully, or apply a garbling, thereby reducing or altering the information content. This decision is determined by a *moderation policy* that the moderator can commit to before the DM determines their action profile. We further distinguish two cases to examine the role of strategic/informational sophistication: (1) an effectively non-strategic setting where the decision maker is unaware of any tampering by the moderator (*naive*) and (2), a setting where the DM takes into account the moderation policy when choosing the action profile (*sophisticated*).

In our model, information processing can be imperfect in two ways: a decision maker might hold inaccurate (*biased*) beliefs about the world and/or incorrectly assess the accuracy of information from the experiment (*misperception*). Both imperfections are motivated by the psychological and experimental literature on beliefs and perception: while the former captures concepts such as over- or underconfidence (Fischoff et al. (1977), Lichtenstein et al. (1982), Moore and Healy (2008)) or motivated beliefs (Epley and Gilovich (2016)), the latter broadly covers directional mistakes, such as confirmation bias (Bruner and Potter (1964), Darley and Gross (1983), Fischoff et al. (1977) Lichtenstein et al. (1982)) or one-sided updating to protect one's ego utility or self-image (Mobius et al. (2014), Eil and Rao (2011)), as well as simple errors. Both perception issues can also arise from a coarse representation of the information environment (Mullainathan (2002), Jakobsen (2022)).

Our model can thus be used to study a wide variety of imperfections in information processing; from random inaccuracies to systematic mistakes. For consistency and ease of interpretation, the analysis is phrased throughout to suggest that the moderator is free of such biases and misperceptions. The model can, however, be equally interpreted as a sender and receiver holding heterogeneous views about the information environment, without taking any stand as to the accuracy of each view.

Biased beliefs and misperception alter the value of experiments, distort posterior beliefs, and shift choices away from the optimal ones. This reduces welfare. Both imperfections have a common channel: they cause non-convexities (in beliefs) in the utility frontier, thus altering the value of experimentation. Beyond this, biased beliefs also affect the utility ranking of actions in the absence of any informative signals. These consequences allow a moderator to beneficially intervene in some cases. Nevertheless, the existence of a superior choice does not imply that the moderator can induce it. The decision maker's choice behaviour and signal perception constrain the influence of the moderator, and these constraints markedly differ between naive and sophisticated types. Exploring the consequences of these constraints is a key focus of the paper.

We characterize when a moderator can have a beneficial impact and what the optimal moderation policy for each type looks like. We find that sophistication allows for

the implementation of beneficial (i.e. utility increasing) moderation policies in more cases but interestingly, these policies might be less effective (i.e., less beneficial) than those for naive decision makers. We pay particular attention to the consequences for the information content of the perceived signals. It is demonstrated that destroying all (relative) information between at least some signals can be superior to a less aggressive garbling. And such moderation policies can be more beneficial for a naive than sophisticated decision maker, pointing to the heterogeneous effects of sophistication. This does, however, require a more complex information environment (non-binary) and/or heterogeneous prior beliefs.

As a key observation, we find that providing decision makers with more accurate information might not always be the only, or even optimal way to counteract inaccuracies in perception. For example, we demonstrate how in settings with more than three states, a decision maker that strictly underestimates the Blackwell informativeness of an experiment might benefit from a further garbling of information. Intensifying a perception issue can have a positive impact if an information environment is more complex. We also examine a particularly interesting case where moderator and decision maker completely disagree about which action should follow which signal. With *complete disagreement*, we mean that a moderator believes an action a should follow one signal, and action b another, with the decision maker holding the completely opposite view. Crucially, complete disagreement occurs 'naturally' in our setting, and not as a result of different preferences over actions or a fundamentally different understanding of the information structure. We show that in all but trivial cases, complete disagreement requires at least three states of nature but can arise without any distortions and only small (ϵ) differences in prior beliefs. Using a geometric approach, we fully characterize when such disagreement between DM and moderator can occur and provide a method for verifying the possibility in a given setting. We further examine the consequences for the optimal moderation policy of each type as well as the welfare implications. If there is complete disagreement, beneficial moderation is always possible but might again be more beneficial for a naive DM, particularly if the signal and choice environment is binary, but the state-space more complex.

If a decision maker is naive, complete disagreement calls for a moderation policy that completely reverses the link between signals and posteriors, leaving the DM completely misinformed, and yet better-off. What would look to an outside observer like malicious misinformation can simply be based on (small) differences in how information is interpreted. If the information environment is more complex, misinformation can occur even if interests are aligned and decision makers act non-strategically. At the same time, this implies that while strategic and informational sophistication can make a decision maker less vulnerable to manipulation by adversarial information sources, it can also negatively limit the ability of a benevolent expert to reduce the ef-

fects of biases and misunderstandings. In other words, raising individuals' awareness of possible manipulations, and thus increasing their resilience to misinformation, can have negative side-effects.

Taking a broader perspective, our results highlight the need for a comprehensive understanding of imperfections in information processing to improve decision-making. One-sided and simplistic approaches that fail to reflect the complexity of the perception issues or of the information and choice environment can have unexpected consequences. Nevertheless, as we demonstrate in this paper, interventions can be feasible and useful. While the analysis is admittedly abstract, it might provide guidance for a more detailed analysis of specific settings.

The remainder of the paper is organized as follows: the next section summarizes the relevant literature. This is followed by a formal description of our model in Section 3. The analysis (Section 4) starts by examining the different constraints imposed on moderation policies by the two types of decision makers (Section 4.2). We proceed by characterizing the optimal moderation policies and explore how they relate to the information and choice environment (Section 4.3). We then turn to cases of complete disagreement (Section 4.4). The last section concludes. All proofs are in the appendix.

2 Relevant literature

Blackwell (1951) formalizes when an information experiment is more informative than another. Marschak and Miyasawa (1968) transfer these statistical ideas to the realm of economics. The key finding is that no rational decision maker would choose to 'garble' their information, i.e. voluntarily introduce noise into experiments. Having more information, however, may not always be beneficial in economic settings, and can cause a disadvantage in some strategic interactions. For example, Hirshleifer (1971) highlights that public information may destroy mutually beneficial insurance possibilities. Information avoidance has also been documented in bargaining (Schelling (1956), Schelling (1960), Conrads and Irlenbusch (2013), Poulsen and Roos (2010)) and holdup problems (Tirole (1986), Rogerson (1992), Gul (2001)). Strategic benefits can also arise when a behavioral agent plays intrapersonal games. Carrillo and Mariotti (2000) and Benabou and Tirole (2002) show that garbling of information can increase the current self's payoff when individuals are time-inconsistent. There is also an extensive literature on psychological reasons to avoid information (Golman et al. (2017)). In our setting, however, benefits from less accurate information are not based on strategic considerations but rely on different interpretations of the information environment. Furthermore, information has a purely instrumental character.

Numerous studies have suggested that people hold incorrect beliefs, especially unrealistically positive views of their traits or prospects. To mention a few, see Weinstein

(1980) for health and salaries, Guthrie et al. (2001) for rates of overturned decisions on appeal by judges, and Fischoff et al. (1977) as well as Lichtenstein et al. (1982) for estimates of ones' own likelihood to answer correctly. Recent papers documented overconfidence in entrepreneurs (Landier and Thesmar (2009)), in CEOs (Malmendier and Tate (2005)), and in laboratory settings (Burks et al. (2013), Charness et al. (2018), Benoit et al. (2015)).

Perception biases have first been documented in the psychology literature, see, for example, Bruner and Potter (1964), Fischoff et al. (1977) Lichtenstein et al. (1982), and Darley and Gross (1983). The literature has explored many ways of modeling such biases, with different implications for learning. For example, Rabin and Schrag (1999) formalized them in a model of confirmation bias. They show how the tendency to misinterpret new information as supportive evidence for one's currently held hypothesis can not only lead to overconfidence in the incorrect hypothesis, but even cause someone to become fully convinced of it. Evidence for such one-sided updating includes Mobius et al. (2014) and Eil and Rao (2011). Perception biases might even deliver a benefit to the decision maker. For instance, Steiner and Stewart (2016) suggest that pure noise inherent in information processing creates problems akin to the winner's curse when unbiased perception strategies are employed. Optimal perception must therefore be biased, correcting for the mistake by inducing more cautious evaluations. Other perception issues may arise from how information is represented and processed in the brain (Brocas (2012)). Consistent with arguments in the latter study, the decision makers in our model have a fundamentally Bayesian approach to decision making, but beliefs and perception can be subject random or systematic inaccuracies.

This paper is closely related to recent theoretical work on persuasion, information design, and strategic communication more generally. While in most cases incentives to misinform arise from different preferences (Crawford and Sobel (1982)), we assume preferences to to be aligned between sender and receiver in order to remove strategic aspects, and instead highlight the role of differences in the interpretation of information. As in Lipnowski and Mathevet (2018), we examine the role of a benevolent expert, who controls the information flow. However, in our setting incentives to reduce informativeness of signals only arise from how information is used, not preferences over information as such. This aspect more closely resembles Alonso and Câmara (2016), who study Bayesian persuasion with heterogeneous priors, as well as de Clippel and Zhang (2022), who focus on non-Bayesian updating more generally. However, in our model the sender cannot freely design the information experiment. This also distinguishes it from Brocas and Carrillo (2007), where the sender can decide how much information to obtain. More importantly, we allow the sender and receiver to disagree over the experiment. As in Kamenica and Gentzkow (2011), the sender has commitment power. This, however, only plays a role for sophisticated decision makers.

3 Model

A decision maker (DM) inhabits a world which is characterized by one of a finite number of possible states Ω. The relevant state is not known to the DM, who instead holds some belief about which state applies. A belief is captured by an element $\boldsymbol{\mu} \in \Delta(\Omega)$, with $\Delta(\Omega)$ the set of all possible probability distributions over Ω. We interpret $\boldsymbol{\mu}$ as a vector, where each entry μ_ω corresponds to the probability the DM assigns to facing a world in state ω. We assume the DM believes all states can occur with positive probability, meaning $\boldsymbol{\mu} > \mathbf{0}$.

The DM faces a simple **decision problem**: they must choose an action from a (countable) set $\mathbb{A}$. The DM's payoff from an action $a \in \mathbb{A}$ depends on the state of the world and is represented by a utility function $u(a|\omega)$. To rule out trivial cases, $\mathbb{A}$ is assumed to contain at least two actions, and no actions that are payoff-equivalent or strictly dominant. To make their choice, the DM does not have to solely rely on their initial ('prior') belief about the world. They can perform an **information experiment** that might reveal additional information about the state. Any such experiment yields a result or **signal** $\boldsymbol{s}$. The relevant aspect of the signal is the probability with which it occurs in each state. Similar to Jakobsen (2022), we identify a signal by its **probability profile** and treat it as a vector $\boldsymbol{s} = (s_\omega)_{\omega\in\Omega}$, where s_ω is the probability that $\boldsymbol{s}$ is observed in state ω. An information experiment itself is characterized by the possible signals it yields with positive probability. For a given experiment X, let S_X denote the set of all such signals. This is assumed to be finite. Analogous to Blackwell (1951), we can consider X a matrix of dimension $|\Omega| \times |S_X|$, where rows correspond to states and columns to signals. An element $x_{\omega,\boldsymbol{s}}$ of this matrix gives the probability that signal $\boldsymbol{s}$ is observed in state ω. Each row of this matrix thus corresponds to the probability distribution over signals for a given state, meaning rows sum up to 1. The space of all such matrices is denoted by $\mathbb{X}$.

The experiment allows the DM to condition choices on signals. Given some $X \in \mathbb{X}$, an **action profile** $\boldsymbol{a} = (a_s)_{s\in S_x}$ is a vector of dimension $|S_X|$ that describes the choice after each possible signal. As a convention, action a_i in the action profile $\boldsymbol{a}$ refers to the action taken after signal $\boldsymbol{s}_i$. An action profile is called *signal sensitive* if at least two actions are distinct. The objective of a DM with access to an experiment X is to choose an action profile that maximizes (von Neumann-Morgenstern) expected utility:

$$\max_{\boldsymbol{a}\in\mathbb{A}^k} E\Big[E\big[u(a|\omega)|\boldsymbol{p}(\boldsymbol{s})\big]\Big], \tag{1}$$

where $k = |S_X|$, and $\boldsymbol{p}(\boldsymbol{s})$ denotes the DM's posterior belief after observing $\boldsymbol{s}$.

Distortions, biases, and Belief updating:

The DM might not be fully aware of the true signal structure of an experiment and

may instead hold a *distorted* view. At this point, we remain agnostic about where this distortion or *misperception* is coming from. For example, the information might be manipulated at the source without the knowledge of the DM, the DM might not understand the signal-generating process correctly, or might suffer from perception limitations. A **signal distortion** is a mapping $d : \mathbb{X} \mapsto \mathbb{X}$. When observing an experiment X with possible signals $S_X = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_k\}$, a DM with a distortion d is under the impression that the experiment yields signals $S_X^d = \{\boldsymbol{s}_1^d, ..., \boldsymbol{s}_n^d\}$. This again can be represented by a matrix $X^d \in \mathbb{X}$, where each column corresponds to a distorted signal.

A DM might also hold a biased view about which state is likely to occur before any information is observed. This is referred to as a **bias in prior**. The idea is that a DM holds a prior belief $\boldsymbol{p} \in \Delta(\Omega)$ that differs from some reference probability distribution $\boldsymbol{\mu}$. One could see $\boldsymbol{\mu}$ as the true probability distribution according to which states realize and $\boldsymbol{p}$ as a biased view due to previous perception mistakes. However, it could also be interpreted as a difference between the assessment of the DM and that of some observer, without any judgment regarding their accuracy.[2] In this case, the DM's belief is 'biased' merely from the perspective of the observer. When these are not identical, $\boldsymbol{p}$ always refers to the belief of the DM, and $\boldsymbol{\mu}$ to the true distribution (actual or presumed). Welfare is evaluated according to $\boldsymbol{\mu}$.

After observing the result of an experiment, the DM updates their prior belief $\boldsymbol{p}$ according to Bayes' rule. However, if the DM holds a distorted view, Bayes' rule is applied not on the basis of a signal $\boldsymbol{s}$, but its distorted version $\boldsymbol{s}^d$. The DM's updated belief can then be computed as:

$$\boldsymbol{p}(\boldsymbol{s}^d) = \frac{\boldsymbol{s}^d \circ \boldsymbol{p}}{\langle \boldsymbol{s}^d, \boldsymbol{p} \rangle} \tag{2}$$

where $\boldsymbol{s}^d \circ \boldsymbol{p}$ denotes the element-wise product, and $\langle \boldsymbol{s}^d, \boldsymbol{p} \rangle$ the dot product of the two vectors. In comparison, the actual signal generated from the experiment is $\boldsymbol{s}$, which would lead a fully Bayesian observer to revise their belief according to (3).

$$\boldsymbol{\mu}(\boldsymbol{s}) = \frac{\boldsymbol{s} \circ \boldsymbol{\mu}}{\langle \boldsymbol{s}, \boldsymbol{\mu} \rangle}. \tag{3}$$

As X and X^d are both experiments, **Bayes' consistency** holds with and without misperception, meaning the expected posterior equals the prior.[3]

As an illustration, consider a decision maker consulting a partisan news source. The DM might view signals to be more informative than they actually are and thus

[2] See Morris (1995) for a detailed discussion of the rationality of heterogeneous priors.

[3] More generally, one could imagine a distortion that results in an X^d that is not a proper experiment, i.e. the elements in a row do not sum to 1 (or a constant). For such an X^d, Bayes consistency would fail. Just from introspection, the DM could identify a problem in the decision-making. For simplicity, we exclude such cases. However, most results would remain unaffected by this generalization.

update more strongly than a neutral observer, who is aware of the partisan slant. Suppose, for instance, there is an (uninformative) signal $\boldsymbol{s}$ but the DM believes the signal contains some relevant information. So while the signal is actually independent of the true state, we have $s^d_\omega \neq s^d_{\omega'}$ for some states $\omega, \omega' \in \Omega$. Starting from a prior $\boldsymbol{p} = \boldsymbol{\mu}$, this leads to posteriors $\boldsymbol{\mu}(\boldsymbol{s}^d) \neq \boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{\mu}$. The DM updates their belief while a neutral observer's posterior belief would remain at their prior.

Information Moderation:

The information from the experiment is first observed by an **information moderator**. Before passing on the signal to the DM, the moderator can influence its content, i.e. 'moderate' the information flow. To avoid probability 0 events, a moderator can only select policies that are consistent with the signals of an experiment X. In particular, given some X, a moderator can change a signal $\boldsymbol{s} \in S_X$ to some $\boldsymbol{s}' \in S_X$, which is perceived by the DM instead. A **moderation policy** is a function

$$m : S_X \mapsto \Delta(S_X), \tag{4}$$

where $\Delta(S_X)$ is the convex hull of S_X. A moderation policy is called *deterministic* if it (effectively) maps to S_X. Since signals are characterized by their probability profile, the actual signals received by the DM given some moderation policy m are denoted by $\boldsymbol{s}^m$. If, for instance, $m(\boldsymbol{s}_i) = \boldsymbol{s}_j = m(\boldsymbol{s}_j)$, then $\boldsymbol{s}^m_j = \boldsymbol{s}_i + \boldsymbol{s}_j$ and $\boldsymbol{s}^m_i = \mathbf{0}$.[4] A moderation policy is a type of garbling and describes how signals are 'swapped'. It can be expressed as a $|S_X| \times |S_X|$ garbling matrix M. Each column i of M gives the probability with which the DM receives each signal in S_X, given that the moderator received signal $\boldsymbol{s}_i$. The experiment effectively becomes:

$$X^m = XM.$$

The moderator is assumed to have no intrinsic incentive to 'misinform' the DM. Preferences are fully aligned. Moreover, the moderator is not aware of the state but relies on the same signal realizations to update the prior. However, the moderator is not subject to distortions and thus (possibly) more sophisticated with regards to processing information or understanding the signal-generating process. The moderator is also assumed to be aware of the DM's distortions and biases, or at least their choice profile. In other words, the moderator updates based on the true X and $\boldsymbol{\mu}$, but is aware of the DM's view X^d and $\boldsymbol{p}$. The moderator may thus choose to moderate information according to some M. If any such garbling *strictly* increases expected utility, it is referred to as **beneficial moderation policy**. The moderator is also assumed to be able to commit to a moderation policy.

[4] For completeness, we assume $\boldsymbol{\mu}(\mathbf{0}) = \boldsymbol{\mu}$. Since this is a probability 0 event, this (or any other assumption on the resulting posterior) remains without consequences.

We further distinguish between two types of decision makers: those oblivious to the moderation policy, which we call *naive*, and those aware of any interference by the moderator, which we call *sophisticated*. In other words, a naive DM updates according to X^d, while a sophisticated one adjust for the moderation policy and instead updates according to $X^d M$. The latter case is also the one where commitment is (potentially) relevant, as the interaction between moderator and sophisticated DM is strategic. It remains innocuous for a naive DM.

This distinction of types allows us to examine more closely the interplay between strategic/informational sophistication, imperfections in decision making, and information moderation. One might, of course, plausibly argue that sophistication should also entail a superior understanding of the information experiment itself, ruling out a distortion d. In the absence of a bias in prior, the choices of a sophisticated DM would then correspond to those of the moderator. A sophisticated DM without a bias in prior could be seen as a benchmark. Instead, we focus on the more general case where both types can suffer from misperception. As will be shown, however, biases and misperception have comparable effects, rendering this assumption qualitatively mostly inconsequential.

Using again the illustration of a biased news source, suppose the source generates only uninformative signals $\boldsymbol{s}$ and $\boldsymbol{t}$, which the DM misperceives, such that $\boldsymbol{\mu}(\boldsymbol{s}^d) \neq \boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{\mu}$. Knowing about the distortion, the moderator might want to intervene. Suppose the action following $\boldsymbol{t}^d$ yields higher expected utility at $\boldsymbol{\mu}$ than that after $\boldsymbol{s}^d$. For a naive DM, the moderator would then want to transform signals $\boldsymbol{s}$ into $\boldsymbol{t}$, meaning $m(\boldsymbol{s}) = \boldsymbol{t}$ (which is perceived as $\boldsymbol{t}^d$). If the DM is sophisticated, any moderation policy that renders signals independent of the state (e.g $m(\boldsymbol{t}) = m(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{t} + \frac{1}{2}\boldsymbol{s}$) would lead the DM to conclude that the (moderated) signals are uninformative. Such a sophisticated DM then chooses the first-best.

Indirect Utility:

When comparing outcomes between decision makers with and without distortions and biases, it is useful to distinguish two cases: the maximum expected utility that can be obtained from the experiment, and the expected utility conditional on some action profile $\boldsymbol{a}$. Denote the maximum expected utility for a prior $\boldsymbol{\mu}$ from an experiment X without distortion by:

$$V(X|\boldsymbol{\mu}) \equiv \max_{\boldsymbol{a}^* \in \mathbb{A}^{|S_X|}} \sum_{\boldsymbol{s} \in S_X} \langle \boldsymbol{\mu}, \boldsymbol{s} \rangle \cdot E\big[u(a_s^*|\omega)|\boldsymbol{\mu}(\boldsymbol{s})\big]. \tag{5}$$

This is also referred to as the **value of the experiment** X. Similarly, the expected utility

from X given some action profile $\boldsymbol{a}$ is denoted by:

$$V(X|\boldsymbol{a},\boldsymbol{\mu}) \equiv \sum_{\boldsymbol{s}\in S_X} \langle \boldsymbol{\mu},\boldsymbol{s}\rangle \cdot E\big[u(a_s|\omega)|\boldsymbol{\mu}(\boldsymbol{s})\big]. \tag{6}$$

Any departure from $V(X|\boldsymbol{\mu})$ arises from a decision maker optimizing subject to distortions and biases, while the actual expected utility is determined by the undistorted X and unbiased $\boldsymbol{\mu}$. There are several key cases: The expected utility a DM subject to distortion d actually obtains from X can be expressed as $V(X|\boldsymbol{a}^d,\boldsymbol{\mu})$, where $\boldsymbol{a}^d$ is the action profile consistent with $V(X^d|\boldsymbol{\mu})$. If the DM has a prior belief $\boldsymbol{p}\neq\boldsymbol{\mu}$, then this DM obtains expected utility $V(X|\boldsymbol{a}^p,\boldsymbol{\mu})$, where $\boldsymbol{a}^p$ is the choice consistent with $V(X|\boldsymbol{p})$. When there is no confusion, the superscript on $\boldsymbol{a}$ will be omitted. Of course, the DM's choice could also be based on a combination of misperception and bias.

If a moderator intervenes, we have to distinguish between the two types of decision maker. Given a choice $\boldsymbol{a}$, the expected utility a naive DM subject to a moderation policy m obtains from X equals $V(XM|\boldsymbol{a},\boldsymbol{\mu})$. In comparison, a sophisticated DM obtains $V(XM|\boldsymbol{a}^*,\boldsymbol{\mu})$, where $\boldsymbol{a}^*$ is the action profile consistent with $V(X^dM|\boldsymbol{p})$, i.e. the optimal choice given a belief $\boldsymbol{p}$, perceived signals S_X^d, and moderation policy m.

Gain from Information

How valuable an experiment is depends on the signal strength and the chosen action profile (and more generally the set of available actions). Misperception and biases affect the perceived signal strength and posterior beliefs. They thus leads to a discrepancy between the value expected by the DM and that of a neutral observer. However, this only negatively impacts expected utility (as judged by an observer) through its effects on choices. Given an action profile $\boldsymbol{a}$, we define the *gain* from an experiment as the difference in expected utility between this choice and the action profile the DM would choose without access to an informative experiment. Since the DM might hold a different prior, choices are made according to $\boldsymbol{p}$ but evaluated against $\boldsymbol{\mu}$. Put differently, the gain from X uses as a reference the best outcome that can be achieved in the absence of informative signals and with the DM aware of this absence of information. It is thus similar to the definition in Kamenica and Gentzkow (2011), except for the conditioning on $\boldsymbol{a}$ and the potential difference in priors. The *conditional gain* makes a similar comparison to an outcome without any informative experiment, but relative to best outcome that can be achieved *among* the actions that are part of $\boldsymbol{a}$. The conditional gain again compares expected utility against a hypothetical setting without information, but with the DM unaware that X is uninformative. If signals don't contain information but induce different actions, which signal would maximize expected utility? The conditional gain uses this as a reference.

Definition 1 (Gain from information)**.** *Given an action profile* $\boldsymbol{a}=(a_1,...,a_l)$ *and prior*

belief $\boldsymbol{p}$*, the* ***conditional gain*** *from an experiment* X *at* $\boldsymbol{\mu}$ *is defined as*

$$V(X|\boldsymbol{a},\boldsymbol{\mu}) - \max_{a\in\{a_1,\dots,a_I\}} E[u(a|\omega)|\boldsymbol{\mu}],$$

and the ***gain*** *from* X *at* $\boldsymbol{\mu}$ *is defined as*

$$V(X|\boldsymbol{a},\boldsymbol{\mu}) - E[u(a^*|\omega)|\boldsymbol{\mu}],$$

with $a^* = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{p}]$.

4 Analysis

We first establish a baseline by characterising the effects of distortions and biases in the absence of a moderator (Proposition 1), particularly concerning the value of an experiment. Based on these findings, we explore when and how a moderator can help a DM reduce the effect of these perception issues and when moderation fails to have a positive impact. Particular emphasis is placed on contrasting the optimal (i.e. utility maximizing) moderation policy for naive decision makers (Proposition 2), with that for a sophisticated DM (Proposition 3), and exploring the consequences for posterior beliefs and expected utility. The aim is to not just explore the implications of imperfections and their potential mitigation, but also to highlight the (ambiguous) consequences of higher sophistication in information processing. We also examine what choices reveal about the type of distortion a DM faces and their implication on beneficial moderation (Proposition 4). Finally, we investigate a phenomenon we refer to as 'complete disagreement', where moderator and DM completely disagree about the implications of signals on optimal choices, despite (potentially) having a comparable qualitative understanding of the experiment. In these instances, the moderator would want to completely misinform a decision maker about at least some of the outcomes of the experiment. Theorem 1 characterizes when this can occur. The consequences for the optimal moderation policy in these cases for either type of decision maker are subsequently examined.

4.1 Distortions, biases and their implications

Unbeknownst to the decision maker, biases and distortions alter the value and gain from experimentation. Depending on the prior and the magnitude of the distortion, this can lead to a different choice of action profile and thus cause a utility loss. More generally, any distortion introduces non-convexities (in beliefs) in the utility frontier. The DM might fail to realise the full gains from an experiment either by relying too much or too little on the outcome. A DM with (only) a bias in prior judges the infor-

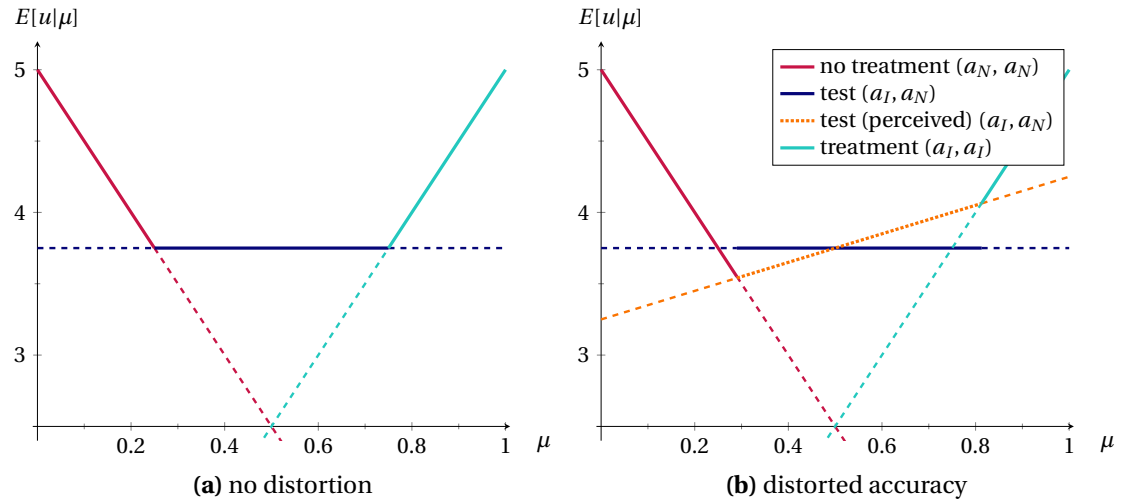
mativeness of signals correctly, but weighs them according to a different prior. This can lead to equally distorted posteriors and thus have similar effects. Moreover, a difference in prior can lead to disagreement over which action is to be taken even in the absence of information, and at beliefs where the utility frontier is locally convex. Example 1 illustrates these observations as well as the setting in general.

Example 1.1. (Diagnostic Testing) Suppose a patient was potentially exposed to an infectious disease. The patient is either infected (ω_I) or not (ω_N), and can immediately seek treatment (a_I), or continue as usual (a_N). To check for infection, the patient can perform a diagnostic test, and react based on the outcome, i.e. $\boldsymbol{a} = (a_I, a_N)$. Of course, it is also possible to ignore the test (which will be interpreted as not taking the test) and take either action independent of the test outcome.

Suppose the test provides informative but not fully revealing signals $\boldsymbol{s}$ and $\boldsymbol{t}$ with $s_I = 0.75 = t_N$. Let $u(a_I|\omega_I) = 5 = u(a_N|\omega_N)$, and 0 otherwise. For simplicity, rather than relying on the vector notation, let μ describe the ex-ante probability of having been infected, i.e. the true state being ω_I. It follows from the symmetry of payoffs that, if the test is taken, the patient performs action a_I or a_N depending on whether the posterior belief after observing the result is greater or smaller than $\frac{1}{2}$.

Figure 1 (a) illustrates the expected utility outcomes as a function of beliefs (here μ could be seen as the prior or the posterior). Profiles (a_I, a_I) and (a_N, a_N) are optimal for more extreme prior beliefs as the information provided by the test is not sufficient to move the posterior below/above $\frac{1}{2}$. If the patient is convinced that they have been infected, it is best to start treatment without relying on the test. The risk of a false-negative result outweighs the risk of unnecessary treatment. Equivalently, if infection is very unlikely, it is best to continue as usual. For intermediate beliefs, the information provided by a result is valuable as the gain from the test is strictly positive. As stated in Proposition 1, the maximum expected utility (bold line segments) is convex in μ.

Suppose now the patient misjudges the accuracy of the test by underestimating chance of a false negative result and overestimating the probability of a false positive. In particular, $s_I = 0.75 < s_I^d = 0.85$ and accordingly $t_I = 0.25 < t_I^d = 0.15$. Furthermore, $t_I = 0.75 > t_I^d = 0.65$, and thus $s_N = 0.25 < s_I^d = 0.35$. The changes are such that the patient underestimates the strength of a positive but overestimates that of a negative test result. This (incorrectly) raises the perceived value of the test for higher prior beliefs. When infection is more likely, an accurate negative signal is valuable, since it affects the chosen action and thus, in expectation, avoids unnecessary treatment. The patient would take the test for a range of prior beliefs, where immediate treatment should be the preferred option. Consequently, convexity of the true expected utility fails in regions where the information is only perceived to be valuable (in the neighbourhood of $\mu = 0.8$). Similarly, by overestimating the probability of false positives, the test appears less valuable to the patient at prior beliefs that put only small weight on a



(a) no distortion **(b)** distorted accuracy

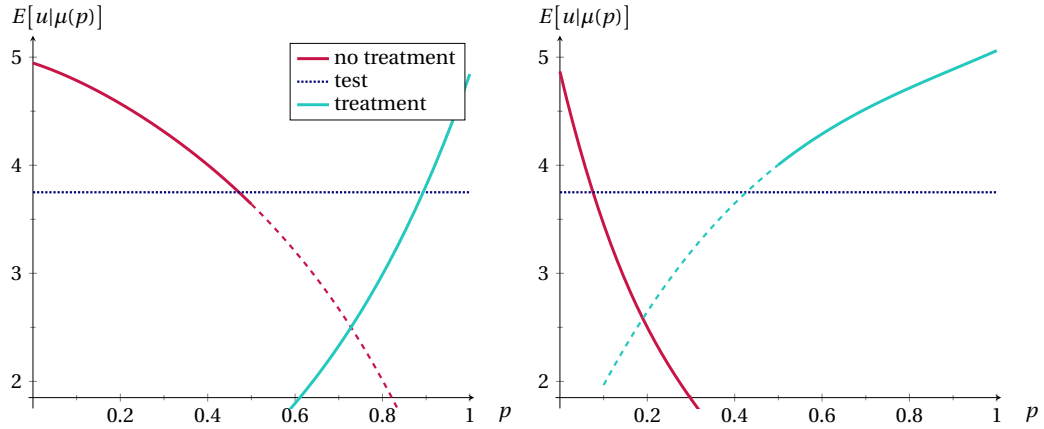
Figure 1: Expected utility of action profiles (Example 1.1)

possible infection. There is a range of prior beliefs where the test is not taken, despite being valuable (in the neighbourhood of $\mu = 0.29$). Again, the utility fails to be convex in that region. ◇

Suboptimal choices can also be caused by biases that affect the prior belief. If $\boldsymbol{\mu} \neq \boldsymbol{p}$, then the DM puts too much (little) weight on one of the states and thus (dis-)favours actions appropriate for that state. Moreover, signals are interpreted against this distorted prior, affecting posterior beliefs. Interestingly, the consequences from such a 'bias' in prior are similar to those of a distortion in signals. This echoes Brandenburger et al. (1992), which establishes the equivalence of distortions and heterogeneous priors in a correlated equilibrium setting.[5]. A key result from Alonso and Câmara (2016) (Proposition 1) shows that for a given difference in priors, there exists a simple relation between posterior beliefs that is independent of the information experiment. When applied in this context, it can be shown that, from the perspective of the moderator, the utility frontier (as generated by the choices of the DM) fails to be convex in beliefs. Similar to the case of signal distortions, the DM (possibly) misjudges gain from an experiment.

Example 1.2. Suppose before taking the test, the patient fails to accurately assess the risk of having been exposed to the disease. In particular, suppose the actual risk is lower than the belief of the patient ($1/5 = \mu < p = 2/5$). Then independent of the accuracy of any test, a negative signal is less and a positive more surprising to the observer than the patient. For a given change in belief of the patient, we can compute the implied signal that would yield this belief. Using this signal, we can calculate the belief an observer would reach. This reveals that - from the perspective of the observer - the expected utility of the patient fails to be convex. And in fact, for a range of posterior

[5]The equivalence, however, requires giving up Bayes' consistency of distorted beliefs



(a) prior beliefs: $\mu = \frac{1}{5}$, $p = \frac{2}{5}$ **(b)** prior beliefs: $\mu = \frac{4}{5}$, $p = \frac{2}{5}$

Figure 2: Expected utility (Ex. 1.2) as a function of posterior beliefs of the patient from the perspective of an observer. A solid line indicate the preferred action of the patient at each p.

beliefs (eg. $p(\boldsymbol{s}) \in (0.5, 0.75)$ in (**a**)), the observer disagrees with the patient over which action should be taken. For an experiment to be considered valuable by the observer at $p = 2/5$, it would need a higher accuracy. A negative signal would need to induce a posterior belief of $\mu(\boldsymbol{s}) > 1/2$ in the observer. This is not the case for the test in question. A similar situation is shown in (b), but here the observer believes ex-ante that infection is more likely. Again, convexity in (posterior) beliefs fails for some range, indicating a disagreement over the value of experiments. ◇

Proposition 1 provides a formal summary of the previous discussions and highlights the shared channel, through which biases and distortions negatively affect choices. Let $\hat{V}(X^d|\boldsymbol{p})$ denote the expected utility a DM with a given bias and distortion actually obtains. In particular, $\hat{V}(X^d|\boldsymbol{p}) = V(X|\boldsymbol{a}^*, \boldsymbol{\mu})$, where $\boldsymbol{a}^*$ is the choice the DM deems optimal (i.e., $\boldsymbol{a}^* = \arg\max_{a \in \mathbb{A}^{|S_X|}} V(X^d|\boldsymbol{a}, \boldsymbol{p})$).

Proposition 1. *For any X and prior $\boldsymbol{\mu}$, indirect utility $V(X|\boldsymbol{\mu})$ is convex in $\boldsymbol{\mu}$. Convexity of $\hat{V}(X^d|\boldsymbol{p})$ in beliefs fails for at least some $\boldsymbol{p}$ or X if one of the following holds:*

- *a DM suffers from a non-trivial distortion ($X^d \neq X$),*
- *a DM holds a biased prior ($\boldsymbol{\mu} \neq \boldsymbol{p}$).*

Without any imperfections, $\hat{V}(X|\boldsymbol{\mu}) = V(X|\boldsymbol{\mu})$, which is convex in beliefs for any X and $\boldsymbol{\mu}$. Intuitively, if a DM was offered some additional information experiment before observing the result of X, this should only increase expected utility. With a bias and/or distortion, this no longer holds. Such additional information can push a DM into a region where misjudging information leads to (further) suboptimal choices and is thus potentially even more costly. Information then has a strictly negative effect.

4.2 Beneficial moderation

A choice of action profile is suboptimal if it fails to realize the maximum value of an experiment. But the existence of a superior choice does not imply that a moderator can actually induce it. The decision maker's choice behaviour and signal perception constrain the influence of the moderator. This section formally identifies this constraint whose implications for the optimal moderation policy are traced out in the subsequent analysis.

Beneficial moderation requires disagreement between the decision maker and moderator regarding the expected utility ranking of an action that is chosen and one that could be induced. This comparison set differs significantly between naive and sophisticated decision makers. As a naive DM is unaware of any moderation, no interference by the moderator can alter the set of chosen actions. The moderator can only affect when these choices are executed, i.e. 'pick' among actions that are chosen by the DM after some signal. If the DM is sophisticated, the moderation policy can induce choices beyond those selected in the absence of moderation. However, the moderator cannot freely reassign chosen actions to signals.

While these relations can be complicated in a setting with many signals and actions, a first key observation is that binary relations play an important role. For a naive DM, an improvement is possible if and only if one action could be beneficially replaced by another, which is already part of the chosen action profile (Lemma 1). While the optimal moderation policy might affect more signals and actions, the existence of such a binary relation is a prerequisite. The situation is more complex for a sophisticated DM, who reacts to the moderation policy. Nevertheless, such a preference relation remains sufficient for beneficial moderation to be possible.

We say the moderator prefers an action profile $\boldsymbol{a}$ over some $\boldsymbol{a}'$, if it achieves higher expected utility. Consider an action profile $\mathbf{a} = (a_1, ..., a_l)$. For some distinct $i, j \in \{1, ..., l\}$, let $\mathbf{a}_{i \to j}$ be a modified $\mathbf{a}$, with a_j replaced by a_i, and all other actions unchanged.

Lemma 1. *Suppose a DM chooses an action profile* $\mathbf{a}$. *There (generically) exists a beneficial moderation policy*

- *for a naive DM, if and only if the moderator prefers* $\mathbf{a}_{i \to j}$ *to* $\mathbf{a}$ *for some* $i, j \in \{1, ..., l\}$,
- *for a sophisticated DM, if the moderator prefers* $\mathbf{a}_{i \to j}$ *to* $\mathbf{a}$ *for some* $i, j \in \{1, ..., l\}$.

Lemma 1 indicates that sophistication allows for beneficial moderation in more settings. This might lead one to (wrongly) conclude that sophistication generally increases the benefit from moderation. While it raises the possibilities for beneficial

moderation, it also curtails the extent of the moderation policy. A sophisticated decision maker takes into account the garbling and chooses the optimal action from their own perspective, i.e. given any distortions and biases. While the existence of some $\boldsymbol{a}_{i\to j}$ that is preferred by the moderator is generically sufficient[6] for both types, the optimal policy for a sophisticated agent might be limited in which actions can be implemented. Replacing a signal $\boldsymbol{s}_i$ entirely with some $\boldsymbol{s}_j$ does not necessarily induce action a_j. A moderation policy might fail to completely substitute one action for another. A moderator might only be able to switch signals with some probability, as the DM otherwise adjusts their action profile. However, this adjustment also allows the moderator to induce actions in $\mathbb{A}$ other than those already part of $\boldsymbol{a}$. Whether or not this benefits moderation depends on whether the alternative choices are superior from the perspective of the moderator. Lemma 2 summarizes this interdependence.

Lemma 2. *Suppose given an experiment X, distortion d, and priors $\boldsymbol{\mu}$, $\boldsymbol{p}$, a DM chooses an action profile* $\mathbf{a}$. *There exists a beneficial moderation policy for a sophisticated DM if and only if:*

(i) There exists an action profile $\boldsymbol{a}^*$ and garbling $M \neq I$ such that $V(XM|\boldsymbol{a}^*, \boldsymbol{\mu}) > V(X|\boldsymbol{a}, \boldsymbol{\mu})$, and

(ii) $\boldsymbol{a}^* = \arg\max_{a\in\mathbb{A}^{|S_X|}} V(X^d M|\boldsymbol{a}, \boldsymbol{p})$.

Lemmas 1 and 2 focus on the existence of beneficial moderation policies. As already mentioned, however, they are moot about the extent to which moderation can help each type. To do so, we address how misperception and biases affect the information contained in an experiment, and more specifically, how they influence the gain a DM realizes from it. Based on this, we characterize the optimal moderation policies and their expected utility outcomes.

4.3 Moderation & the gain from information

Moderation is a form of garbling and such such, any non-trivial moderation policy m renders an experiment (weakly) less Blackwell informative. While the moderator is aware of the true signals and thus reaches the same posterior whether or not there is moderation, the decision maker only has access to the moderated signals. Choices need to be evaluated according to XM. Moderation can only be beneficial if reducing the informativeness of signals increases the value of an experiment. When the maximum expected utility is convex in beliefs, this cannot be the case. The aim is thus to determine when biases and misperception, and more specifically the choices they

[6] Generically, in the sense that it only requires a strict preference for a_i over a_j at the belief $\boldsymbol{p}(\boldsymbol{s}_i^d)$. With countable actions, this is satisfied for almost all beliefs.

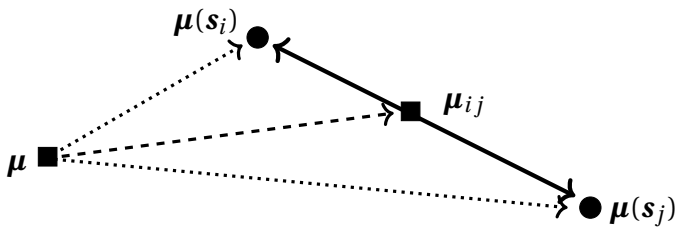


Figure 3: Starting from a prior $\boldsymbol{\mu}$, the experiment X results in a posterior $\boldsymbol{\mu}(\boldsymbol{s}_i)$ if signal $\boldsymbol{s}_i$ occurs (equivalently for $\boldsymbol{s}_j$). The solid line illustrates the 'relative information' between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Starting from $\boldsymbol{\mu}_{ij}$, experiment $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ leads to a posterior $\boldsymbol{\mu}(\boldsymbol{s}_i)$ or $\boldsymbol{\mu}(\boldsymbol{s}_j)$.

induce, create non-convexities (as described in Proposition 1) that can be addressed by garbling signals, and to characterize the optimal garbling.

It will again be useful not to look at the entire experiment, but to take a binary perspective and only consider the 'relative' information contained in two signals. One can think of this as the information that remains from the experiment, knowing that one of the two signals has occurred. Figure 3 provides a schematic representation. Denote by $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ the experiment generated from X, with a signal $\boldsymbol{s}_i \circ [\boldsymbol{s}_i + \boldsymbol{s}_j]^{-1}$ replacing $\boldsymbol{s}_i$, signal $\boldsymbol{s}_j \circ [\boldsymbol{s}_i + \boldsymbol{s}_j]^{-1}$ replacing $\boldsymbol{s}_j$, and all others equal $\mathbf{0}$. These hypothetical signals are a rescaling of $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ such that for each state, probability ratios are preserved, but probabilities sum to 1. The related intermediate (i.e., conditional) belief, knowing that either $\boldsymbol{s}_i$ or $\boldsymbol{s}_j$ has occurred, is denoted by $\boldsymbol{\mu}_{ij}$ (and $\boldsymbol{p}_{ij}$ from the perspective of the DM). Note that if $|S_X| = 2$, then this simply reduces to X, $\boldsymbol{\mu}$, and $\boldsymbol{p}$ respectively. Lemma 3 briefly justifies the relevance of this binary perspective in the context of utility maximization.

Lemma 3. *If the (conditional) gain from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ for some action profile $\boldsymbol{a}$ is negative at $\boldsymbol{\mu}_{ij}$, then $\boldsymbol{a}$ does not maximize expected utility at $\boldsymbol{\mu}$.*

Deterministic moderation

A failure to effectively use the relative information in two signals necessarily leads to a suboptimal choice. If this choice is 'sufficiently far' from the optimum, the gain from this relative information becomes negative. As Proposition 2 shows, these are exactly the instances when a deterministic moderation policy, i.e. one that destroys all relative information between at least some signals, has a a positive impact on expected utility. Furthermore, Corollary 2.1 establishes that for naive agents, these are the only instances where moderation is beneficial, meaning the optimal moderation policy is always deterministic.

Proposition 2. *There exists a beneficial deterministic moderation policy if and only if there are signals $\boldsymbol{s}_i, \boldsymbol{s}_j \in S_X$ such that:*

- *(naive DM) the conditional gain from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is negative at $\boldsymbol{\mu}_{ij}$.*

- *(sophisticated DM) the gain from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is negative at $\boldsymbol{\mu}_{ij}$.*

A deterministic moderation policy replaces one or several signals with another one from S_X. If, for instance, $\boldsymbol{s}_i$ is replaced with $\boldsymbol{s}_j$ (i.e., $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$), then after observing $\boldsymbol{s}_j$, an unbiased observer can only conclude that either of the signals has occurred, leading to a belief $\boldsymbol{\mu}_{ij}$. The information contained in $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is removed by the moderation policy, meaning $X^m_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is uninformative. A naive DM nevertheless reaches a posterior belief $\boldsymbol{p}(\boldsymbol{s}^d_j)$ and thus executes a_j after both of the actual signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. This is beneficial if the conditional gain from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is negative at $\boldsymbol{\mu}_{ij}$, which implies not just that the value from X is not maximized by $\boldsymbol{a}$, but that the DM is better-off if some information is removed entirely. While the optimal policy might remove the relative information from more than just two signals, a binary perspective is nevertheless sufficient to identify cases where optimal moderation is possible, which is the key aspect of Proposition 2. In contrast, a sophisticated DM is aware the removal of this relative information. Such a DM then chooses the optimal action at $\boldsymbol{p}_{ij}$ (the conditional belief knowing that either $\boldsymbol{s}^d_i$ or $\boldsymbol{s}^d_j$ has realized). This is beneficial if the gain from $X^m_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ at $\boldsymbol{\mu}_{ij}$ is negative. The key question for the moderator is thus whether this optimal choice at $\boldsymbol{p}_{ij}$ is also superior (to the original choices) at $\boldsymbol{\mu}_{ij}$.

A deterministic moderation policy is, of course, extreme in its effect on the information content of signals. Nevertheless, as Corollary 2.1 shows, the optimal moderation policy for a naive decision maker is always deterministic. In other words, for a naive DM, a negative gain from information is not just sufficient but also necessary for *any* beneficial moderation policy to exist.

Corollary 2.1. *The optimal moderation policy for a naive DM is deterministic.*

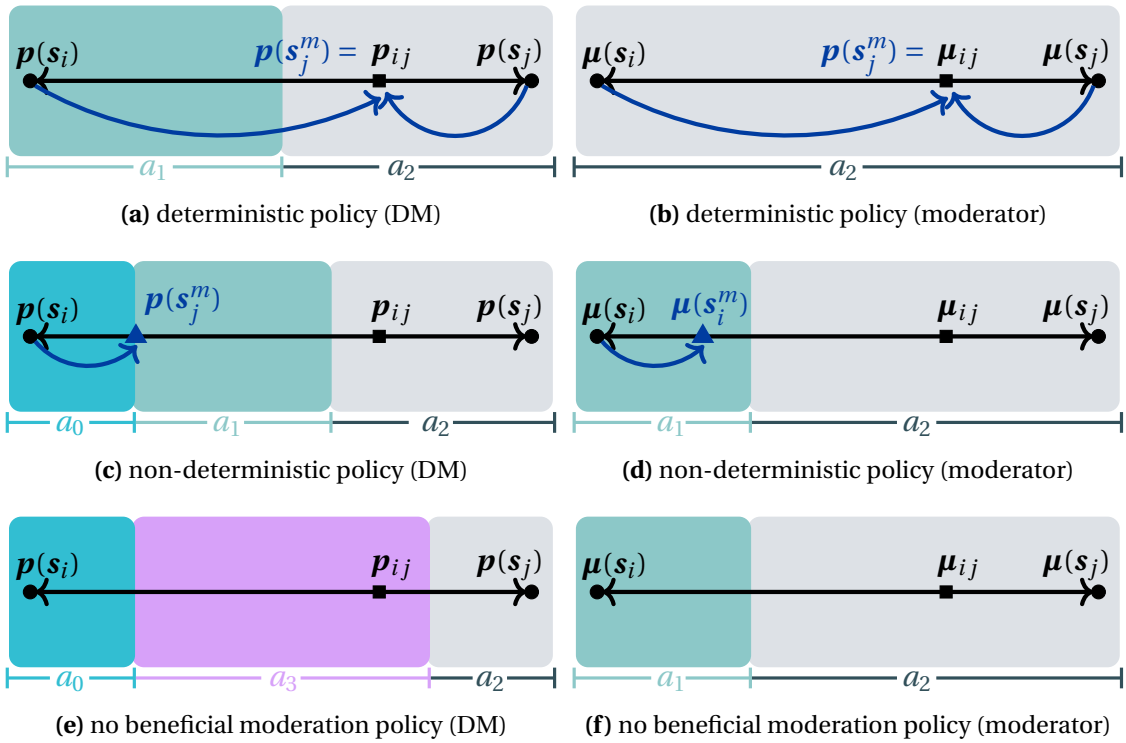
Sophistication & optimal moderation

For a sophisticated decision maker, a deterministic moderation policy can be equally optimal. Figure 4 (**a**), (**b**) schematically illustrates such a case. Panel (**a**) highlights which choices are optimal from the DM's perspective, and (**b**) what the moderator considers optimal. After a signal $\boldsymbol{s}_i$, the DM and moderator disagree about the optimal action: the moderator prefers action a_2 over a_1. Note that a_2 is also the preferred action by both moderator and DM at $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{p}_{ij}$ respectively. We can easily verify that the gain from information is negative. With a deterministic policy $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$ (or equivalently $m(\boldsymbol{s}_j) = \boldsymbol{s}_i$), the moderator eliminates the relative information between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Aware of this garbling, the DM can then only conclude from observing $\boldsymbol{s}^m_j$ that one of the two signals has occurred, but not which one. This leads the DM to take action a_2, which is preferred by the moderator. This is, however, not the unique optimal policy. The moderator could achieve the same outcome with a non-deterministic policy by randomly garbling both signals into each other, destroying (enough of) the underlying information. As was already foreshadowed by the discussion of Lemma 1,

a deterministic moderation policy might neither be the only nor the optimal way to beneficially influence the choices of a sophisticated decision maker.

Figure 4 (**c**), (**d**) depicts a scenario where a non-deterministic policy is uniquely optimal. By garbling some $\boldsymbol{s}_j$ signals into $\boldsymbol{s}_i$, the moderator can generate a posterior belief just close enough to $\boldsymbol{p}_{ij}$ that a_1 becomes optimal from the perspective of the DM. The posterior belief after observing $\boldsymbol{s}_j$ remains unaffected (even though the belief itself becomes less likely). Note that a deterministic moderation policy might also be beneficial here: if $E[u(a_2|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)] > E[u(a_0|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$, the gain from information at $\boldsymbol{\mu}_{ij}$ is negative, implying the existence of a beneficial deterministic policy. But since the moderator prefers a_1 to a_2 at $\boldsymbol{\mu}(\boldsymbol{s}_i^m)$, this cannot be optimal. In this particular case, a sophisticated DM benefits more from moderation than a naive one.

Finally, 4 (**e**), (**f**) demonstrates when there might be no beneficial moderation policy. If the moderator prefers a_0 to a_3 at $\boldsymbol{\mu}(\boldsymbol{s}_i)$ as well as $\boldsymbol{\mu}_{ij}$, then the gain from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is unambiguously positive at $\boldsymbol{\mu}_{ij}$. Any garbling that induces the DM to take a_3 after a signal $\boldsymbol{s}_i$ cannot be beneficial. If the moderator also prefers a_0 to a_2 at $\boldsymbol{\mu}(\boldsymbol{s}_i)$, then any garbling between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is suboptimal. There exists no beneficial moderation policy. Interestingly, if instead the moderator prefers a_2 to a_0 at $\boldsymbol{\mu}(\boldsymbol{s}_i)$, then for a sophisticated DM there exists a beneficial non-deterministic moderation policy but no deterministic one. In contrast, for a naive DM, the deterministic policy $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$ would not just be beneficial, but also achieve a strictly better outcome. A naive DM would be strictly better off.



(a) deterministic policy (DM)

(b) deterministic policy (moderator)

(c) non-deterministic policy (DM)

(d) non-deterministic policy (moderator)

(e) no beneficial moderation policy (DM)

(f) no beneficial moderation policy (moderator)

Figure 4: (In-)feasibility of beneficial moderation with a sophisticated DM

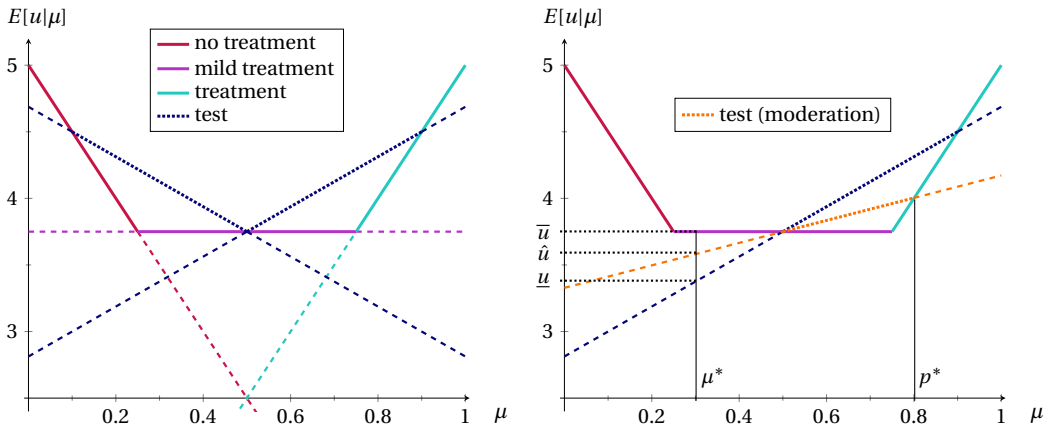
This discussion leads us to a conclusion about the optimal moderation policy that sharply differs from what we established for a naive DM. In contrast to Corollary 2.1, Proposition 3 shows that if a decision maker is sophisticated, a non-deterministic moderation policy always achieves a weakly better outcome. Furthermore, the optimal policy is uniquely non-deterministic in some cases.

Proposition 3. *Suppose the DM is sophisticated. Then the optimal moderation policy might not be deterministic. Furthermore, for every beneficial deterministic moderation policy, there exists a non-deterministic moderation policy that achieves weakly higher expected utility.*

Example 1.3 illustrates the differences in optimal moderation policies in a specific case and provides evidence for the mentioned ambiguous effect of sophistication, particularly when there is a bias in prior.

Example 1.3. Continuing with the previous example of diagnostic testing, suppose beyond the possibilities of treatment (a_I) or no treatment (a_N), a patient also has the option of a less aggressive (but somewhat less effective) treatment (a_M). Suppose $u(a_M|\omega_I) = u(a_M|\omega_N) = 3.75$, with all other payoffs as before. The previous action profile (a_I, a_N), which also has an expected payoff of 3.75, is no longer strictly optimal. However, the patient can still strictly benefit from the test by choosing a profile (a_N, a_M) for low/intermediate priors, and (a_M, a_I) for higher priors (see Figure 5 (**a**)). Now suppose the doctor concludes $\mu \in (0.25, 0.5)$, while the patient believes $p \in (0.5, 0.75)$. The doctor considers the aggressive treatment option strictly inferior to the intermediate one (at the prior and each posterior). The (conditional) gain from the test given the profile (a_M, a_I) is negative. This leaves the possibility to beneficially moderate the test result; either with a deterministic policy or with one that completely garbles both signals into white noise. For any such policy, a patient aware of the doctor's effort to obscure the result is then willing to resort to the more conservative treatment. In contrast, the optimal moderation policy for a naive patient is uniquely deterministic (always return a negative result). Nevertheless, the expected utility outcome is the same for both types. The doctor cannot, however, achieve the first best: no such patient (sophisticated or naive) is willing to forgo treatment completely after a negative test result for any moderation policy.

If the patient exaggerates the risk of infection even further ($p \in (0.75, 0.9)$), then the patient is still willing to take the test, but their preferred default (at the prior) is the aggressive treatment. If the patient is sophisticated, the optimal moderation policy must be non-deterministic (see Figure 5 (**b**)). It turns positive into negative test results with a just high enough probability, such that the patient is indifferent between (a_I, a_I) and (a_M, a_I). For priors μ^* and p^*, the optimal moderation policy yields $\hat{u}$ instead of



(a) no moderation **(b)** with moderation

Figure 5: Expected utility of action profiles (Example 1)

$\underline{u}$. For a naive patient, however, the optimal policy is such that all positive results are converted into negative ones. This yields $\overline{u}$. A naive patient strictly better off. ◇

As Figure 4 and Example 1.3 illustrate, an important aspect in determining whether sophistication poses a (dis-)advantage is whether or not the moderator an DM agree about the default action, i.e. the action taken when all relative information is removed. In Figure 4 (**e**) and (**f**), as well as in the Example 1.3 / Figure 5 (**b**), there is disagreement over which action is best if $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ (and its distorted counterpart) is uninformative. In contrast, Figure 4 (**c**) and (**d**) illustrates a setting where the moderator and DM agree on the default action. Since any disagreement over the default action requires $\boldsymbol{\mu}_{ij} \neq \boldsymbol{p}_{ij}$, a sophisticated DM can only be unambiguously better-off if there is no such discrepancy in conditional beliefs. Corollary 3.1 formalizes this latter observation.

Corollary 3.1. *If $\boldsymbol{\mu} = \boldsymbol{p}$, $|S_X| = 2$, and $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)] \geq E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$ for at least one of the signals $\boldsymbol{s}_i \in S_X$, then the optimal moderation policy for a sophisticated DM achieves (weakly) higher expected utility than that for a naive one.*

In a binary setting, sophistication proves an unambiguous advantage if priors are aligned. The adjustments in choices by the DM in response to moderation enhance the benefit from moderation. With only two signals and no bias in prior, the moderator and sophisticated DM trivially agree about the conditional belief $\boldsymbol{\mu}_{ij}$, since this corresponds to the prior. Accordingly, they agree about the best action in the absence of any information. If the distortion does not completely reverse the correlation between signals and states, meaning a DM is no better-off by switching both actions, then the optimal policy of a sophisticated decision maker achieves a (weakly) better outcome than that for a naive one.

While these conditions might appear restrictive, the characterisation is tight. Relaxing any of the three conditions can lead a sophisticated DM to be strictly worse-off.

The effect of sophistication become ambiguous if distortions and/or the information environment become more complex. With more than three signals, even if $\boldsymbol{\mu} = \boldsymbol{p}$, we can have $\boldsymbol{\mu}_{ij} \neq \boldsymbol{p}_{ij}$, since a distortion also affects the conditional belief. This can lead the moderator and decision maker to disagree over which is the best action when all relative information between signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is removed. Sophistication can then negatively affect the benefit from moderation. In this sense, a binary setting is not representative. For $\boldsymbol{\mu} \neq \boldsymbol{p}$, this disagreement over conditional beliefs is trivially possible. Finally, a distortion and/or bias in prior can cause actions to be chosen that a moderator would prefer to symmetrically swap. Section 4.4 explores this in detail and shows how this benefits naive DM (weakly) more than a sophisticated one. **Beliefs, choices, and beneficial moderation**

As mentioned in the beginning of this section, the scope of beneficial moderation hinges on whether a reduction in Blackwell informativeness can have a positive impact. This, in turn, depends on the relation between posterior beliefs, and more particularly, their implications on choices. This connection is now explored further to gain a better understanding of which settings lend themselves to disagreement between moderator and DM, and consequently, beneficial moderation.

If there are only two states of the world, any distortion can be described as either an over- or underestimation of signal strength. Furthermore, correlations between signals and states are either retained or reversed. For each $\boldsymbol{s}$, the DM updates 'too much' ($|\boldsymbol{\mu}(\boldsymbol{s}^d) - \boldsymbol{\mu}| > |\boldsymbol{\mu}(\boldsymbol{s}) - \boldsymbol{\mu}|$), or 'too little' ($|\boldsymbol{\mu}(\boldsymbol{s}^d) - \boldsymbol{\mu}| < |\boldsymbol{\mu}(\boldsymbol{s}) - \boldsymbol{\mu}|$), and possibly in the wrong direction. With $n > 2$ states, even if there are still only two signals, such a binary comparison of beliefs is no longer suitable, as the distorted signal can result in both 'too much' updating for some states and 'too little' for others. Nevertheless, we can still define a notion that captures the relevant effects of under- and overestimation of signal strength on choices, particularly when looking at the relative information in two signals, i.e. $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$.

Definition 2 (Misestimation of signal strength)**.** *For an experiment X, prior $\boldsymbol{\mu}$, and chosen action profile $\boldsymbol{a} = (a_1, ..., a_l)$, suppose there is some $a_i \neq \arg\max_{a \in \mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$. We say a_i is consistent with an **underestimation of signal strength** at $\boldsymbol{\mu}_{ij}$ if there exists an $\alpha \in [0,1)$, such that $a_i = \arg\max_{a \in \mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}_\alpha]$, for some $\boldsymbol{\mu}_\alpha = \alpha \cdot \boldsymbol{\mu}(\boldsymbol{s}_i) + (1-\alpha) \cdot \boldsymbol{\mu}_{ij}$. It is consistent with an **overestimation of signal strength** at $\boldsymbol{\mu}_{ij}$ otherwise.*

To see the relevance of Definition 2, note that whether moderation is possible depends not on the beliefs directly, but the actions they induce. The critical question becomes whether actions are consistent with an unambiguous reduction in (relative) informativeness (the distorted posterior lies on a straight line through some intermediate belief and undistorted posterior) or whether actions can only be rationalized with signals that contain additional information. Suppose a suboptimal action is cho-

sen after a signal $\boldsymbol{s}_i$. If this choice is optimal for a belief (of the moderator) that can be written as a convex combination of $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and some conditional belief $\boldsymbol{\mu}_{ij}$, then we say it is consistent with a underestimation of signal strength at $\boldsymbol{\mu}_{ij}$ (or simply *relative* underestimation). The choices can be rationalized with a $X_{\triangle}(\boldsymbol{s}_i^d, \boldsymbol{s}_j^d)$ that is less informative than $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$. Note that this definition also implies that correlations between states and signals are not completely reversed, since this would imply that the relative information in the distorted signal is inconsistent with that of the undistorted one. Since choices must be optimal at some belief, any other choice implies that the DM wrongly believes the signal contains some additional information (at least relative to some signal). In this case, we say choices are consistent with a (relative) overestimation of signal strength at at $\boldsymbol{\mu}_{ij}$.

Proposition 4 establishes that the latter is a requirement for beneficial moderation to be feasible. It is subsequently demonstrated that the 'relative' perspective is crucial.

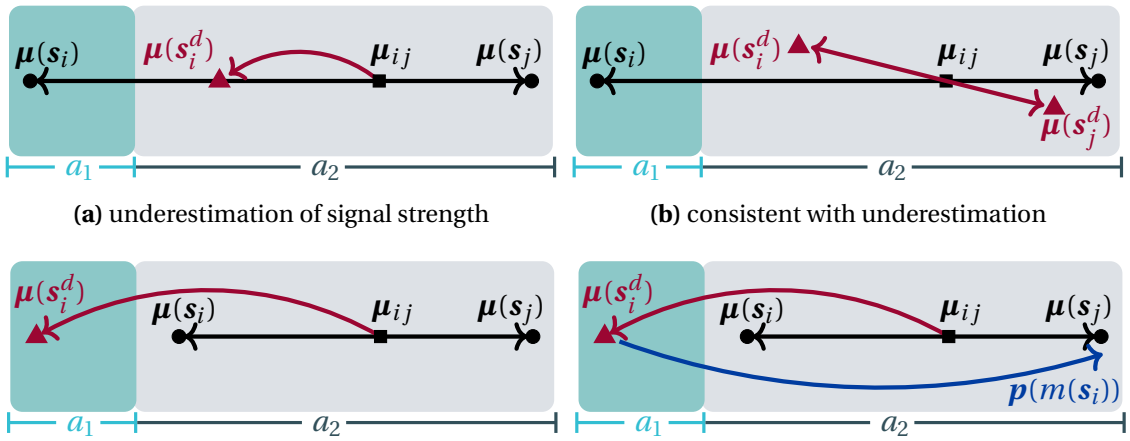
Proposition 4. *For any distortion d and priors $\boldsymbol{\mu}$, $\boldsymbol{p}$, there exists a beneficial moderation policy only if there is a choice a_i in $\boldsymbol{a}$ and signals $\boldsymbol{s}_i, \boldsymbol{s}_j \in S_X$ such that a_i is consistent with an overestimation of signal strength at $\boldsymbol{\mu}_{ij}$.*

Figure 6 visualizes some key cases. A relative underestimation of the signal strength of $\boldsymbol{s}_i$ can lead to a suboptimal choice (action a_2), which cannot be improved upon with moderation (**a**). It is irrelevant whether the signal strength is actually underestimated or choices are merely consistent with such an underestimation (**b**). Notice how (**b**) is not simply an underestimation of relative signal strength. The posteriors for the distorted signals are pointing in a different direction than those for the undistorted signals, which indicates that they contain additional (or different) information with regards to some state. When the choice after $\boldsymbol{s}_i$ is not consistent with some posterior between $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and $\boldsymbol{\mu}_{ij}$, the DM must overestimate the informativeness of the signal in at least some direction (**c**). In this case, beneficial moderation might be possible (**d**), e.g., the moderated-signal restores action a_2.

As an immediate implication, in a binary setting with only two states and two signals, any distortion that leads to an underestimation of signal strength (without reversing the correlation between signals and states)[7] does not allow for beneficial moderation.

Corollary 4.1. *Suppose $|\Omega| = |S_X| = 2$ and $\boldsymbol{\mu} = \boldsymbol{p}$. Then for any distortion that causes an underestimation of signal strength without reversing the correlation between signals and states, there exists no beneficial moderation policy.*

[7] Formally, we say a distortion causes an underestimation of signal strength without reversing the correlation between signals and states if for all $\boldsymbol{s} \in S_X$ and $\omega \in \Omega$, $|\boldsymbol{\mu}(\boldsymbol{s}_i^d) - \boldsymbol{\mu}| \leq |\boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}|$ with $\mu_\omega \in [\mu_\omega, \mu_\omega(\boldsymbol{s})]$ (or $[\mu_\omega(\boldsymbol{s}), \mu_\omega]$).



(a) underestimation of signal strength

(b) consistent with underestimation

(c) overestimation of signal strength

(d) beneficial moderation (naive DM)

Figure 6: (In-)feasibility of beneficial moderation

As moderation implies a 'destruction' of information, is seems unsurprising that it cannot be helpful in cases where the informativeness of a signal is already underestimated. However, with $|S_X| > 2$ signals, this argument is not as straight-forward. Suppose the distortion is simply a (perceived) garbling between a signal $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. The informativeness of $\boldsymbol{s}_i$ at $\boldsymbol{\mu}_{ij}$ is underestimated. However, beneficial moderation might still be possible. The distortion generally causes the DM's intermediate belief $\boldsymbol{p}_{ik}$ to differ from $\boldsymbol{\mu}_{ik}$ for a signal $\boldsymbol{s}_k \neq \boldsymbol{s}_i, \boldsymbol{s}_j$. At $\boldsymbol{\mu}_{ik}$, the distortion might then be inconsistent with an underestimation of signal strength, since $\boldsymbol{s}_i^d$ contains some information from $\boldsymbol{s}_j^d$. Proposition 4 only rules out beneficial moderation if a distortion is consistent with underestimation of signal strength relative to all other signals. Again, a binary setting is of limited representativeness. Example 2.1 illustrates the point.

Example 2.1. Suppose there are three states, $\{\omega_1, \omega_2, \omega_3\}$, with each state is equally likely ex ante. There are two actions, a_1 and a_2. Payoffs are such that $u(a_1, \omega_1) \geq u(a_1, \omega_2) > u(a_1, \omega_3)$, and $u(a_2, \omega_1) = u(a_2, \omega_3) > u(a_2, \omega_2)$. Suppose further the experiment X is such that there are three symmetric signals $S_X = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$, with $\boldsymbol{s}_i$ having a higher probability in state ω_i than in the other two states (and with equal probability in both other states). As can be seen from Figure 6, a_1 is optimal after signals $\boldsymbol{s}_1$, and $\boldsymbol{s}_2$, while a_3 is optimal after $\boldsymbol{s}_3$.[8] Now suppose the (naive) DM instead perceives a distorted experiment. Each signal is symmetrically garbled with the other two such that $\boldsymbol{s}_i^d$ still has higher probability in state ω_i, but the likelihood ratios relative to the other states reduced. The DM would then prefer a_2 after signal $\boldsymbol{s}_1$, with all other choices unaltered. As X^d is a garbling of X, the DM clearly underestimates the informativeness

[8] The example is based on the following numerical values: Payoffs are such that $u(a_1, \omega_1) = 10, u(a_1, \omega_2) = 5, u(a_1, \omega_3) = 0$, and $u(a_2, \omega_1) = u(a_2, \omega_3) = 9, u(a_2, \omega_2) = 0$. The probability of observing $\boldsymbol{s}_i$ in state ω_i is $8/10$, and $1/10$ in each of the other states.

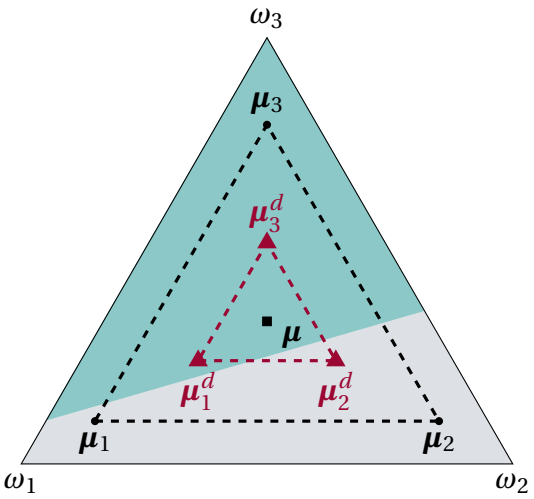


Figure 7: Posteriors of three distorted and undistorted signals in a belief simplex (Example 2). Posteriors after a signal $\boldsymbol{s}_i$ are denoted by $\boldsymbol{\mu}_i$. The DM underestimates the signal strength of every signal. Beneficial moderation remains possible as a_2 is not consistent with an underestimation of $\boldsymbol{s}_1$ relative to $\boldsymbol{s}_2$.

of the entire experiment.[9] The posterior beliefs of the DM are contained in the convex hull of the posteriors of the moderator. Nevertheless, this is not consistent with a relative underestimation of signals. Action a_2 is not optimal for any belief in the set $\{\alpha \cdot \boldsymbol{\mu}(\boldsymbol{s}_2) + (1-\alpha) \cdot \boldsymbol{\mu}_{ij},\ \alpha \in [0,1]\}$. And in fact, a beneficial moderation policy exists: $m(\boldsymbol{s}_1) = \boldsymbol{s}_2$, and m equal the identity mapping otherwise. ◇

Example 2.1 leads to an interesting conclusion: with more than two states, a moderator can beneficially destroy information, even if the decision maker already (strictly) underestimates the informativeness of all signals. Amplifying misperception can reduce the utility loss.

The previous results where mostly concerned with a strict reduction in informativeness of signals. For a naive DM, this implies rather 'heavy handed' moderation policies. The optimal garblings are such that all relative information between affected signals is destroyed. In fact, for all signals that are garbled into each other, a naive DM reaches the same posterior (which corresponds to one of the posteriors in the absence of moderation). The DM is being misinformed in at least some cases. Interventions for a sophisticated DM are (weakly) less aggressive. As sophisticated DMs also take into account the reduced informativeness, posteriors are less distorted (relative to the beliefs the DM would reach otherwise). Nevertheless, there are cases where the complete destruction of information is more beneficial than a partial one. And furthermore, it can be more effective for a naive than a sophisticated DM - implying that a more dis-

[9] The distorted experiment is obtained by right-multiplying X with the garbling matrix $\begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}$.

torted posterior can be beneficial. As follows from the discussion of Corollary 3.1, this requires more than two signals or a bias in prior.

The remaining analysis is concerned with one other possibility: cases where the optimal moderation policy leaves more information to a naive decision maker and yet renders this DM completely misinformed (but potentially better-off). This again requires that the information environment is non-binary, lending further support to the conclusion that, as the number of states and signals increases, the scope for moderation does as well, while simultaneously rendering the effects of sophistication more ambiguous.

4.4 Complete disagreement

The final part of the analysis turns to a potentially counterintuitive and yet particularly instructive case: the moderator and decision maker 'completely' disagree about which action should be taken after which signal. With *complete disagreement*, we mean that a moderator believes an action a should follow a signal $\boldsymbol{s}$, and action b a signal $\boldsymbol{t}$, with the decision maker holding the completely opposite view. This creates an incentive for the moderator to fully misinform a (naive) decision maker by manipulating each signal and thus inducing 'reversed' posteriors. This would be hardly surprising in a sender-receiver game when preferences are not aligned but here the 'sender' (i.e. moderator) and receiver (i.e. DM) hold identical preferences. Of course, if a signal distortion were to completely reverse the information content of signals, this would be equally trivial. But as will be made precise, complete disagreement can occur with distortions that leave moderator and DM, at least in principle, in agreement about the qualitative information content of signals. In fact, we show that such disagreement can arise in settings without any distortions at all, and instead be the result of a bias in prior. We fully characterise when complete disagreement can occur and provide a method for verifying if such a possibility exists in a given setting. We then show the implications for the optimal moderation policy.

Let $\boldsymbol{a}_{i\leftrightarrow j}$ denote an action profile identical to $\boldsymbol{a}$, except action a_i is replaced with a_j and vice versa. In other words, while $a_{i\to j}$ denotes a single substitution, $\boldsymbol{a}_{i\leftrightarrow j}$ describes a symmetric one.

Definition 3. *Given a chosen action profile $\boldsymbol{a}$, a moderator and DM are in complete disagreement if there exists $\boldsymbol{a}_{i\leftrightarrow j}$ such that*

$$V(X|\boldsymbol{a}_{i\leftrightarrow j},\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu}).$$

and

$$V(X|\boldsymbol{a}_{i\leftrightarrow j},\boldsymbol{\mu}) > \max\{V(X|\boldsymbol{a}_{i\to j},\boldsymbol{\mu}), V(X|\boldsymbol{a}_{j\to i},\boldsymbol{\mu})\}.$$

We say they are in complete disagreement over a_i and a_j.

Lemma 4 fully characterizes geometrically, when a given information environment allows for complete disagreement.

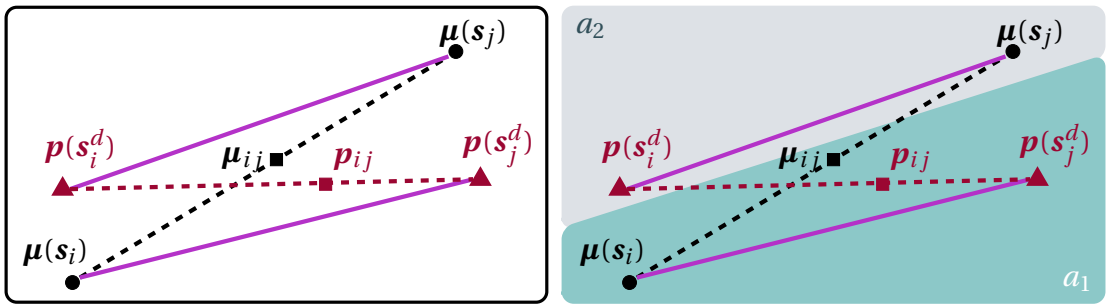
Lemma 4. *Given an experiment X, distortion d, and priors $\boldsymbol{\mu}$ and $\boldsymbol{p}$, there exists preferences such that the DM and moderator are in complete disagreement if and only if for some $\boldsymbol{s}_i, \boldsymbol{s}_j \in S_X$, the line segments between $\boldsymbol{\mu}(\mathbf{s}_i)$ and $\boldsymbol{p}(\mathbf{s}_j^d)$, as well as $\boldsymbol{\mu}(\mathbf{s}_j)$ and $\boldsymbol{p}(\mathbf{s}_i^d)$, do not intersect.*

Intuitively, complete disagreement becomes a possibility if a distortion and/or bias in prior create a sufficient rotation between the posterior beliefs of the moderator and the DM, i.e. if the direction of the updating is not sufficiently aligned between them. Figure 8 schematically depicts the two cases. In (**a**), the distortion and bias lead to a clockwise rotation of the posterior beliefs relative to the moderator's. The relevant line segments (solid lines) do not cross. As shown in (**b**), there are preferences such that the beliefs $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and $\boldsymbol{p}(\boldsymbol{s}_j^d)$ lie on one side of the indifference curve, and $\boldsymbol{\mu}(\boldsymbol{s}_j)$ and $\boldsymbol{p}(\boldsymbol{s}_j^d)$ on the other. There is complete disagreement. But this does not stem from a reversed correlation between signals. Signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_i^d$ lead to a qualitatively comparable updating of beliefs, indicating that they provide evidence towards the same states. The alternative scenario is depicted in (**c**), where the equivalent line segments cross. Here, there is also a clockwise rotation of beliefs, but this is small relative to the (vertical) shift in beliefs. (**d**) depicts a particular example where there is no complete disagreement and (at least between these two signals) no beneficial moderation. While the latter consequence is specific to the example, there are no preferences that would lead to complete disagreement in this case.

The characterization in Lemma 4 is necessary and sufficient, but such a geometric condition is not always easy to interpret or verify, particularly if the state space contains more than three states. To provide a more convenient approach for analysing complete disagreement, we utilise the following idea: the magnitude and direction of an update of beliefs can be described by a vector. How updating differs between a DM and moderator is then reflected by differences between corresponding vectors. We can identify necessary and sufficient conditions for complete disagreement based on the set of these vectors and the space they span.

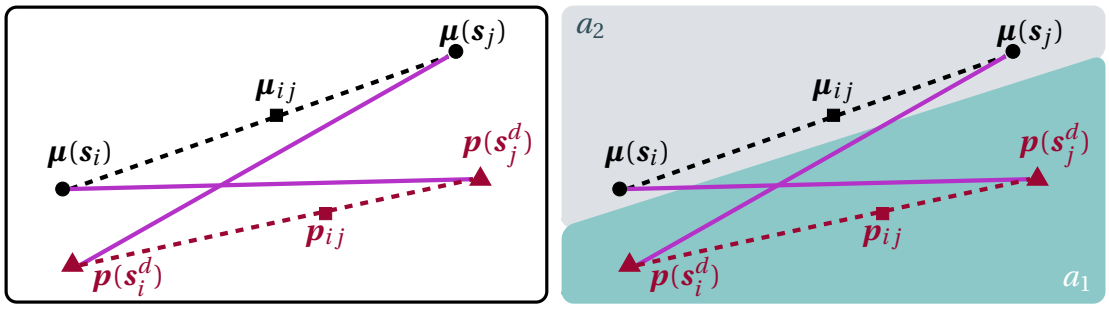
For a given experiment, distortion, and bias, all possible posterior beliefs (both for the moderator and DM) can be described by elements of a vector space that originates at some prior belief. We refer to this as the 'belief space'. As a convention, we use $\boldsymbol{\mu}$ as a reference (i.e. origin), even though the DM's prior $\boldsymbol{p}$ could equally be used.

Definition 4 (Belief space)**.** *The **belief space** of an experiment X relative to $\boldsymbol{\mu}$, given biased prior $\boldsymbol{p}$ and distortion d, is the linear space spanned by the vectors*



(a) line segments do not intersect

(b) complete disagreement



(c) line segments intersect

(d) no complete disagreement

Figure 8: Possibility of complete disagreement

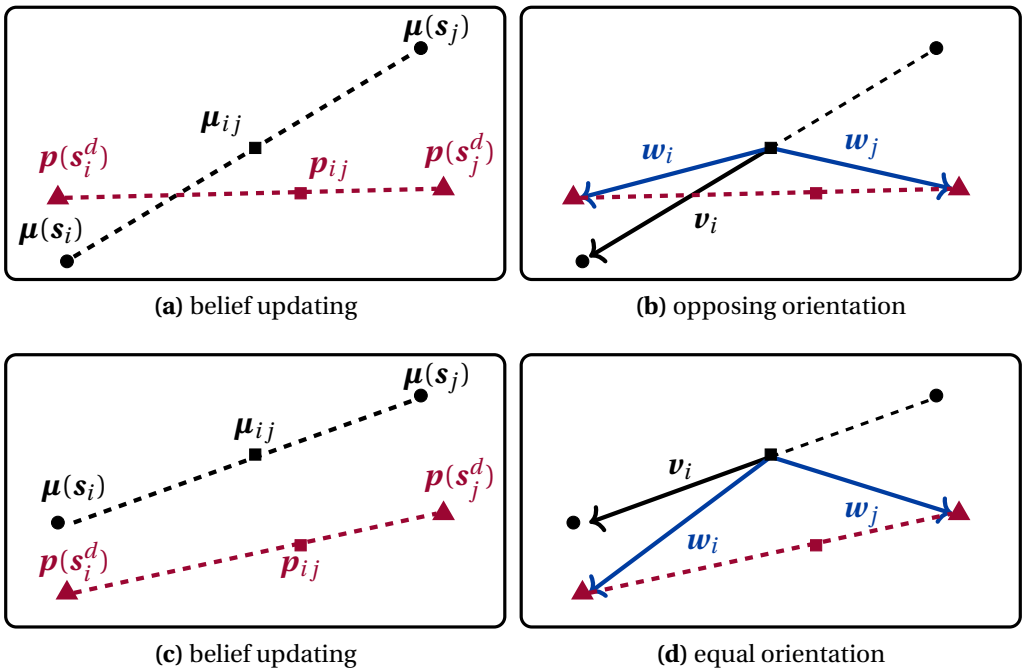
$$\{\boldsymbol{v}_1, ..., \boldsymbol{v}_l, \boldsymbol{w}_1, ..., \boldsymbol{w}_l\}$$

where $\boldsymbol{v}_i = \boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}$ *and* $\boldsymbol{w}_i = \boldsymbol{p}(\boldsymbol{s}_i) - \boldsymbol{\mu}$.

When referring to vectors in such a belief space, we adopt the notational convention that $\boldsymbol{v}_i = \boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}$ and $\boldsymbol{w}_i = \boldsymbol{p}(\boldsymbol{s}_i^d) - \boldsymbol{\mu}$. These vectors also allow us to formalize the notion of a DM not completely misjudging the correlation between signals and states: we say beliefs satisfy *non-reversal*, if $v_i \neq -\alpha w_i$ for any $\alpha > 0$ and all $i \in \{1, ..., l\}$, i.e. the direction of the update is not fully reversed. Naturally, the definitions for belief space and non-reversal can be applied to any $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$. As Theorem 1 shows, this binary perspective again provides a convenient simplification. The dimensions of the belief space for a given $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ has direct implications for the possibility for complete disagreement. Beyond the dimension of this vector space, the introduction of an additional property (Definition 5) allows for a full characterization.

Definition 5 (Opposing orientation)**.** *Let* $\boldsymbol{v}$*,* $\boldsymbol{w}$*, and* $\boldsymbol{x}$ *be vectors in a 2-dimensional vector space. We say* $\boldsymbol{v}$ *and* $\boldsymbol{w}$ *have opposing orientation relative to* $\boldsymbol{x}$ *if the sets* $\{\boldsymbol{x}, \boldsymbol{v}\}$ *and* $\{\boldsymbol{x}, \boldsymbol{w}\}$ *both span and have different orientation, meaning the unique linear transformation* L*, with* $\{\boldsymbol{x}, \boldsymbol{w}\} = \{L\boldsymbol{x}, L\boldsymbol{v}\}$*, is such that* $\det(L) < 0$.

Simply put, two vectors satisfy *opposing orientation* if they point to a different side relative to a third vector. This formalizes the notion of a 'sufficient rotation' in posterior beliefs, which is the basis for complete disagreement. Take, for instance, an



(a) belief updating **(b)** opposing orientation

(c) belief updating **(d)** equal orientation

Figure 9: Opposing orientation illustrated

experiment $X_\triangle(s_i, s_j)$. The belief update from some μ_{ij} to $\mu(s_i)$, as shown in Figure 9 (**a**), can be described by the vector $v_i = \mu(s_i) - \mu_{ij}$. For the DM, the same signal (perceived distortedly) leads to the posterior $p(s_i^d)$. In the belief space, this is captured by the vector $w_i = p(s_i^d) - \mu_{ij}$. The equivalent is true for s_j. If w_i and w_j satisfy opposing orientation relative to v_i, then - loosely speaking - one points to the left of v_i and the other to the right. If we think of a hypothetical line through $\mu(s_i)$ and $\mu(s_j)$ (and thus also μ_{ij}), then $p(s_i^d)$ must be on one side, and $p(s_j^d)$ on the other. This is the case in (**b**). Vectors w_i and w_j have opposing orientation relative to v_i. In a sense, the distortion introduces perceived information that is not just orthogonal to the experiment $X_\triangle(s_i, s_j)$, but that acts in opposing directions on the posteriors after s_i^d and s_j^d. This creates a rotation in the posterior beliefs relative to μ_{ij}. Panel (**c**) shows a similar case, but here the rotation is small compared to the shift in beliefs. Vectors w_i and w_j have the same orientation relative to v_i.

Figure 9 depicts two cases already illustrated in Figure 8. The constellation of posterior beliefs in (**a**) allows for complete disagreement, while no such disagreement is possible in (**c**). In the former case, illustrated in Figure 9 (**b**), w_i and w_j have opposing orientation. In the latter case, the orientation relative to v_i is the same (see (**d**)). This points towards the key role of orientation for complete disagreement. Theorem 1 formalizes this relation and completely characterizes (non-trivial) cases of complete disagreement solely with properties of the set of vectors representing posterior beliefs.

Recall that for a belief space of an experiment $X_\triangle(s_i, s_j)$ relative to μ_{ij}, we denote

by $\boldsymbol{v}_i$ the vector (in that space) corresponding to $\boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}_{ij}$. And by $\boldsymbol{w}_i$ the vector corresponding to $\boldsymbol{p}(\boldsymbol{s}_i^d) - \boldsymbol{\mu}_{ij}$.

Theorem 1. *Given an experiment X, priors $\boldsymbol{\mu}$, $\boldsymbol{p}$, and distortion d, suppose the belief space of $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ relative to $\boldsymbol{\mu}_{ij}$ has dimension z. Let $\boldsymbol{a} = (a_1, ..., a_l)$ be the DM's chosen action profile.*

- *If $z = 1$ and beliefs satisfy non-reversal, there cannot be complete disagreement over a_i and a_j for any preferences.*
- *If $z = 2$, there exist preferences such that there is complete disagreement over a_i and a_j if and only if one of the following applies:*
  - *(i) $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ have the opposing orientation property relative to $\boldsymbol{v}_i - \boldsymbol{v}_j$, or*
  - *(ii) $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ have the opposing orientation property relative to $\boldsymbol{w}_i - \boldsymbol{w}_j$, or*
  - *(iii) $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ have the opposing orientation property relative to $\boldsymbol{\mu}_{ij} - \boldsymbol{p}_{ij}$.*
- *If $z = 3$, there always exist preferences such that there is complete disagreement.*

If the belief space of some $X_\Delta(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is 1-dimensional, then beliefs cannot be rotated relative to each other. Hence, there cannot be complete disagreement. Except, of course, if a distortion completely reverses the correlation between signals and states. Here, such a 180° rotation is ruled out by non-reversal. If this assumption were relaxed, then a 1-dimensional space allows for complete disagreement in these (trivial) cases. All other results relating to higher dimensions would remain unaffected. As an immediate implication, non-trivial cases of complete disagreement require Ω to contain at least three states (Corollary 4.2).

Corollary 4.2. *Under non-reversal, there can be complete disagreement only if $|\Omega| \geq 3$.*

With only two states, the belief space of any experiment can only be 1-dimensional, ruling out (non-trivial) rotations. In a 2-dimensional space belief space, which requires $|\Omega| \geq 3$, such a rotation is possible (but not necessary). Hence complete disagreement is a possibility in some cases. These are captured by the property of opposing orientation which is necessary and sufficient for complete disagreement to be possible. To easily verify opposing orientation, the appendix (see Section A.2) provides a method that relies solely on the determinant of 2-dimensional matrices constructed from the relevant vectors. If $|\Omega| > 3$, the belief space can be 3-dimensional (the highest possible dimension, given that there are at most three linearly independent vectors). If $z = 3$, each of the four posteriors is necessarily rotated out of the plane spanned by the other three. This is sufficient to guarantee the possibility for complete disagreement.

Moderating complete disagreement

Since Bernoulli utilities are aligned between decision maker and moderator, complete disagreement is based on a different understanding how information is to be interpreted. The preferred choices of the DM and moderator are as if correlations between states, and some signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, are completely reversed. With $|\Omega| \geq 3$, however, an actual reversal is not required. In fact, a small distortion that leads to a rotation of beliefs is sufficient. Alternatively, again for $|\Omega| \geq 3$, this can arise from differences in prior beliefs without any distortion present. In both cases, choices are not consistent with an underestimation of signal strength. But with complete disagreement, beneficial moderation does not just become a possibility (as in Proposition 4), it is generically possible for naive and sophisticated DMs (Proposition 5). Most importantly, while complete disagreement implies a negative (conditional) gain from some experiment $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ at $\boldsymbol{\mu}_{ij}$ (Lemma 5 in Appendix A.1), it also implies that an alternative action profile can be constructed from the choices in $\boldsymbol{a}$ that achieves strictly positive conditional gain. This profile achieves strict (local) convexity in posterior beliefs. The moderator does not want to destroy relative information between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, but rather change its interpretation.

For a naive DM, such a reinterpretation is possible through a relabeling of signals, e.g. with a moderation policy m such that $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$ and $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$. No information is destroyed since $s_i^m = s_j$ and $s_i^m = s_j$. Experiments X and X^m are equally (Blackwell) informative, and yet the naive DM ends up completely misinformed in terms of posterior beliefs. For a sophisticated decision maker, such a relabeling is not feasible. The DM would simply swap the labels again. Sophistication constrains the moderator and forces a moderation policy that destroys relative information. If no better alternative choice can be induced (e.g. if $V(X|\boldsymbol{\mu}) = V(X|\boldsymbol{a}_{i\leftrightarrow j}\boldsymbol{\mu})$), then sophistication poses a binding constraint to the moderator. Given the respective optimal moderation policies, a naive DM might be strictly better off (Corollary 5.1). While sophistication makes a decision maker harder to manipulate, it also limits beneficial interventions when prior beliefs deviate - to the detriment of the decision maker.

Proposition 5. *Suppose given a chosen action profile $\boldsymbol{a}$, the moderator and DM are in complete disagreement over some a_i and a_j. Then there generically exists a beneficial moderation policy for both a naive, and a sophisticated DM.*

Suppose there is complete disagreement over some actions a_i and a_j, in the sense that a moderator would prefer a_i after $\boldsymbol{s}_j$, and equivalently for a_j. Then if these choices do not correspond to the utility-maximizing actions at the undistorted posterior beliefs (i.e. a_j is not utility maximizing at $\boldsymbol{\mu}(\boldsymbol{s}_i)$), sophistication can be an advantage. The 'constraint' imposed by sophistication, as laid out in Lemma 2, can potentially

lead to a better choice. If, however, both actions are nevertheless optimal (meaning from the perspective of the moderator, they are simply chosen after the wrong signal), then sophistication is a liability. If the information provided by at least one of the corresponding signals is also strictly valuable relative to any other signal, meaning $V(X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_k)|\boldsymbol{\mu}_{ik}) > \max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}_{ik}]$ for all $\boldsymbol{s}_k \neq \boldsymbol{s}_i$, then sophistication causes a strict loss in expected utility. [10] In both cases, simply (symmetrically) swapping signals implements the first-best for a naive DM. Corollary 5.1 formalizes this for the particular case described, i.e. $\boldsymbol{a}_{i\leftrightarrow j}$ is the (unconstrained) optimal action profile and a_i or a_j is unique in $\boldsymbol{a}$.[11]

Corollary 5.1. *Suppose given a chosen action profile $\boldsymbol{a} = (a_1, ..., a_l)$, we have $V(X|\boldsymbol{\mu}) = V(X|\boldsymbol{a}_{i\leftrightarrow j}\boldsymbol{\mu})$, and either a_i or a_j not equal to any other $a_k \in \{a_1, ..., a_l\}$. Then the optimal moderation policy for a naive DM generically achieves strictly higher expected utility than the optimal policy for a sophisticated DM.*

This allows a complementarity perspective to Corollary 3.1, which established that if signals are binary ($|S_X| = 2$), priors are aligned, and there is no complete disagreement, sophistication cannot be detrimental. As can be easily verified, if the choice environment is also binary ($|A| = 2$), then moderation achieves the same expected utility whether the DM is naive or sophisticated. In contrast, when both choices and signals are binary, but there is complete disagreement, sophistication is a strict disadvantage:

Corollary 5.2. *Suppose $|S_X| = |A| = 2$ and there is complete disagreement. Then the optimal moderation policy for a naive DM achieves strictly higher expected utility than the optimal policy for a sophisticated DM.*

Complete disagreement follows from a sufficiently distinct interpretation of signals between moderator and decision maker. Proposition 6 shows that it does not require any distortion at all, but can occur with only a bias in prior. Moreover, if $|\Omega| \geq 3$ and signals have distinct probability ratios across states, then a bias in prior that can lead to complete disagreement necessarily exists. Furthermore, we can find such a bias with the difference in beliefs only ϵ-small (Lemma 6,).

Proposition 6. *Let $|\Omega| > 2$ and suppose X is an experiment with two non-identical signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ that have distinct probabilities for at least 3 states. Then there exist preferences and a pair of prior beliefs $\boldsymbol{\mu}, \boldsymbol{p} \in \Delta(\Omega)$ such that there is complete disagreement.*

Example 3.1 demonstrates a case where complete disagreement arises (only) from a difference in priors. Furthermore, in line with Corollary 5.2, it illustrates another

[10] Note that if several signals induce the same action, then the information provided by the corresponding signals is not strictly valuable, in the sense that there is a garbled X that yields the same choices and expected utility.

[11] This would, for instance, be the case if there is complete disagreement and $|S_X| \leq 3$.

instance where a naive decision maker benefits strictly more from moderation.

Example 3.1. Suppose a firm is considering an applicant for a position and can either hire them (a_H) or not (a_N). The firm hires an outside HR consultancy to conduct an assessment and provide a recommendation. They can ask the applicant to take an assessment test on which the candidate can score highly (signal $\boldsymbol{s}$) or poorly (signal $\boldsymbol{t}$). There are three states: the candidate is highly skilled and experienced at taking assessment tests (ω_1), highly skilled and inexperienced at taking tests (ω_2), or does not have the required skills (ω_3). The assessment (X) is such that a positive result ($\boldsymbol{s}$) is most likely if the candidate is skilled and experienced at tests, but inexperienced applicants perform poorly on average. Probabilities are as follows:

$$\boldsymbol{s} = \begin{pmatrix} s_{\omega_1} \\ s_{\omega_2} \\ s_{\omega_3} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{2} \end{pmatrix} \qquad \boldsymbol{t} = \begin{pmatrix} t_{\omega_1} \\ t_{\omega_2} \\ t_{\omega_3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{1}{2} \end{pmatrix} \qquad X = \begin{pmatrix} | & | \\ \boldsymbol{s} & \boldsymbol{t} \\ | & | \end{pmatrix}.$$

The firm's payoff only depends on the candidate's skill, not their test-taking ability. Hiring a skilled candidate yields a payoff of 1, not hiring an a candidate yields 0, and hiring an unskilled candidate yields -2. Based on the applicant's profile, both the firm as well as the HR consultancy assign probability 0.7 to the candidate being skilled. But while the firm believes the candidate to be skilled but inexperienced with assessments ($\boldsymbol{p}$), the consultancy is under the impression that the candidate is experienced with tests ($\boldsymbol{\mu}$). The priors are:

$$\boldsymbol{\mu} = \begin{pmatrix} 0.6 \\ 0.1 \\ 0.3 \end{pmatrix} \qquad \boldsymbol{p} = \begin{pmatrix} 0.1 \\ 0.6 \\ 0.3 \end{pmatrix}.$$

Figure 10 visualizes the possible posterior beliefs and payoff-maximizing actions in the belief simplex. Firm and consultancy agree that hiring (a_H) is the optimal course of action in the absence of any test. However, the test leads to complete disagreement. The firm interprets a negative result as (further) evidence for a skilled but inexperienced test taker and would still prefer to hire. The consultancy, however, takes a negative result for face value and sees it as evidence for an unskilled applicant. The reactions to a positive test result are necessarily symmetrically opposed.

The HR consultancy has an incentive to misinform the firm about the test result: delivering a negative result leads the firm to hire the candidate (as the firm then believes sufficiently strongly in their limited experience with tests). Equivalently, a positive result induces action a_N, the preferred course of action from the perspective of the consultancy after a poor test performance. Reversing the test results ($m(s) = t$,

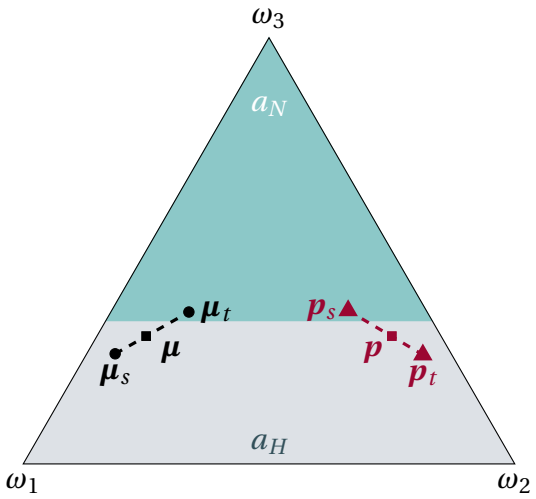


Figure 10: Prior and posterior beliefs of physician (moderator) and patient (DM). The diagnostic test induces complete disagreement.

$m(t) = s$) leads to the first-best, or at least maximizes expected utility from the consultancy's point of view. If the consultancy is employed by a firm aware of any such tempering, then the best course of action is to fully garble the outcome (i.e., not perform the test) despite considering the test results as valuable information. As in Example 1.3, a sophisticated client is strictly worse-off than a naive one. ◇

Maybe surprisingly, complete disagreement arises despite the agreement over the likelihood of facing a skilled applicant, and the identical course of action whether the candidate is experienced or inexperienced at assessment tests. States ω_1 and ω_2 are payoff equivalent. The example thus also highlights that - since the dimension of the belief space is crucial for the possible differences in beliefs and actions - combining seemingly identical states is not without loss when modelling such decision problems.

5 Conclusion

We analyzed the effects of two fundamental mistakes in information processing and how they can be mitigated by a moderator or gatekeeper that has a better understanding of the information environment (i.e., does not suffer from any mistakes). Even though preferences between this moderator and the decision maker are assumed to be aligned, a decision maker can be better-off if information is garbled and destroyed.

As a key observation, knowledge of such manipulations can have a heterogeneous impact on expected utility. A decision maker that is strategically sophisticated enough to recognize interference by the moderator can be strictly better-off, as it can lead to a beneficial adjustment of the DM's course of action. However, it can also lead to a strict loss in expected utility, particularly so if the choice environment is binary but there are more than two states. A more sophisticated decision maker is harder to manipulate by

an adversarial agent as they use the available information optimally based on their beliefs. However, beliefs can differ due to misunderstandings and biases of the decision maker. Sophistication can thus pose a binding constraint to a benevolent expert who is trying to correct the effects of such biases and misperceptions. In other words, raising individuals' awareness of possible manipulations and thus making them more resilient to misinformation can have a negative side-effect.

We also identify and characterize settings in which a moderator would want to completely misinform a decision maker (i.e., alter which signal induces which posterior). Even small differences in priors and/or the perception of an information experiment can lead to what we call 'complete disagreement'. The moderator and decision maker hold an opposing view on which action should follow which signal. To an outside observer, this might look like the moderator and decision maker have opposing interest (i.e. different preferences). But this can be caused by only small differences in how new information is interpreted. This goes to show that disagreement and misinformation can arise even between individuals that have identical interests and very similar beliefs.

A Auxiliary Results

A.1 Complete disagreement

Lemma 5. *Suppose given an action profile $\boldsymbol{a}$, DM and moderator are in complete disagreement over some a_i and a_j. Then the moderator prefers $\mathbf{a}_{i\to j}$ and $\mathbf{a}_{j\to i}$ to $\mathbf{a}$.*

Proof. WLOG, let $V(X|\boldsymbol{a}_{i\to j},\boldsymbol{\mu}) \geq V(X|\boldsymbol{a}_{j\to i},\boldsymbol{\mu})$. Complete disagreement over a_i and a_j requires that

$$V(X|\boldsymbol{a}_{i\leftrightarrow j},\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu}).$$

Together they imply $E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big] > E\big[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big]$, and hence $V(X|\boldsymbol{a}_{i\to j},\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu})$. Moreover, again by definition of complete disagreement,

$$V(X|\boldsymbol{a}_{i\leftrightarrow j},\boldsymbol{\mu}) > \max\big\{V(X|\boldsymbol{a}_{i\to j},\boldsymbol{\mu}), V(X|\boldsymbol{a}_{j\to i},\boldsymbol{\mu})\big\}.$$

It follows that $E\big[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)\big] > E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)\big]$, and thus $V(X|\boldsymbol{a}_{j\to i},\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu})$, as required. □

Lemma 6. *Let $|\Omega| > 2$. For any signal $\boldsymbol{s}$ from an experiment X, there (i) exist a pair of distinct prior beliefs $\boldsymbol{\mu},\boldsymbol{p} \in \Delta(\Omega)$ such that $\operatorname{sign}(\boldsymbol{\mu}_\omega(\boldsymbol{s}) - \boldsymbol{\mu}_\omega) \neq \operatorname{sign}(\boldsymbol{p}_\omega(\boldsymbol{s}) - \boldsymbol{p}_\omega)$ for any non-extreme state ω, i.e., $\omega \notin \arg\min_{\omega'} \boldsymbol{s}_{\omega'}$ and $\omega \notin \arg\max_{\omega'} \boldsymbol{s}_{\omega'}$. Indeed, (ii) for any $\epsilon > 0$, there exists such a such pair of beliefs with $||\boldsymbol{\mu} - \boldsymbol{p}|| < \epsilon$. Furthermore, (iii) such a pair of beliefs exists with $\boldsymbol{\mu}_\omega = \boldsymbol{p}_\omega$.*

Proof. Denote the states for which signal $\boldsymbol{s}$ is least and most informative by $l \equiv \arg\min_{\omega'} \boldsymbol{s}_{\omega'}$ and $h \equiv \arg\max_{\omega'} \boldsymbol{s}_{\omega'}$.

Wlog, assume these two states are unique for notational convenience. Let $\Omega^- \equiv \Omega \setminus \{h, l\}$.

We start by first showing part (i) of the lemma as it generates a useful condition despite the fact that part (ii) implies part (i). For any state $\omega \in \Omega^-$, the decision maker updates upwards if and only if $\boldsymbol{\mu}_\omega(\boldsymbol{s}) = \frac{\boldsymbol{\mu}_\omega \cdot \boldsymbol{s}_\omega}{\langle \boldsymbol{s}, \boldsymbol{\mu} \rangle} > \boldsymbol{\mu}_\omega$, which is true whenever the likelihood of the signal in state ω, $\boldsymbol{s}_\omega$, exceeds the likelihood of receiving $\boldsymbol{s}$. Rewrite this inequality as

$$\begin{aligned} \boldsymbol{s}_\omega &> \boldsymbol{\mu}_\omega \boldsymbol{s}_\omega + \boldsymbol{\mu}_h \boldsymbol{s}_h + \boldsymbol{\mu}_l \boldsymbol{s}_l + \sum_{\omega' \in \Omega^- \setminus \omega} \boldsymbol{\mu}_{\omega'} \boldsymbol{s}_{\omega'} \\ 0 &> \boldsymbol{\mu}_h \cdot (\boldsymbol{s}_h - \boldsymbol{s}_\omega) - \boldsymbol{\mu}_l \cdot (\boldsymbol{s}_\omega - \boldsymbol{s}_l) + \sum_{\omega' \in \Omega^- \setminus \omega} \boldsymbol{\mu}_{\omega'} \cdot (\boldsymbol{s}_{\omega'} - \boldsymbol{s}_\omega) \end{aligned} \tag{7}$$

Set $\boldsymbol{\mu}_{\omega'} = 0$ for all $\omega' \in \Omega^- \setminus \omega$, the RHS becomes $\boldsymbol{\mu}_h \cdot (\boldsymbol{s}_h - \boldsymbol{s}_\omega) - \boldsymbol{\mu}_l \cdot (\boldsymbol{s}_\omega - \boldsymbol{s}_l)$ and so the DM updates upwards if

$$\frac{\boldsymbol{\mu}_l}{\boldsymbol{\mu}_h} > \frac{\boldsymbol{s}_h - \boldsymbol{s}_\omega}{\boldsymbol{s}_\omega - \boldsymbol{s}_l} \tag{8}$$

and downward otherwise. As the relative signal ratio $\frac{\boldsymbol{s}_h - \boldsymbol{s}_\omega}{\boldsymbol{s}_\omega - \boldsymbol{s}_l}$ is a positive finite number, there always exist a pair of prior beliefs $\boldsymbol{\mu}, \boldsymbol{p}$ with sufficient weights on state h and l such that one prior belief ratio exceeds it and the other falls short of it. Note, this inequality can be used to directly check the direction of updating with 3-states.[12]

(ii) To show that this can be true for two arbitrarily close prior beliefs, we first find a prior belief $\hat{\boldsymbol{q}}$ for which inequality (7) is an equality. Since, $\boldsymbol{s}_h > \boldsymbol{s}_\omega > \boldsymbol{s}_l$, such $\hat{\boldsymbol{q}}$ exists, i.e., $\boldsymbol{s}_\omega = \hat{\boldsymbol{q}}_h \boldsymbol{s}_h + \hat{\boldsymbol{q}}_h \boldsymbol{s}_l$ and which places probability 0 on all other states. But then, a strictly positive prior $\boldsymbol{q}$ also exists. To obtain, $\boldsymbol{\mu}$ and $\boldsymbol{p}$ that are arbitrarily close, simply shift a sufficiently small probability from state h to l and from l to h respectively.

(iii) It is easily verified that the result goes through with the additional restriction of $\boldsymbol{\mu}_\omega = \boldsymbol{p}_\omega$. □

A.2 Belief space & orientation

Lemma 7. *Let $\boldsymbol{v}$, $\boldsymbol{w}$, and $\boldsymbol{x}$ be vectors $\mathbb{R}^2$. Then $\boldsymbol{v}$ and $\boldsymbol{w}$ have opposing orientation relative to $\boldsymbol{x}$ if and only if the matrices*

[12] While this part of the proof only relies on the prior for two states, it obviously extends to strictly positive prior beliefs (see also part (ii)). To see this, note that shifting any weighs from the prior of states with $\boldsymbol{s}_{\omega'} > \boldsymbol{s}_\omega$ to those with $\boldsymbol{s}_{\omega''} < \boldsymbol{s}_\omega$ lowers the RHS (and vice versa for states with relatively lower signal strength).

$$A = \begin{pmatrix} | & | \\ \boldsymbol{x} & \boldsymbol{v} \\ | & | \end{pmatrix} \qquad B = \begin{pmatrix} | & | \\ \boldsymbol{x} & \boldsymbol{w} \\ | & | \end{pmatrix}$$

are such that $\det(A) < 0 < \det(B)$*, or* $\det(B) < 0 < \det(A)$.

Proof. The opposing orientation requires that the sets $\{\boldsymbol{x}, \boldsymbol{v}\}$ and $\{\boldsymbol{x}, \boldsymbol{w}\}$ both span and there exists a matrix L with $\{\boldsymbol{x}, \boldsymbol{w}\} = \{L\boldsymbol{x}, L\boldsymbol{v}\}$ and $\det(L) < 0$. Since they span, $\det(A), \det(B) \neq 0$. It follows from the product rule of determinants that this can hold if and only if $\text{sign}(\det(A)) = -\text{sign}(\det(B))$. The result follows. □

Let $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ for all $i \in \{0,1,2\}$ be distinct vectors in a 2-dimensional Euclidean space. Let V_i and W_i for all $i \in \{0,1,2\}$ denote the corresponding points. Suppose the vectors are such that V_0 lies on the line segment $\overline{V_1V_2}$ (i.e., $\boldsymbol{v}_0$ is a convex combination of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$), W_0 lies on the line segment $\overline{W_1W_2}$, but not all points lie on a single line. Further define $\Delta\boldsymbol{v} \equiv \boldsymbol{v}_1 - \boldsymbol{v}_0$, $\Delta\boldsymbol{w} \equiv \boldsymbol{w}_1 - \boldsymbol{w}_0$, and $\Delta\boldsymbol{x} \equiv \boldsymbol{w}_0 - \boldsymbol{v}_0$.

Proposition 7 (Vector Orientation)**.** *The line segments* $\overline{V_1W_2}$ *and* $\overline{V_2W_1}$ *do not cross if and only if there exists* $a_1, a_2 \in \mathbb{R}$*,* $b_1 < 0 < b_2$*, and a vector* $\boldsymbol{u}$ *such that at least one of the following holds:*

(i) $\langle \boldsymbol{u}, \Delta\boldsymbol{v} \rangle = 0$*, while* $\boldsymbol{w}_1 - \boldsymbol{v}_2 = a_1\Delta\boldsymbol{v} + b_1\boldsymbol{u}$*, and* $\boldsymbol{w}_2 - \boldsymbol{v}_1 = a_2\Delta\boldsymbol{v} + b_2\boldsymbol{u}$*, or*

(ii) $\langle \boldsymbol{u}, \Delta\boldsymbol{w} \rangle = 0$*, while* $\boldsymbol{w}_1 - \boldsymbol{v}_2 = a_1\Delta\boldsymbol{w} + b_1\boldsymbol{u}$*, and* $\boldsymbol{w}_2 - \boldsymbol{v}_1 = a_2\Delta\boldsymbol{w} + b_2\boldsymbol{u}$*, or*

(iii) $\langle \boldsymbol{u}, \Delta\boldsymbol{x} \rangle = 0$*, while* $\Delta\boldsymbol{w} = a_1\Delta\boldsymbol{x} + b_1\boldsymbol{u}$*, and* $\Delta\boldsymbol{v} = a_2\Delta\boldsymbol{x} + b_2\boldsymbol{u}$.

Proof. **Sufficiency:**
First note that the line segment $\overline{V_1W_2}$ can be characterized by the set of points $\mathbb{P}_1 = \left\{V_1 + \lambda(\boldsymbol{w_2} - \boldsymbol{v_1}) | \lambda \in [0,1]\right\}$, while the line segment $\overline{V_2W_1}$ can be characterized by $\mathbb{P}_2 = \left\{V_2 + \lambda(\boldsymbol{w_1} - \boldsymbol{v_2}) | \lambda \in [0,1]\right\}$. Now suppose indeed that a vector $\boldsymbol{u}$ exists, so that (i) holds. We can construct matrices A and B such that:

$$A = \begin{pmatrix} | & | \\ \Delta\boldsymbol{v} & \boldsymbol{w}_2 - \boldsymbol{v}_1 \\ | & | \end{pmatrix} \qquad B = \begin{pmatrix} | & | \\ \Delta\boldsymbol{v} & \boldsymbol{w}_1 - \boldsymbol{v}_2 \\ | & | \end{pmatrix}$$

where

$$\Delta\boldsymbol{v} = \begin{pmatrix} \Delta v_1 \\ \Delta v_2 \end{pmatrix} \qquad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \qquad \boldsymbol{w}_i - \boldsymbol{v}_j = \begin{pmatrix} a_i\Delta v_1 + b_i u_1 \\ a_i\Delta v_2 + b_i u_2 \end{pmatrix}.$$

We can compute $\det(A) = b_1(\Delta v_1 u_2 - \Delta v_2 u_1)$, and $\det(B) = b_2(\Delta v_1 u_2 - \Delta v_2 u_1)$. As $b_1 < 0 < b_2$, $\det(A)$ and $\det(B)$ are of opposing sign. They thus describe linear maps of

different orientation (Lemma 7). Consequently, the cross products $\Delta\boldsymbol{v} \times (\boldsymbol{w}_2 - \boldsymbol{v}_1)$ and $\Delta\boldsymbol{v} \times (\boldsymbol{w}_1 - \boldsymbol{v}_2)$ have opposite signs (i.e. point to opposite sides relative to $\Delta\boldsymbol{v}$). Note further that since $\boldsymbol{v}_0$ is a convex combination of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, we have $V_2 = V_1 + \kappa\Delta\boldsymbol{v}$, for some $\kappa \in \mathbb{R}$. We can write $\mathbb{P}_2$ as $\{V_1 + \kappa\Delta\boldsymbol{v} + \lambda(\boldsymbol{w_1} - \boldsymbol{v_2}) | \lambda \in [0,1]\}$. As $V_1 \neq V_2$, and hence $\kappa \neq 0$, it follows that the sets $\mathbb{P}_1$ and $\mathbb{P}_2$ describing the line segments are necessarily disjoint, i.e. the line segments cannot cross. As the naming of vectors was arbitrary and the statements are symmetric, sufficiency of (ii) follows.

Next we prove sufficiency of (iii): Suppose such a $\boldsymbol{u}$ exists. Since $\boldsymbol{v}_0$ is assumed to be a convex combination of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, we can write $\boldsymbol{v}_2 - \boldsymbol{v}_0 = -\kappa\Delta\boldsymbol{v}$ for some $\kappa > 0$. We can thus write $\boldsymbol{v}_2 - \boldsymbol{v}_0 = a_3\Delta\boldsymbol{x} + b_3\boldsymbol{u}$, where $b_3 < 0$ (i.e. the same sign as b_1). Construct the matrices

$$C = \begin{pmatrix} | & | \\ \Delta\boldsymbol{x} & \boldsymbol{w}_2 - \boldsymbol{w}_0 \\ | & | \end{pmatrix} \qquad D = \begin{pmatrix} | & | \\ \Delta\boldsymbol{x} & \boldsymbol{v}_1 - \boldsymbol{v}_0 \\ | & | \end{pmatrix}.$$

Computing the determinants, we can conclude that $\text{sign}(\det(C)) = \text{sign}(\det(D))$. But then the cross products $\Delta\boldsymbol{x} \times (\boldsymbol{w}_2 - \boldsymbol{w}_0)$ and $\Delta\boldsymbol{x} \times (\boldsymbol{v}_1 - \boldsymbol{v}_0)$ have the same sign (i.e. point to the same side relative to $\Delta\boldsymbol{v}$). Equivalently, we can establish that the cross products $\Delta\boldsymbol{x} \times (\boldsymbol{w}_1 - \boldsymbol{w}_0)$ and $\Delta\boldsymbol{x} \times (\boldsymbol{v}_2 - \boldsymbol{v}_0)$ also have the same sign, but the opposite sign compared to the previous two cross products. This means V_1 and W_2 lie on the same side of the line segment $\overline{V_0W_0}$, while V_2 and W_1 both lie on the opposite side. Set $\mathbb{P}_1$ and $\mathbb{P}_2$ are disjoint as required.

Necessity:

Suppose (i), (ii), and (iii) are not satisfied. It then follows from the previous argument that V_1 and W_1 must lie on the same side of the line segment $\overline{V_0W_0}$, with V_2 and W_2 both on the opposite side. Then either (a) the line segments $\overline{V_1W_1}$, $\overline{W_1W_2}$, $\overline{V_2W_2}$, and $\overline{V_1V_2}$ form the edges of a quadrilateral. Or (b), the line segments $\overline{V_1W_1}$, $\overline{W_1V_2}$, $\overline{V_2W_2}$, and $\overline{V_1W_2}$ form the edges of a quadrilateral. But note that (b) requires that W_1 and W_2 lie on opposite sides of the line through V_1 and V_2. This would require that for any $\boldsymbol{u}$ with $\langle\boldsymbol{u},\boldsymbol{v}\rangle = 0$, we can find a_1, a_2, b_1, b_2 such that $\boldsymbol{w}_1 - \boldsymbol{v}_2 = a_1\Delta\boldsymbol{v} + b_1\boldsymbol{u}$ and $\boldsymbol{w}_2 - \boldsymbol{v}_1 = a_2\Delta\boldsymbol{v} + b_2\boldsymbol{u}$, with b_2 the opposite sign of b_1. But this would mean (i) is satisfied. A contradiction. Hence the edges (a) form a quadrilateral. It follows further from the violation of (ii) that relative to $\Delta\boldsymbol{w}$, vectors $\boldsymbol{w}_1 - \boldsymbol{v}_2$ and $\boldsymbol{w}_2 - \boldsymbol{v}_1$ must have the same orientation. This implies that V_1 and V_2 lie on the same side of the line segment $\overline{W_1W_2}$. We can conclude that the line segments $\overline{V_1W_1}$, $\overline{W_1W_2}$, $\overline{V_2W_2}$, and $\overline{V_1V_2}$ form the edges of a *convex* quadrilateral. It then follows from the Crossbar Theorem that the diagonals $\overline{V_1W_2}$ and $\overline{V_2W_1}$ cross as required. □

B Proofs

Proof of Proposition 1:

Proof. Convexity in posterior beliefs: Expected utility $E[u(a|\omega)|\boldsymbol{\mu}]$ is linear and hence weakly convex in $\boldsymbol{\mu}$. The expected utility from each action as a function of $\boldsymbol{\mu}$ can be seen as a hyperplane in Δ^n. The maximum value from any experiment can also be written as a linear function in $\boldsymbol{\mu}$:

$$V(X|\boldsymbol{\mu}) = \max_{a\in\mathbb{A}^{|S_X|}} \sum_{s_i\in S_X}\sum_{\omega\in\Omega} \langle \boldsymbol{s}_i, \boldsymbol{\mu}\rangle u(a_i|\omega)\frac{\boldsymbol{s}_i\boldsymbol{\mu}_\omega}{\langle \boldsymbol{s}_i, \boldsymbol{\mu}\rangle}.$$

This again describes a hyperplane in Δ^n. By the properties of the maximum, the combination of such hyperplanes that achieves maximum expected utility is necessarily convex in $\boldsymbol{\mu}$.

Non-convexity from distortions: Recall that $\mathbb{A}$ is assumed to be such that actions are not payoff equivalent. Then for any non-trivial experiment, distortion, and signal-sensitive $\boldsymbol{a}$, we have $V(X^d|\boldsymbol{a},\boldsymbol{\mu}) \neq V(X|\boldsymbol{a},\boldsymbol{\mu})$. Let a^* be such that $a^* = \arg\max_{a\in\mathbb{A}} u(a|\omega)$ for some $\omega \in \Omega$. It follows from linearity of V in beliefs that for any non-trivial X^d, there exists a belief $\boldsymbol{\mu}^*$ such that: (i) $E[u(a^*|\omega)|\boldsymbol{\mu}^*] = V(X^d|\boldsymbol{a},\boldsymbol{\mu}^*) = V(X^d|\boldsymbol{\mu}^*)$ and (ii) the profile $\boldsymbol{a}$ is signal sensitive. As $\boldsymbol{a}$ is signal sensitive, $V(X|\boldsymbol{a},\boldsymbol{\mu}^*) \neq V(X^d|\boldsymbol{a},\boldsymbol{\mu}^*)$. As V is linear and thus continuous in beliefs, for any $\epsilon > 0$, we can find $\boldsymbol{\mu}_\epsilon$ with $||\boldsymbol{\mu}_\epsilon - \boldsymbol{\mu}^*|| < \epsilon$, such that $V(X^d|\boldsymbol{\mu}_\epsilon) = E[u(a^*|\omega)|\boldsymbol{\mu}_\epsilon]$. As the action profile chosen at this belief is not signal sensitive, we have $\hat{V}(X|\boldsymbol{\mu}_\epsilon) = V(X^d|\boldsymbol{\mu}_\epsilon)$. At the same time, we can find $\boldsymbol{\mu}'_\epsilon$ with $||\boldsymbol{\mu}'_\epsilon - \boldsymbol{\mu}^*|| < \epsilon$ such that $V(X^d|\boldsymbol{\mu}'_\epsilon) = V(X^d|\boldsymbol{a},\boldsymbol{\mu}'_\epsilon) > E[u(a^*|\omega)|\boldsymbol{\mu}_\epsilon]$. As $V(X|\boldsymbol{a},\boldsymbol{\mu}^*) \neq V(X^d|\boldsymbol{a},\boldsymbol{\mu}^*)$, it follows that $\lim_{\epsilon\to 0}\hat{V}(X|\boldsymbol{\mu}'_\epsilon) \neq \lim_{\epsilon\to 0}\hat{V}(X|\boldsymbol{\mu}_\epsilon)$. There is a discontinuity at $\boldsymbol{\mu}^*$. Convexity in $\boldsymbol{\mu}$ necessarily fails.

Non-convexity from biased prior:
Take an arbitrary experiment X with signals S_X. It follows from Alonso and Câmara (2016) [Proposition 1] that the posteriors between moderator and DM can be related as follows:

$$p_\omega(\boldsymbol{s}_i)\cdot\langle \boldsymbol{\mu}(\boldsymbol{s}_i), \boldsymbol{p}\circ\boldsymbol{\mu}^{-1}\rangle = \mu_\omega(\boldsymbol{s}_i)\cdot\frac{p_\omega}{\mu_\omega}$$

where $\boldsymbol{p}\circ\boldsymbol{\mu}^{-1} = \left(\frac{p_\omega}{\mu_\omega}\right)_{\omega\in\Omega}$. Let ω^* be such that $\frac{p_{\omega^*}}{\mu_{\omega^*}} = \min\left\{\frac{p_\omega}{\mu_\omega} \mid \omega\in\Omega\right\}$. Let $\alpha_1,...,\alpha_l$ be weights such that $\sum_{i=1}^{l}\alpha_i\mu_\omega(\boldsymbol{s}_i) = \boldsymbol{\mu}$. These correspond to the probabilities with which each signal in S_X is observed from the perspective of the moderator.

We can write:

$$\sum_i \alpha_i p_{\omega^*}(\boldsymbol{s}_i)\langle \boldsymbol{\mu}(\boldsymbol{s}_i), \boldsymbol{p}\circ\boldsymbol{\mu}^{-1}\rangle = \frac{p_{\omega^*}}{\mu_{\omega^*}}\sum_i \alpha_i \mu_{\omega^*}(\boldsymbol{s}_i). \tag{9}$$

Observe that $\sum_{\omega\in\Omega}\mu_\omega(\boldsymbol{s}_i) = 1$. It follows that for all $\boldsymbol{s}_i \in S_X$:

$$\sum_{\omega\in\Omega}\mu_\omega(\boldsymbol{s}_i)\frac{p_\omega}{\mu_\omega} > \frac{p_{\omega^*}}{\mu_{\omega^*}}.$$

Note that the left-hand side equals $\langle \boldsymbol{\mu}(\boldsymbol{s}_i), \boldsymbol{p}\circ\boldsymbol{\mu}^{-1}\rangle$. Define

$$\kappa_i \equiv \langle \boldsymbol{\mu}(\boldsymbol{s}_i), \boldsymbol{p}\circ\boldsymbol{\mu}^{-1}\rangle \cdot \frac{\mu_{\omega^*}}{p_{\omega^*}}$$

and note that $\kappa_i > 1$ for all $i \in \{1,\dots l\}$. Substituting this into (9), we we obtain:

$$\sum_i \alpha_i p_{\omega^*}(\boldsymbol{s}_i)\cdot\kappa_i = \sum_i \alpha_i \mu_{\omega^*}(\boldsymbol{s}_i).$$

For convexity to hold from the perspective of the moderator for any X, we require $\sum_i \alpha_i p_{\omega^*}(\boldsymbol{s}_i)\cdot\kappa_i = \sum_i \alpha_i p_{\omega^*}(\boldsymbol{s}_i) = p_\omega$. But since $\kappa_i > 1$ for all i, this cannot be the case.

□

B.1 Beneficial moderation

Proof of Lemma 1:

Proof. Naive DM - sufficiency:
For a naive DM, the policy $m(\boldsymbol{s}_j) = \boldsymbol{s}_i$ and m equals to the identity mapping for all other signals achieves expected utility equal to:

$$\begin{aligned} V(X^m|\boldsymbol{a},\boldsymbol{\mu}) &= \sum_{t=1}^{l}\langle \boldsymbol{\mu}, \boldsymbol{s}_t^m\rangle \cdot E\big[u(a_t|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_t^m)\big] \\ &= \sum_{t\neq i,j}^{l}\langle \boldsymbol{\mu}, \boldsymbol{s}_t\rangle \cdot E\big[u(a_t|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_t)\big] + \langle \boldsymbol{\mu}, \boldsymbol{s}_i+\boldsymbol{s}_j\rangle \cdot E\big[u(a_i|\omega)|\boldsymbol{\mu}_{ij}\big] \end{aligned}$$

where we used that $\boldsymbol{s}_j^m = \boldsymbol{0}$, and $\boldsymbol{s}_i^m = \boldsymbol{s}_i + \boldsymbol{s}_j$. It follows from linearity of E and Bayes consistency that

$$\langle \boldsymbol{\mu}, \boldsymbol{s}_i+\boldsymbol{s}_j\rangle \cdot E\big[u(a_i|\omega)|\boldsymbol{\mu}_{ij}\big] = \langle \boldsymbol{\mu}, \boldsymbol{s}_i\rangle \cdot E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)\big] + \langle \boldsymbol{\mu}, \boldsymbol{s}_j\rangle \cdot E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big].$$

As $\boldsymbol{a}_{i\to j}$ is preferred to $\boldsymbol{a}$, we can conclude that $V(X^m|\boldsymbol{a},\boldsymbol{\mu}) = V(X|\boldsymbol{a}_{i\to j},\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu})$ as required.
Naive DM - necessity:
Suppose there exists such a moderation policy. Then there exists a signal $\boldsymbol{s}_j$ with $m(\boldsymbol{s}_j) = \sum_1^l p_t \boldsymbol{s}_t$, where $\sum_1^l p_t = 1$ and, as m is non trivial, $p_i > 0$ for some $i \neq j$. For this

to increase expected utility, there must be at least some such $i \neq j$ with $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)] > E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$. But then $\boldsymbol{a}_{i\to j}$ must be preferred to $\boldsymbol{a}$.

Sophisticated DM - sufficiency:
The policy $m(\boldsymbol{s}_j) = \epsilon \cdot \boldsymbol{s}_i + (1-\epsilon)\boldsymbol{s}_j$ and m equal to the identity mapping for all other signals achieves expected utility equal to:

$$V(X^m|\boldsymbol{a},\boldsymbol{\mu}) = \sum_{t\neq i,j}^{l} \langle \boldsymbol{\mu}, \boldsymbol{s}_t\rangle \cdot E\big[u(a_t|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_t)\big] + \langle \boldsymbol{\mu}, \boldsymbol{s}_i + \epsilon \cdot \boldsymbol{s}_j\rangle \cdot E\big[u(a^*|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i^m)\big] + \langle \boldsymbol{\mu}, (1-\epsilon)\boldsymbol{s}_j\rangle \cdot E\big[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big]$$

where a^* is the DM's choice after $\boldsymbol{s}_i^m$, i.e. at $\boldsymbol{p}(\boldsymbol{s}_i^m)$. Note that for $\epsilon \to 0$, $\boldsymbol{s}_i^m \to \boldsymbol{s}_i$. Then generically ($a_i$ being strictly preferred after $\boldsymbol{s}_i$), for small enough ϵ, continuity ensures that $a^* = a_i$, and due to linearity of E,

$$E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i^m)\big] = (1-\gamma)E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)\big] + \gamma \cdot E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big],$$

where $1-\gamma = \frac{\langle \boldsymbol{\mu}, \boldsymbol{s}_i\rangle}{\langle \boldsymbol{\mu}, \boldsymbol{s}_i + \epsilon\cdot \boldsymbol{s}_j\rangle}$. Since $\boldsymbol{a}_{i\to j}$ is preferred to $\boldsymbol{a}$, $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)] > E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$, which means this m strictly increases expected utility. □

Proof of Lemma 2:

Proof. *Sufficiency:* If such a profile $\boldsymbol{a}^*$ and garbling M exist, then there exists a moderation policy m, such that the DM chooses $\boldsymbol{a}^*$. This m is by definition beneficial.
Necessity: Suppose a beneficial moderation policy m with associated garbling M exists. Then there must exist some $\boldsymbol{a}^* \in \mathbb{A}^{|S_X|}$ such that $V(X|\boldsymbol{a}^*,\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu})$. Furthermore, for any $\boldsymbol{a}^* \neq \boldsymbol{a}$ to be chosen by a sophisticated DM, it must be that $\boldsymbol{a}^* = \operatorname{argmax}_{a\in\mathbb{A}^l} V(X^d M|\boldsymbol{a},\boldsymbol{p})$ and M must be non-trivial. For this to be beneficial, we need that the choices of the DM, given the true underlying signals, increase expected utility from the perspective of the moderator. Hence, it must be that $V(XM|\boldsymbol{a}^*,\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu})$, as required. □

B.2 Moderation & the gain from information

Proof of Lemma 3:

Proof. An action profile $\boldsymbol{a}$ maximizes expected utility at $\boldsymbol{\mu}$ if $V(X|\boldsymbol{a},\boldsymbol{\mu}) \geq V(X|\boldsymbol{a}',\boldsymbol{\mu})$ for all $\boldsymbol{a}' \in \mathbb{A}^{|S_X|}$. Write $V(X|\boldsymbol{a},\boldsymbol{\mu})$ as:

$$V(X|\boldsymbol{a},\boldsymbol{\mu}) = \sum_{\boldsymbol{s}\in S_X} \langle \boldsymbol{\mu}, \boldsymbol{s}\rangle \cdot E\big[u(a_s|\omega)|\boldsymbol{\mu}(\boldsymbol{s})\big] = \sum_{\boldsymbol{s}\in S_X\setminus\{\boldsymbol{s}_i,\boldsymbol{s}_j\}} \langle \boldsymbol{\mu}, \boldsymbol{s}\rangle \cdot E\big[u(a_s|\omega)|\mu(\boldsymbol{s})\big] + [\langle \boldsymbol{\mu}, \boldsymbol{s}_i\rangle + \langle \boldsymbol{\mu}, \boldsymbol{s}_j\rangle] \cdot V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij}]$$

where the last equality follows from the definition of V and Bayes consistency. If $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij}] < E[u(a|\omega)|\boldsymbol{\mu}_{ij}]$ for some $a \in \mathbb{A}$, then convexity of V in posterior beliefs is violated. Replacing a_i and a_j with a strictly increases expected utility. The equivalent result for the conditional gain can be obtained by replacing a with $a^* = \arg\max_{a\in\{a_i,a_j\}} E[u(a|\omega)|\boldsymbol{\mu}_{ij}]$. □

Proof of Proposition 2:

Proof. Suppose the DM chooses an action profile $\boldsymbol{a}$. We start with the naive DM:
Sufficiency:
Suppose such signals $\boldsymbol{s}_i,\boldsymbol{s}_j$ exist. Then, $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij}) - E[u(a_j|\omega)|\boldsymbol{\mu}_{ij}] < 0$. It follows from Bayes' consistency and linearity that $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a}_{j\to i},\boldsymbol{\mu}_{ij}) = E[u(a_j|\omega)|\boldsymbol{\mu}_{ij}]$, where we assumed wlog that a_j is the preferred action at $\boldsymbol{\mu}_{ij}$. Using the argument from Lemma 3, we can conclude the moderator prefers $\boldsymbol{a}_{j\to i}$ to $\boldsymbol{a}$. It follows from Lemma 1 that a beneficial moderation policy exists.
Necessity:
Suppose the moderator chooses a non-trivial moderation policy. Then again using Lemma 1, there must be actions a_i,a_j in $\boldsymbol{a}$, such that $\boldsymbol{a}_{j\to i}$ is preferred to $\boldsymbol{a}$. As the action profile $\boldsymbol{a}_{j\to i}$ is constant for signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, we have $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a}_{j\to i},\boldsymbol{\mu}_{ij}) = E[u(a_j|\omega)|\boldsymbol{\mu}_{ij}]$. As $E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$ is unchanged, it must be that $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)] < E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$ and hence $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a}_{j\to i},\boldsymbol{\mu}_{ij}) > V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij})$. Given $\boldsymbol{a}$, the conditional gain from $X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)$ is negative at $\boldsymbol{\mu}_{ij}$.

We continue with the sophisticated DM.
Sufficiency:
Let $\boldsymbol{s}_i,\boldsymbol{s}_j \in S_X$ be such signals. By definition there exists $a^* \in \mathbb{A}$, such that $E[u(a^*|\omega)|\mu_{ij}] > V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij})$. As follows from Lemma 3, $\boldsymbol{a}$ cannot be optimal and replacing a_i,a_j with a^* achieves strictly higher expected utility. Now consider a moderation policy with $m(\boldsymbol{s}_i) = m(\boldsymbol{s}_j) = \boldsymbol{s}_j$. By Bayes' rule, $\boldsymbol{p}(\boldsymbol{s}_j^m) = \boldsymbol{p}_{ij}$. As a^* maximizes expected utility at $\boldsymbol{p}_{ij}$, a sophisticated DM chooses a^* after observing $\boldsymbol{s}_j^m$. It follows from Lemma 2 that a beneficial moderation policy exists.
Necessity:
Suppose there exists a beneficial moderation policy with $m(s_i) = s_j$. For a sophisticated DM, we have $\boldsymbol{p}(\boldsymbol{s}_j^m) = \boldsymbol{p}_{ij}$. Generically, it follows that if $\boldsymbol{a}^*$ is chosen given this m, then $a_i^* = a_j^* = a^*$, for some $a^* \in \mathbb{A}$. Due to linearity, this is only beneficial if $E[u(a^*|\omega)|\boldsymbol{\mu}_{ij}] > V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij})$ as required. □

Proof of Corollary 2.1:

Proof. This follows from the necessity argument in Proposition 2. For beneficial moderation to exist, there must be some actions a_i,a_j that are part of $\boldsymbol{a} = (a_1,...,a_l)$, such

that $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a}_{j\to i},\boldsymbol{\mu}_{ij}) > V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a},\boldsymbol{\mu}_{ij})$. Then $m(s_i) = s_j$ achieves higher expected utility than any non-deterministic moderation policy, unless there is another a_k that dominates a_j at $\boldsymbol{\mu}(\boldsymbol{s}_i)$. But then the optimal policy has $m(s_i) = s_k$. With countable actions, the maximum is generically unique. Iterating this over all a_i gives the desired result. □

Proof of Proposition 3:

Proof. We first proof the second statement. Let m be a deterministic moderation policy given an experiment X with signals $S_X = \{\boldsymbol{s}_1,...,\boldsymbol{s}_l\}$. Let $\boldsymbol{\mu}(\boldsymbol{s}_i^m)$ denote the posterior belief for some $\boldsymbol{s}_i^m$ that occurs with positive probability. Furthermore, let $S_X^i \subseteq S_X$ denote the subset that includes all signals in S_X that are mapped into $\boldsymbol{s}_i$ (i.e. all $\boldsymbol{s}_k \in S_X$ with $m(\boldsymbol{s}_k) = \boldsymbol{s}_i$). As m is deterministic, this $\boldsymbol{s}_i^m$ can be written as $\boldsymbol{s}_i^m = \sum_{\boldsymbol{s}\in S_X^i} \boldsymbol{s}$. Following Bayes' rule:

$$\boldsymbol{\mu}(\boldsymbol{s}_i^m) = \frac{\boldsymbol{s}_i^m \circ \boldsymbol{\mu}}{\langle \boldsymbol{s}_i^m, \boldsymbol{\mu}\rangle}.$$

Noting that all $\boldsymbol{s}_k^m$ with $\boldsymbol{s}_k \in S_X^i$ (i.e. all signals that are mapped into $\boldsymbol{s}_i$) are perceived by the DM with probability 0 given m, the posterior $\boldsymbol{\mu}(\boldsymbol{s}_i^m)$ is reached with probability $\langle \boldsymbol{s}_i^m, \boldsymbol{\mu}\rangle$, i.e. the ex-ante probability of $\boldsymbol{s}_i^m$ being perceived by the DM.

Now consider an alternative moderation policy $\hat{m}$ such that $\hat{m}(\boldsymbol{s}_i) = \frac{1}{K}\sum_{\boldsymbol{s}\in S_X^i} \boldsymbol{s}$, where $K = |S_X^i|$. Note that $\hat{m}(\boldsymbol{s}_i) = \frac{1}{K}\boldsymbol{s}_i^m$. It follows that

$$\boldsymbol{\mu}(\boldsymbol{s}_i^{\hat{m}}) = \frac{\frac{1}{K}\boldsymbol{s}_i^m \circ \boldsymbol{\mu}}{\langle \frac{1}{K}\boldsymbol{s}_i^m, \boldsymbol{\mu}\rangle} = \frac{\boldsymbol{s}_i^m \circ \boldsymbol{\mu}}{\langle \boldsymbol{s}_i^m, \boldsymbol{\mu}\rangle} = \boldsymbol{\mu}(\boldsymbol{s}_i^m).$$

Furthermore, $\boldsymbol{\mu}(\boldsymbol{s}_i^{\hat{m}}) = \boldsymbol{\mu}(\boldsymbol{s}_k^{\hat{m}})$ for all $\boldsymbol{s}_k \in S_X^i$. Each of these posteriors is generated with (ex-ante) probability $\langle \boldsymbol{s}_k^{\hat{m}}, \boldsymbol{\mu}\rangle = \langle \boldsymbol{s}_i^{\hat{m}}, \boldsymbol{\mu}\rangle = \frac{1}{K}\langle \boldsymbol{s}_i^m, \boldsymbol{\mu}\rangle$. Hence $\boldsymbol{\mu}(\boldsymbol{s}_i^{\hat{m}})$ is perceived with (overall) probability $K \cdot \langle \boldsymbol{s}_i^{\hat{m}}, \boldsymbol{\mu}\rangle = \langle \boldsymbol{s}_i^m, \boldsymbol{\mu}\rangle$. Iterating over all distinct S_X^j allows us to conclude that $\hat{m}$ and m generate the same distribution over posteriors and thus yield the same expected utility.
Example 1.3 provides a specific example where a non-deterministic policy achieves strictly higher expected utility. This completes the proof. □

Proof of Corollary 3.1:

Proof. The optimal moderation policy for a naive DM must be deterministic (Corollary 2.1). As $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)] \geq E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$, it must be that $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)] > E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$ for there to be beneficial moderation for a naive DM (Proposition 2). Otherwise, the result follows trivially. If the optimal moderation policy for a naive DM is non-trivial, then $\boldsymbol{s}_i^m = \boldsymbol{s}_i + \boldsymbol{s}_j$. For a naive DM, this achieves expected utility of $\langle \boldsymbol{\mu}, \boldsymbol{s}_i^m\rangle \cdot E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i^m)] = E[u(a_i|\omega)|\boldsymbol{\mu}]$, where the last equality follows from Bayes'

consistency since $|S_X| = 2$. For the same moderation policy, expected utility for a sophisticated DM is

$$E[u(a^*|\omega)|\boldsymbol{\mu}], \quad \text{where } a^* = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}].$$

Clearly, the sophisticated DM is weakly better-off. □

Proof of Proposition 4:

Proof. Trivially, a_i needs to be suboptimal for there to exist a beneficial moderation policy. Suppose every suboptimal action part of $\boldsymbol{a}$ is only consistent with an underestimation of signal strength (relative to any $\boldsymbol{\mu}_{ij}$). Let a_i be such a suboptimal choice.
Naive DM: Compare a_i to any a_j, the choice after some signal $\boldsymbol{s}_j$. It follows from the premise that a_j must be either optimal or at also consistent with underestimation. It is thus the optimal choice for some belief in $\{\beta\boldsymbol{\mu}(\boldsymbol{s}_j)+(1-\beta)\boldsymbol{\mu}_{ij}|\beta\in[0,1]\}$. It then follows from linearity of expected utility in beliefs that since $a_i = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}_\alpha]$, for some $\alpha\in[0,1)$ that $E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)] < E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$. Equivalently, $E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)] > E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$. We can conclude that the conditional gain from X at $\boldsymbol{\mu}_{ij}$ must be positive. By Proposition 2 and Corollary 2.1, there cannot be a beneficial moderation policy for a naive DM.
Sophisticated DM: For beneficial moderation, there needs to exist an action $a^*\in\mathbb{A}$ such that $E[u(a^*|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i^m)] > E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$, and $a^* = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{p}(\boldsymbol{s}_i^{d,m})]$, where $\boldsymbol{s}_i^{d,m}$ is the distorted signal after moderation. Furthermore, if a_i is only consistent with an underestimation of signal strength, then for every $\boldsymbol{s}_j\in S_X$ with $j\neq i$, there exists an $\alpha_j\in[0,1]$ such that $a_i = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}_{\alpha_j}]$, where $\boldsymbol{\mu}_{\alpha_j} = \alpha_j\boldsymbol{\mu}(\boldsymbol{s}_i)+(1-\alpha_j)\cdot\boldsymbol{\mu}_{ij}$. As expected utility is linear in beliefs, and as $a_i\neq\arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$, we can find a set $\{\boldsymbol{\mu}_{\alpha_j}|j\neq i, j\in\{1,...,l\}\}$ of such beliefs that all lie on a hyperplane in Δ^{n-1}. By definition of $\boldsymbol{\mu}_\alpha$, this hyperplane divides the belief simplex Δ^{n-1} in two subsets: one containing $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and the other containing all other posteriors for signal in S_X. Denote the former by B_i. It follows from convexity in beliefs that moderation can only be beneficial, if a^* is optimal for some belief $\boldsymbol{\mu}^*\in B_i$. Again by linearity of expected utility, the set of beliefs for which a^* achieves higher expected utility than a_i must be a convex subset of B_i. Denote this by B_i^*. Now note that for all $\boldsymbol{s}_j^d\neq\boldsymbol{s}_i^d$, we have $\boldsymbol{p}(\boldsymbol{s}_j^d)\notin B_i^*$. This follows from the premise that all a_j they must be consistent with underestimation of signal strength at each $\boldsymbol{\mu}_{ji}$. Convexity of expected utility then implies that it cannot be that a_j is optimal at some $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}(\boldsymbol{s}_j)+(1-\alpha)\cdot\boldsymbol{\mu}_{ji}$ and at some $\boldsymbol{\mu} = \beta\boldsymbol{\mu}(\boldsymbol{s}_i)+(1-\beta)\cdot\boldsymbol{\mu}_{ji}$, but not at some belief strictly in between. As any moderation policy garbles signals, the resulting posteriors must lie in the convex hull of the original (distorted) posteriors $\{\boldsymbol{p}(\boldsymbol{s}_k^d)|\boldsymbol{s}_k^d\in S_{X^d}\}$. But $B_1^*\cap\{\boldsymbol{p}(\boldsymbol{s}_k^d)|\boldsymbol{s}_k^d\in S_{X^d}\} = \emptyset$. No moderation policy can generate a belief $\boldsymbol{p}(\boldsymbol{s}_i^{d,m})\in B_i^*$ that induces a choice a^*. No

beneficial moderation policy exists. □

Proof of Corollary 4.1:

Proof. In a binary setting, a less extreme posterior without a reversal of the correlation between signals and states implies that a_1 is optimally chosen for some $\boldsymbol{\mu}_\alpha = \alpha \cdot \boldsymbol{\mu}(\boldsymbol{s}_i) + (1-\alpha) \cdot \boldsymbol{\mu}$, $\alpha \in [0,1]$, since $\mu_1 = 1 - \mu_2$. The result then follows from Proposition 4. □

B.3 Complete disagreement

Proof of Lemma 4:

Proof. Let $\Phi^{\sim}(a_i, a_j) \subset \Delta^{n-1}$ be the set of beliefs for which the DM (and moderator) is indifferent between a_i and a_j. Similarly, let $\Phi^{+}(a_i, a_j)$ be the set of beliefs for which the DM strictly prefers a_i to a_j. Since expected utility is linear in beliefs, expected utility for each action $a \in \mathbb{A}$ forms a hyperplane in $\mathbb{R}^n$. The indifference set for any two actions a_i and a_j is thus geometrically defined by the intersection of two such hyperplanes. This means $\Phi^{\sim}(a_i, a_j)$ can be described by an indifference manifold in $\mathbb{R}^n$ (still referred to as a curve). Linearity implies that this has dimension $n-2$ and is itself a hyperplane of the simplex $\Delta^{n-1} \subset \mathbb{R}^{n-1}$. Furthermore, $\Phi^{+}(a_i, a_j)$ is a subset of Δ^{n-1}.

Now fix some experiment X, distortion d, and prior beliefs $\boldsymbol{\mu}$ and $\boldsymbol{p}$. For complete disagreement over some a_i and a_j, there needs to exist signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, such that the moderator strictly prefers a_j after $\boldsymbol{s}_i$, and a_i after $\boldsymbol{s}_j$. Preferences need to be such that $\boldsymbol{\mu}(\boldsymbol{s}_j), \boldsymbol{p}(\boldsymbol{s}_i^d) \in \Phi^{+}(a_i, a_j)$, and $\boldsymbol{\mu}(\boldsymbol{s}_i), \boldsymbol{p}(\boldsymbol{s}_j^d) \in \Phi^{+}(a_j, a_i)$. By definition, these are disjoint and separated by $\Phi^{\sim}(a_i, a_j)$. As these sets are convex, this can only be the case if the smallest convex set containing $\boldsymbol{\mu}(\mathbf{s}_i)$ and $\boldsymbol{p}(\mathbf{s}_j^d)$, i.e.

$$\left\{\mu \in \Delta^{n-1} : \mu = \alpha \cdot \boldsymbol{\mu}(\mathbf{s}_i) + (1-\alpha) \cdot \boldsymbol{p}(\mathbf{s}_j^d),\ \alpha \in [0,1]\right\}$$

is disjoint from the smallest convex set containing $\boldsymbol{\mu}(\mathbf{s}_j)$ and $\boldsymbol{p}(\mathbf{s}_i^d)$, which is

$$\left\{\mu \in \Delta^{n-1} : \mu = \alpha \cdot \boldsymbol{\mu}(\mathbf{s}_j) + (1-\alpha) \cdot \boldsymbol{p}(\mathbf{s}_i^d),\ \alpha \in [0,1]\right\}.$$

These are the line segments in question. If the line segments do not cross, then the sets are disjoint. The hyperplane separation theorem then guarantees the existence of a separating hyperplane in Δ^{n-1}. Let this be $\Phi^{\sim}(a_i, a_j)$. The remaining subsets of Δ^{n-1} are disjoint and convex. Let these be $\Phi^{+}(a_i, a_j)$ and $\Phi^{+}(a_j, a_i)$. It is easy to verify that preferences consistent with these sets must exist. With these preferences, there is complete disagreement. If the line segments cross, the sets are not disjoint. No separating hyperplane can exist, which precludes the required preference relation. □

Proof of Theorem 1:

Proof. **Case** $z = 1$: If the belief space is 1-dimensional, then any two vectors in the belief space are linearly dependent. The points $\boldsymbol{\mu}(s_i)$, $\boldsymbol{\mu}(s_j)$ and $\boldsymbol{p}(s_i^d)$, $\boldsymbol{p}(s_j^d)$ all lie on a line. Non-reversal guarantees that $\boldsymbol{p}(\boldsymbol{s}_i^d)$ cannot lie between $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and $\boldsymbol{p}(\boldsymbol{s}_i^d)$, and equivalently for $\boldsymbol{p}(\boldsymbol{s}_j^d)$. It follows from Lemma 4 that complete disagreement is not possible, since the line segments between $\boldsymbol{\mu}(s_i)$, $\boldsymbol{p}(s_j^d)$ and $\boldsymbol{p}(s_i^d)$, $\boldsymbol{\mu}(s_j)$ must necessarily intersect/coincide.

Case $z = 2$: The result follows almost immediately from Proposition 7. Let L denote the belief space. If the belief space is 2-dimensional, it can be equivalently represented in $\mathbb{R}^2$, i.e. there is an isomorphism $A : L \mapsto \mathbb{R}^2$. Note that such an isomorphism is either orientation-preserving or reversing (Guillemin and Pollack, 1974, p. 96). This means if two vectors have the opposing orientation property in L, this also holds in R^2. Let V_i be the point in $\mathbb{R}^2$ corresponding to $\boldsymbol{\mu}(\boldsymbol{s}_i)$, W_i corresponding to $\boldsymbol{p}(\boldsymbol{s}_i^d)$, and V_0 (W_0) corresponding to $\boldsymbol{\mu}_{ij}$ ($\boldsymbol{p}_{ij}$). The result follows from 7.

Case $z = 3$: If the belief space is 3-dimensional, then $\boldsymbol{\mu}(s_i)$, $\boldsymbol{\mu}(s_j)$ and $\boldsymbol{p}(s_i^d)$ lie on a plane that does not contain $\boldsymbol{p}(s_j^d)$. The line segment between $\boldsymbol{\mu}(s_i)$ and $\boldsymbol{p}(s_j^d)$ cannot cross with the line segment between $\boldsymbol{\mu}(s_j)$ and $\boldsymbol{p}(s_i^d)$. It follows from Lemma 4 that there exist preferences such that complete disagreement is possible. □

Proof of Corollary 4.2:

Proof. Following Theorem 1, complete disagreement is not possible if the belief space for two signals is 1-dimensional. But with $|\Omega| = 2$, all beliefs are contained in Δ^1, the 1-dimensional simplex. Any belief space thus has at most dimension 1. □

Proof of Proposition 5:

Proof. It follows from Lemma 5 that the moderator prefers $\mathbf{a}_{i\to j}$ to $\mathbf{a}$. It then follows directly from Lemma 1 that a beneficial moderation policy must exist for both a naive and sophisticated DM. □

Proof of Corollary 5.1:

Proof. For a naive DM, the moderation policy $m(s_i) = s_j$ and $m(s_j) = s_i$ achieves expected utility

$$V(X^m|\boldsymbol{a}_{i\leftrightarrow j}, \boldsymbol{\mu}) = V(X|\boldsymbol{a}_{i\leftrightarrow j}, \boldsymbol{\mu}) = V(X|\boldsymbol{\mu}).$$

By definition, a sophisticated DM could at most achieve equal expected utility.

It follows from Proposition 3 that the optimal policy is non-deterministic.

Given the assumption on either a_i or a_j, it is WLOG to assume a_i is the unique in $\{a_1,...,a_l\}$. Suppose further that $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_k)] < E[u(a_k|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_k)]$ for all $k \in 1,...,l$, $k \neq i$. This is generically true, if a_i is unique in $\{a_1,...,a_l\}$.

If $\boldsymbol{s}_j^m \neq \mathbf{0}$, then optimality requires that

$$a_i = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{p}(\boldsymbol{s}_j^m)].$$

and yet $\boldsymbol{s}_j^m = \boldsymbol{s}_j$.

The last equality follows from the fact that if a_i is chosen after $\boldsymbol{s}_j^m$, any mixture with another signal leads to a strict loss in expected utility relative to $V(X|\boldsymbol{\mu})$. This is a contradiction.

Suppose now $\boldsymbol{s}_j^m = \mathbf{0}$. Then using the previous argument, we need that

$$a_i = \arg\max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{p}(\boldsymbol{s}_k^m)].$$

with $\boldsymbol{s}_k^m = \boldsymbol{s}_j$, for some $\boldsymbol{s}_k \neq \boldsymbol{s}_j$. This is again a contradiction. The result follows. □

Proof of Corollary 5.2:

Proof. Complete disagreement and $|A| = 2$ imply that $V(X|\boldsymbol{\mu}) = V(X|\boldsymbol{a}_{i\leftrightarrow j}\boldsymbol{\mu})$. Furthermore, if $|S_X| = 2$, then each action is necessarily unique in $\boldsymbol{a}$. The result follows from Corollary 5.1. □

Proof of Proposition 6:

Proof. Suppose $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are such signals. Let $\tilde{\boldsymbol{s}}_i = \frac{\boldsymbol{s}_i}{\boldsymbol{s}_i+\boldsymbol{s}_j}$ and equivalently for $\tilde{\boldsymbol{s}}_j$; i.e. these are the signals associated with $X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)$. It then follows from Lemma 6 that there exist beliefs $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{p}_{ij}$ (strictly inside $\Delta(\Omega)$), such that $\text{sign}\big(\mu_\omega(\tilde{\boldsymbol{s}}_i)-\mu_\omega\big) = -\text{sign}\big(p_\omega(\tilde{\boldsymbol{s}}_i)-p_\omega\big)$, while $\mu_{\omega.ij} = p_{\omega.ij}$ for some state $\omega \in \Omega$. As the probabilities in each signal are distinct, the belief space is at least 2-dimensional. If it is 3-dimensional, there exist preferences such that there is complete disagreement (Proposition 1).

Suppose now it is 2-dimensional. As $\mu_{\omega.ij} = p_{\omega.ij}$, it follows that the vectors $\boldsymbol{v}_i = \boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}_{ij}$ and $\boldsymbol{w}_i = \boldsymbol{p}(\boldsymbol{s}_i) - \boldsymbol{p}_{ij}$ have opposing orientation relative to $\boldsymbol{p}_{ij} - \boldsymbol{\mu}_{ij}$. To see this, denote $\Delta\boldsymbol{x} = \boldsymbol{p}_{ij} - \boldsymbol{\mu}_{ij}$. Then $\Delta u_\omega = 0$, while any vector $\boldsymbol{u}$ that is orthogonal to $\boldsymbol{p}_{ij} - \boldsymbol{\mu}_{ij}$ necessarily has $u_\omega \neq 0$. We can write $\boldsymbol{v}_i$ as the linear combination $\boldsymbol{v}_i = \alpha_1\Delta\boldsymbol{x} + \beta_1\boldsymbol{u}$, and equivalently $\boldsymbol{w}_i = \alpha_2\Delta\boldsymbol{x} + \beta_2\boldsymbol{u}$. As $\text{sign}\big(\mu_\omega(\tilde{\boldsymbol{s}}_i) - \mu_\omega\big) = -\text{sign}\big(p_\omega(\tilde{\boldsymbol{s}}_i) - p_\omega\big)$, we have $v_{\omega.i} > 0 > w_{\omega.i}$ or $v_{\omega.i} < 0 < w_{\omega.i}$. Accordingly, $\text{sign}(\beta_2) = -\text{sign}(\beta_1)$. The result follows from Proposition 7, noting that beliefs $\boldsymbol{\mu}$ and $\boldsymbol{p}$ that lead to the conditional posteriors $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{p}_{ij}$ for the two signals necessarily exist, since the conditional beliefs are assumed to be strictly in the interior of $\Delta(\Omega)$. □

References

Ricardo Alonso and Odilon Câmara. Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706, 2016.

Roland Benabou and Jean Tirole. Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3):817–915, 2002.

Jean-Pierre Benoit, Juan Dubra, and Don A. Moore. Does the better-than-average effect show that people are overconfident? Two experiments. *Journal of the European Economic Association*, 13(2):293–329, 2015.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press, 1951.

Adam Brandenburger, Eddie Dekel, and John Geanakoplos. Correlated equilibrium with generalized information structures. *Games and Economic Behavior*, 4(2):182–201, 1992.

Isabelle Brocas. Information processing and decision-making: evidence from the brain sciences and implications for economics. *Journal of Economic Behavior & Organization*, 83(3):292–310, 2012.

Isabelle Brocas and Juan D Carrillo. Influence through ignorance. *The RAND Journal of Economics*, 38(4):931–947, 2007.

Jerome S. Bruner and Mary C. Potter. Interference in visual recognition. *Science*, 144 (3617):424–425, 1964.

Stephen V. Burks, Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini. Overconfidence and social signalling. *The Review of Economic Studies*, 80(3):949–983, 2013.

Juan D. Carrillo and Thomas Mariotti. Strategic ignorance as a self-disciplining device. *The Review of Economic Studies*, 67(3):529–544, 2000.

Gary Charness, Aldo Rustichini, and Jeroen Van de Ven. Self-confidence and strategic behavior. *Experimental Economics*, 21(1):72–98, 2018.

Julian Conrads and Bernd Irlenbusch. Strategic ignorance in ultimatum bargaining. *Journal of Economic Behavior and Organization*, 92:104–115, 2013.

Vincent Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–51, 1982.

John M. Darley and Paget H. Gross. A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1):20–33, 1983.

Geoffroy de Clippel and Xu Zhang. Non-bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642, 2022.

David Eil and Justin M. Rao. The good news–bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 2(3):114–138, 2011.

Nicholas Epley and Thomas Gilovich. The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40, September 2016.

Baruch Fischoff, Paul Slovic, and Sarah Lichtenstein. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3(4):552–564, 1977.

Russell Golman, David Hagmann, and George Loewenstein. Information avoidance. *Journal of Economic Literature*, 55(1):96–135, 2017.

Victor Guillemin and Alan Pollack. *Differential Topology*. Prentice Hall, 1974.

Faruk Gul. Unobservable investment and the hold-up problem. *Econometrica*, 69(2): 343–376, 2001.

Chris Guthrie, Jeffrey Rachlinski, and Andrew Wistrich. Inside the judicial mind: Heuristics and biases. *Cornell Law Review*, 56:777–830, 2001.

Jack Hirshleifer. The private and social value of information and the reward to inventive activity. *The American Economic Review*, 61(4):561–574, 1971.

Alexander Jakobsen. Coarse bayesian updating. Working paper, 2022.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Augustin Landier and David Thesmar. Financial contracting with optimistic entrepreneurs. *Review of Financial Studies*, 22(1):177–150, 2009.

Heidi J Larson, Louis Z Cooper, Juhani Eskola, Samuel L Katz, and Scott Ratzan. Addressing the vaccine confidence gap. *The Lancet*, 378(9790):526 – 535, 2011.

Sarah Lichtenstein, Baruch Fischoff, and Phillips Lawrence. Calibration of probabilities: The state of the art to 1980. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgement under uncertainty: Heuristics and biases*, pages 306–334. Cambridge University Press, Cambridge, 1982.

Elliot Lipnowski and Laurent Mathevet. Disclosure to a psychological audience. *American Economic Journal: Microeconomics*, 10(4):67–93, 2018.

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

Jacob Marschak and Koichi Miyasawa. Economic comparability of information systems. *International Economic Review*, 9(2):137–174, 1968.

Markus M. Mobius, Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. Managing self-confidence. *working paper*, 2014.

Don A. Moore and Paul J. Healy. The trouble with overconfidence. *Psychological Review*, 115(2):502–517, 2008.

Stephen Morris. The common prior assumption in economic theory. *Economics & Philosophy*, 11(2):227–253, 1995.

Matthew Motta, Timothy Callaghan, and Steven Sylvester. Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine*, 211:274–281, 2018.

Sendhil Mullainathan. Thinking through categories. Working paper, 2002.

Gregory A. Poland and Robert M. Jacobson. Understanding those who do not understand: a brief review of the anti-vaccine movement. *Vaccine*, 19(17):2440 – 2445, 2001.

Anders U. Poulsen and Michael W. M. Roos. Do people make strategic commitments? experimental evidence on strategic information avoidance. *Experimental Economics*, 13(2):206–225, 2010.

Mathew Rabin and Joel Schrag. First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, 114(2):37–82, 1999.

William P. Rogerson. Contractual solutions to the hold-up problem. *Review of Economic Studies*, 59(4):777–793, 1992.

Thomas C. Schelling. An essay on bargaining. *American Economic Review*, 46(3): 281–306, 1956.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.

Jakub Steiner and Colin Stewart. Perceiving prospects properly. *American Economic Review*, 106(7):1601–1631, 2016.

Jean Tirole. Procurement and renegotiation. *Journal of Political Economy*, 94(2): 235–259, 1986.

Neil D. Weinstein. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5):806–820, 1980.