The Benefits of Being Misinformed: Information Moderation under Misperception*

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

Manuel Staab
manuelstaab@gmail.com
Aix-Marseille University, AMSE

February 6, 2023

Abstract

We explore how two fundamental mistakes in information processing - incorrect beliefs 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 analyze a simple sender-receiver model in which a moderator (i.e., sender) can garble signals about payoff-relevant states. 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). 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 the moderator and decision maker completely disagree about which action should follow which signal. If there are at least three states, this 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, communication, persuasion, heterogeneous priors

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 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 individuals' views on how to interpret and act on information.

In light of these observations, we analyze a general belief-updating and decision problem 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 - called the *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), Green and Stokey (2007), 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 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 fully aligned with the DM. The moderator, 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

[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..

DM determines their action profile. We further distinguish two cases to examine the role of strategic/informational sophistication: (1) a 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), Rabin and Schrag (1999)) 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.

Biased beliefs and misperception distort posteriors and shift choices away from the optimal ones. Both imperfections have a common channel: they cause non-convexities (in beliefs) in the utility frontier, thus altering the value of information. Beyond this, biased beliefs also affect the utility ranking of actions in the absence of any informative signals. These consequences render interactions between moderator and DM effectively strategic and 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 behavior 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. 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 signals) 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 de-

cision maker who 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 in more complex information environments.

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. 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 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 effects 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.

For consistency and ease of interpretation, the analysis is phrased throughout to suggest that the moderator is free of such biases and misperceptions. However, given a suitable adjustment to the perspective on Welfare, it can also be interpreted as a sender and receiver holding heterogeneous views about the information environment, without taking any stand as to the accuracy of each view. The result can be seen as highlighting instances were a DM is more vulnerable to interference by a well-meaning but potentially destructive moderator. A third interpretation relates to the interaction of different mistakes around information processing, and highlights when the DM can be better off from suffering more severe perception issues. Specifically, if the moderator finds it optimal to destroy information, then a DM would be better off if their information processing was subject to noise or distortions, such as incorrectly recalling past information, or a tendency to dismiss certain types of signals, as long as these distortions mimic a beneficial moderation policy.

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.

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 and can cause a disadvantage in 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)), holdup problems (Tirole (1986), Rogerson (1992), Gul (2001)), and even intra-personal games of behavioral decision maker (Carrillo and Mariotti (2000), Benabou and Tirole (2002)). There is also an extensive literature on psychological reasons to avoid information (Kőszegi (2006), Golman et al. (2017)). In our setting, benefits from less accurate information are not based on strategic or psychological considerations, but rely only on different interpretations of the information environment. Information is purely instrumental.

Numerous studies have suggested that people hold incorrect beliefs, where beliefs can range from objective (economic) quantities to individual traits or prospects.[2] For example, Weinstein (1980) document unrealistically positive views for health and salaries. There is also evidence for overconfidence in entrepreneurs (Landier and Thesmar (2009)), as well as CEOs (Malmendier and Tate (2005)), who as a consequence are more likely to pursue risky actions. Overconfidence may, however, not always have a negative impact, as shown theoretically, for instance in the context of job search problems (Dubra (2004)). Furthermore, what may in empirical investigations *appear* as overconfidence - or more generally as some form of mistake in the belief (-formation) - may (partially) be a result of the way beliefs are elicited (Gigerenzer and Hoffrage (1995)) or the way the problem is presented (Gigerenzer et al. (1988)). Benoit and Dubra (2011) highlight that data based on median comparisons, e.g., that a majority of people view themselves as better than the median, does not imply

[2]While individuals tend be better able to asses everyday economic quantities such as the price of gas (Ansolabehere et al. (2013)), their beliefs are often incorrect even for quantities that are of direct importance to them, with measurable impact on outcomes. In the labor market, for example, Spinnewijn (2015) documents that 80% of job seekers underestimate the length of their unemployment spell, leading to too little search effort. Potter (2021) shows that job seekers overestimate their job-finding prospects by roughly 60% at the time of job loss. For a recent review on expectation in the labor market, see Mueller and Spinnewijn (2022).

they are overconfident. Instead, such beliefs are entirely consistent with Bayesian updating.[3] This paper makes a similar contribution in this regard, showing that a a fully Bayesian decision maker may update in completely opposite directions as a result of a (small) change in prior beliefs.[4] More generally, while there is debate regarding the source and existence of systematic biases in beliefs, there at least seems to be significant evidence that decision makers may not always hold accurate beliefs.

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 welfare and learning. For example, distorted perception can deliver a benefit in some settings (e.g., Steiner and Stewart (2016)) but may prove harmful in others (Rabin and Schrag (1999)). More specifically, Rabin and Schrag (1999) show one-sided updating due to confirmation bias can lead a decision maker to become fully convinced of a wrong hypothesis. Evidence for such updating behavior includes Mobius et al. (2014) and Eil and Rao (2011). Perception issues may also 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 to 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 be fully aligned between sender and receiver in order to remove strategic aspects, and instead highlight the role of differences in the interpretation of information. This resembles Green and Stokey (2007), who allow for differences in prior, but assume agents otherwise agree on the information structure. Furthermore, our approach focuses on different types of decision makers and includes the case where the sender can commit to an information structure. 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, Tsakas and Tsakas (2021), who analyze the impact of exogenous noise layered on top of the sender's signal, 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 potential signals but is constrained by the available informa-

[3] In response, lab experiments that document overconfidence and are robust to the critique in Benoit and Dubra (2011) are reported in Burks et al. (2013), Benoit et al. (2015), and Charness et al. (2018).

[4] Interestingly, while the information provision experiments by Armantier et al. (2016) (consumer inflation expectations) and Coibion et al. (2018) (macroeconomic beliefs of firms) suggest that decision makers hold incorrect beliefs, they also highlight that they tend to revise beliefs in a manner qualitatively consistent with Bayesian updating. See Haaland et al. (2023) for a review of the relevant literature.

tion experiment as well as the DMs view on it. 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 $\boldsymbol{\mu} \in \Delta(\Omega)$, 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 a 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 finite 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**. The relevant aspect of the signal is the probability with which it occurs in each state. 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 X is characterized by its finite set of possible signals S_X. Analogous to Blackwell (1951), we can consider X a row-stochastic 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 thus corresponds to the probability distribution over signals for a given state. 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 $\boldsymbol{s}_i$. The objective of a DM with access to an experiment X is to choose an action profile that maximizes (von Neumann-Morgenstern) expected utility:

$$\max_{\boldsymbol{a}\in\mathbb{A}^{|S_X|}} E\Big[U\big(a|\boldsymbol{\mu}(\boldsymbol{s})\big)\Big|\boldsymbol{\mu}\Big], \tag{1}$$

where $U\big(a|\boldsymbol{\mu}(\boldsymbol{s})\big) = E\big[u(a|\omega)|\boldsymbol{\mu}(\boldsymbol{s})\big]$ and $\boldsymbol{\mu}(\boldsymbol{s})$ denotes the 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 subject to distortion d is under the impression that the experiment yields signals $S_X^d = \{\boldsymbol{s}_1^d, ..., \boldsymbol{s}_k^d\}$. 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, i.e., holds a prior belief $\boldsymbol{p} \in \Delta(\Omega)$ that differs from some reference belief $\boldsymbol{\mu}$. This is referred to as a **bias in prior**. 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. It could alternatively be interpreted as a difference between the assessment of the DM and that of an observer, without any judgment regarding their accuracy.[5] In this case, the DM's belief is 'biased' merely from the perspective of the observer. When these are not identical, $\boldsymbol{p}$ 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 the experiment, the DM updates their prior belief according to Bayes' rule but on the basis of the distorted signal $\boldsymbol{s}^d$:

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

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, an unbiased, fully Bayesian observer revises their belief as follows:

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

As X and X^d are both experiments, **Bayes' consistency** holds with and without misperception, meaning the expected posterior equals the prior.[6]

Information Moderation:

Before a signal reaches the decision maker, it is first observed by an **information moderator**, who can influence its content, i.e., 'moderate' the information flow. In particular, given

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

[6] More generally, one could imagine a distortion that results in an X^d that is not a proper experiment, i.e., is not row-stochastic. 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.

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 $\boldsymbol{s}_j^m = \boldsymbol{s}_i + \boldsymbol{s}_j$ and $\boldsymbol{s}_i^m = \mathbf{0}$.[7] 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 $M = \left(m_{ij}\right)_{1 \leq i,j \leq k}$, where m_{ij} is the probability with which the DM perceives $\boldsymbol{s}_j^d$ given that the moderator observed signal $\boldsymbol{s_i}$. The experiment effectively becomes:

$$X^m = XM.$$

The moderator can commit to a moderation policy but has no intrinsic incentive to misinform the DM. Preferences are assumed to be identical. 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 skilled 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 regarding X^d and $\boldsymbol{p}$. The moderator may thus choose to implement a non-trivial moderation policy ($M \neq I$). If this *strictly* increases expected utility, it is referred to as *beneficial moderation*.

We distinguish between two types of decision makers: those oblivious to the moderation policy, who we call *naive*,[8] and those aware of the moderator's interference, who we call *sophisticated*. 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. In line with the persuasion literature, we are interested in the *sender-preferred Perfect Bayesian Equilibrium* outcomes for the sophisticated DM. The distinction of types allows us to examine more closely the interplay between strategic/informational sophistication, imperfections in decision making, and information moderation.

Indirect Utility:

When comparing outcomes between decision makers with and without distortions and bi-

[7] 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.

[8] This is also sometimes referred to as a 'credulous' type. See, for instance, Kartik et al. (2007).

ases, 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 undistorted experiment X by:

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

We also refer to this 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 $\boldsymbol{\mu}$. For instance, the expected utility that a DM subject to distortion d and biased prior $\boldsymbol{p}$ actually obtains from X can be expressed as $V(X|\boldsymbol{a}^*, \boldsymbol{\mu})$, where $\boldsymbol{a}^*$ is the action profile consistent with $V(X^d|\boldsymbol{p})$. If a moderator intervenes, we have to distinguish between the two types of decision maker. The expected utility a naive DM subject to a moderation policy m obtains from X equals $V(XM|\boldsymbol{a}^*, \boldsymbol{\mu})$, where again $\boldsymbol{a}^*$ is the action profile consistent with $V(X^d|\boldsymbol{p})$. In comparison, a sophisticated DM obtains $V(XM|\hat{\boldsymbol{a}}, \boldsymbol{\mu})$, where $\hat{\boldsymbol{a}}$ is the action profile consistent with $V(X^d M|\boldsymbol{p})$, i.e., the optimal choice given a belief $\boldsymbol{p}$, perceived signals S_X^d, and 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 $\boldsymbol{\mu}$.[9] 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.

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

[9] This is similar to the definition in Kamenica and Gentzkow (2011), except for the conditioning on $\boldsymbol{a}$ and the potential difference in priors.

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

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

*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

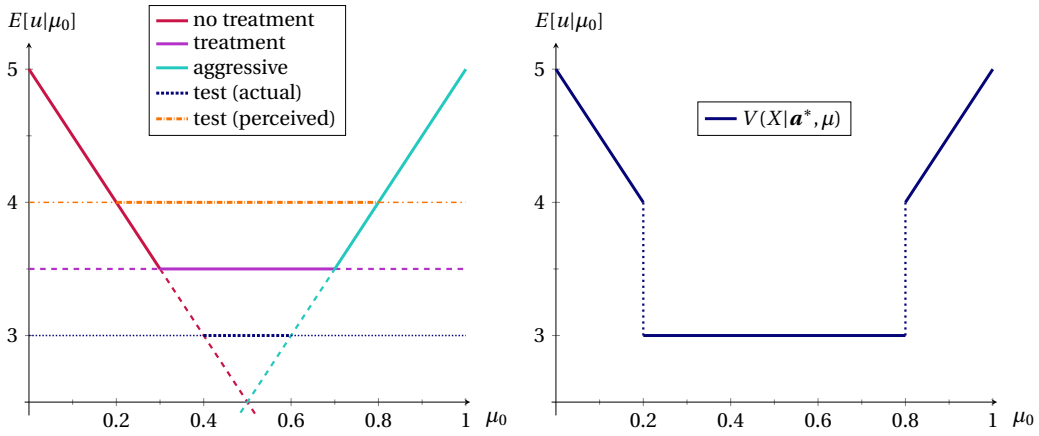
We first analyse the effects of distortions and biases. We then characterise the general optimization problem from the perspective of the moderator (Theorem 1) and subsequently explore when and how a moderator can help reduce the effect of perception issues and when moderation fails to have a positive impact. Particular emphasis is placed on contrasting the optimal (i.e., utility maximizing) moderation policy for naive decision makers (Proposition 1), with that for a sophisticated DM (Proposition 2), and exploring the ambiguous consequences of higher sophistication. We further examine what choices reveal about the type of distortion a DM faces and their implication on beneficial moderation (Proposition 3). 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 2).

4.1 Distortions, biases and their implications

Consider a patient consulting a doctor for advice. The interaction is summarized by the following game, which is adopted from Farrell and Rabin (1996), where rows reflect the doctor's information after a diagnostic test, and columns the patient's actions:

		patient		
		a_H	a_L	a_M
doctor's information	$\boldsymbol{\mu}_s$	$(3,4)$	$(2,1)$	$(\frac{7}{2},\frac{7}{2})$
	$\boldsymbol{\mu}_t$	$(2,1)$	$(3,4)$	$(\frac{7}{2},\frac{7}{2})$

The doctor ('she') considers a_M the best course of action, while the patient ('he') prefers to take a_H if the doctor's information is $\boldsymbol{\mu}_s$, and a_L otherwise. In a cheap talk setting, there are effectively two distinct equilibria: one that yields actions a_L and a_H in the appropriate



(a) expected utility of action profiles

(b) indirect utility given misperception

Figure 1: Misperception leads to non-convexities in indirect utility as a function of μ_0 (i.e., $V(X|\boldsymbol{a}^*,\mu_0)$, with $\boldsymbol{a}^*$ the patient's perceived optimal choice at each μ_0).

cases, and a 'babbling' equilibrium that results in a_M regardless of the message of the doctor. Farrell and Rabin convincingly argue that the former is somewhat implausible. It is only sustained because the patient interprets any message as indicating one of the two information states. The doctor should make it clear that she is not conveying information and thus achieve her preferred outcome. We argue, however, that both outcomes can result from a doctor optimally moderating the information flow. The interaction can be interpreted as a doctor facing a patient misperceiving the accurateness of the exam. If the patient is naive, this results in the revealing equilibrium, while if the patient is sophisticated, the doctor can induce the babbling equilibrium.

Example 1.1. Suppose the strategic game is a condensed representation of the following setting: the patient is either healthy (ω_L) or has a disease (ω_H). He can continue as usual without treatment (a_L), take an appropriate mild treatment but otherwise continue as normal (a_M), or follow an aggressive course of action (a_H) with additional changes in lifestyle, etc. The doctor conducts a diagnostic test that yields signals $\boldsymbol{s}$ and $\boldsymbol{t}$ with $s_H = 0.6 = t_L$. Let payoffs be $u(a_H|\omega_H) = 5 = u(a_L|\omega_L)$, $u(a_M|\omega_H) = u(a_M|\omega_L) = \frac{7}{2}$, and 0 otherwise. For simplicity, rather than relying on the vector notation, let μ_0 and p_0 describe the ex-ante probability of having the disease, i.e., the true state being ω_H. If $\mu_0 = p_0 = \frac{1}{2}$, the test is not informative enough to condition actions on the results. If, however, the patient (incorrectly) believes the accuracy is higher, i.e., $s_H^d = 0.8 = t_L^d$, he will strictly prefer to condition actions on the test, with a_H and a_L taken after the corresponding result. Figure 1 visualizes the actual and perceived expected utilities.

The patient knows the doctor has observed one of the outcomes. If he is naive, he doesn't take into account any possible moderation and interprets every message as indicating one of those results. The doctor's best course of action is to convey the result accurately ($m(\boldsymbol{s}) = \boldsymbol{s}$, $m(\boldsymbol{t}) = \boldsymbol{t}$) despite the patient's misperception and suboptimal action choices. There exists no

beneficial moderation policy. If, however, the patient is sophisticated, the doctor can send a sufficiently garbled message, so that the patient no longer considers it accurate enough to condition outcomes on it (e.g., $m(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{s} + \frac{1}{2}\boldsymbol{t} = m(\boldsymbol{t})$). This leads to the best course of action from the doctor's point of view. ◇

The beneficial intervention becomes possible due to the discontinuities and hence non-convexities in $V(X|\boldsymbol{a}^*, \mu_0)$. This is analysed in more detail in the Online Appendix (C.1). In particular, Result A.1 shows how both misperception and biases cause such non-convexities for at least some beliefs.[10] As a result, expected utility is non-monotone in the Blackwell-informativeness of X, meaning more information is not unambiguously beneficial.[11] Forwarding all information accurately is no longer necessarily the dominant strategy. Misperception renders the interaction strategic, even though preferences are fundamentally aligned. Naivete and sophistication correspond to different variations of the resulting strategic communication game.

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. As Example 1.1 already demonstrated, the decision maker's choice behavior and signal perception constrain the influence of the moderator. This section formally identifies this constraint as well as the moderator's optimization problem.

Upon observing the outcome of an experiment X, a moderator with a prior $\boldsymbol{\mu}$ reaches a posterior belief in the set $\{\boldsymbol{\mu}(\boldsymbol{s}_i)\}_{i=1}^k$. For the same experiment, but given a distortion d and a prior $\boldsymbol{p}$, a DM reaches a belief in $\{\boldsymbol{p}(\boldsymbol{s}_i^d)\}_{i=1}^k$. Let P denote the convex hull of the possible posterior beliefs of the DM, and Q for the moderator. A moderation policy can only yield posterior beliefs inside these two sets. Take an element $\hat{\boldsymbol{q}} \in Q$. By definition, this can be written as a convex combination of beliefs in $\{\boldsymbol{\mu}(\boldsymbol{s}_i)\}_{i=1}^k$. Denote the corresponding convex weights by the column vector $\boldsymbol{w}^{\boldsymbol{d}} = (w_1^d, ..., w_k^d)$. Any such weights can also be applied to $\{\boldsymbol{p}(\boldsymbol{s}_i^d)\}_{i=1}^k$, which yields a belief $\hat{\boldsymbol{p}} \in P$, noting that the sets have the same cardinality. And for any belief, there is at least one action $\hat{a} \in \mathbb{A}$ that maximizes expected utility from the perspective of the DM (i.e., $U(\hat{a}|\hat{\boldsymbol{p}})$). We can thus define a choice correspondence that maps each convex weight $\boldsymbol{w}^{\boldsymbol{d}}$ to the utility maximizing choice(s) at the corresponding belief in P:

$$C(\boldsymbol{w}^d) \equiv \Big\{\hat{a} \in \mathbb{A} \;\Big|\; \hat{a} = \arg\max_{a \in A} U(a|\hat{\boldsymbol{p}}),\ \hat{\boldsymbol{p}} = \sum_{i=1}^{k} w_i^d\, \boldsymbol{p}(\boldsymbol{s}_i^d)\Big\}. \tag{7}$$

[10] This echoes Brandenburger et al. (1992), which establishes the equivalence of distortions and heterogeneous priors in a correlated equilibrium setting. The equivalence, however, requires giving up Bayes' consistency of distorted beliefs.

[11] The unambiguous benefit of information refers to a case where observing X is costless. Conditions for a negative marginal return of information if its acquisition is costly have been discussed in Radner and Stiglitz (1984) and Chade and Schlee (2002).

However, a garbling of X given $\boldsymbol{\mu}$ might not yield the same convex combination of posterior beliefs as the same garbling of X^d given $\boldsymbol{p}$. Nevertheless, the moderated experiments XM and X^dM pin down the posterior beliefs and corresponding convex combinations. There thus exists a function that relates the convex weights, and hence posterior beliefs, between moderator and DM. Theorem 1 below identifies this relation. If the weights to obtain each element in Q were unique, any element of Q could be directly associated with exactly one element in P and the utility maximizing choices consistent with it. However, this is only the case if $|S_X| = \dim(X) \leq |\Omega|$. In general, the element in P corresponding to some $\boldsymbol{q} \in Q$ depends on the weights used to obtain $\boldsymbol{q}$.

As $C(\boldsymbol{w}^d)$ does not necessarily yield a singleton, we need to define the (maximum) expected utility of a set of actions $\hat{A} \subseteq A$ given a belief $\boldsymbol{q}$:

$$\overline{U}(\hat{A}|\boldsymbol{q}) \equiv \max\{U(a|\boldsymbol{q}) \mid a \in \hat{A}\}. \tag{8}$$

This allows us to look at the problem of finding a beneficial moderation policy entirely from the perspective of the moderator. For any set of beliefs in the convex hull of the moderator's posterior beliefs $\{\boldsymbol{\mu}(\boldsymbol{s}_i)\}_{i=1}^k$, we can find convex weights that generate these. Using the corresponding weights for the DM, the choice correspondence and $\overline{U}$ (sophisticated), or the original choices and U (naive), we can compute the maximum expected utility (from the moderator's perspective) at each of these beliefs. Note that the sender-preferred equilibrium allows the moderator to induce their most-preferred action from the correspondence.

Given a distribution over this set of beliefs, we can compute the expected utility at the prior. It is well understood that for any set of beliefs that contains the prior in its convex hull, we can find a distribution that is Bayes' consistent (e.g., Shmaya and Yariv (2016)). To establish the link to the corresponding beliefs of the DM, however, we need to identify these by the garbling they are obtained with. These are then necessarily Bayes' consistent and can be achieved with a moderation policy. If for any of these distributions expected utility exceeds the one from the original experiment and choices (i.e., $V(X|\boldsymbol{a},\boldsymbol{\mu})$, where $\boldsymbol{a}$ are the DM's utility maximizing choices given X^d and $\boldsymbol{p}$), then there exists a beneficial moderation policy. The optimal moderation policy maximizes expected utility among these distributions.

Theorem 1. *Suppose given an experiment X, distortion d, and priors $\boldsymbol{p}$ and $\boldsymbol{\mu}$, the DM chooses an action profile $\boldsymbol{a}$. Then there exists a beneficial moderation policy if and only if there exists a row-stochastic matrix $M = (m_{ij})_{1\leq i,j\leq k} \neq I$ such that:*

- *(naive)* $\sum_{i=1}^k \pi_i \cdot U(a_i|\boldsymbol{q}_i) > V(X|\boldsymbol{a},\boldsymbol{\mu})$,
- *(sophisticated)* $\sum_{i=1}^k \pi_i \cdot \overline{U}(C(\boldsymbol{w}_i^d)|\boldsymbol{q}_i) > V(X|\boldsymbol{a},\boldsymbol{\mu})$,

where $\pi_i = \sum_{j=1}^k m_{ji}\langle\boldsymbol{\mu},\boldsymbol{s}_j\rangle$, $\boldsymbol{q}_i = \sum_{j=1}^k w_{ji}\boldsymbol{\mu}(\boldsymbol{s}_j)$, $w_{ji} = \frac{m_{ji}\langle\boldsymbol{\mu},\boldsymbol{s}_j\rangle}{\sum_{j=1}^k m_{ji}\langle\boldsymbol{\mu},\boldsymbol{s}_j\rangle}$, $w_{ji}^d = \frac{m_{ji}\langle\boldsymbol{p},\boldsymbol{s}_j^d\rangle}{\sum_{j=1}^k m_{ji}\langle\boldsymbol{p},\boldsymbol{s}_j^d\rangle}$, *and* $\boldsymbol{w}_i^d$ *the column-vector* $(w_{1i}^d, ..., w_{ki}^d)$.

The problem of finding the optimal moderation policy is thus equivalent to finding a Bayes' consistent distribution over posteriors in Q, with the chosen actions dictated by the corresponding beliefs of the DM. And those beliefs can be obtained through the garbling that generates the moderator's distribution over beliefs (Proposition 8 in Appendix B completes this argument based on Blackwell (1951)). This problem would look identical if preferences were not aligned and the moderator would simply want to induce actions more beneficial to them. But as Example 1.1 demonstrated and will be shown in greater detail, even for aligned preferences, solutions will often be non-trivial. Nevertheless, it should be noted that given the beliefs, distortions, and biases, solutions are not necessarily consistent with both cases. In other words, for a given set of beliefs, some moderation policies might only be consistent with misaligned preferences. Results in Section 4.4, for instance, allow for such a distinction.

Theorem 1 also highlights the key difference between a naive and a sophisticated decision maker. Unaware of any moderation, no interference by the moderator can modify the set of chosen actions of a naive DM. The moderator can only affect when and how often these choices are executed, i.e., 'pick' among the actions that are chosen by the DM after some signal. In the most extreme case, by garbling one signal entirely into another, actions can be effectively removed but no new action can be introduced. The situation is more complex for a sophisticated DM who reacts to the moderation policy. Any convex combination of beliefs (subject to the garbling restriction) can be mapped to a corresponding choice arising from DM's maximization problem and signal perception. This alters not only the relative frequency of actions but also (potentially) affects the set of choices. This gives the moderator additional freedom in terms of actions but simultaneously constrains when and how frequently they can be induced. It is therefore not immediately clear if sophistication increases the benefit from moderation. Whether or not this benefits moderation depends on whether the alternative choices are superior from the perspective of the moderator.

As a first key observation, despite the complexity of the problem, binary relations play an important role. For some distinct $i, j \in \{1, ..., k\}$, let $\mathbf{a}_{i\to j}$ be the modified action profile based on $\mathbf{a} = (a_1, ..., a_k)$, where a_i replaces a_j, and all other actions remain unchanged. We say the moderator prefers action profile $\boldsymbol{a}$ over some $\boldsymbol{a}'$, if it achieves higher expected utility.

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

Beneficial moderation requires disagreement between the DM and moderator regarding the expected utility ranking of an action that is chosen and one that could be induced. For a naive DM, all instances of beneficial moderation can be identified by comparing the DM's preferred action to that of the moderator, with the set of actions restricted to those chosen

by the DM at some belief. For a sophisticated DM, cases of such disagreement is generically sufficient but not necessary for beneficial moderation to be feasible.[12] A moderator can switch some signals, and hence actions, with a sufficiently low probability without altering the action profile itself. But there are instances where this is not beneficial and yet inducing another action might be. Lemma 1 thus reveals that a moderator can beneficially intervene in more instances if the DM is sophisticated.

In the next section, we explore how this relates to the underlying signals. This allows us to address more specifically how misperception and biases affect the gain from information, what this implies for the optimal moderation policy, and the extent to which moderation can benefit each type.

4.3 Moderation & the gain from information

Moderation is a form of garbling and as such renders an experiment (weakly) less Blackwell informative.[13] When the maximum expected utility is convex in beliefs, this cannot increase the value of an experiment. The aim is thus to determine when biases and misperception create non-convexities that can be addressed by garbling signals, and to further characterize the optimal garbling.

It it is 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 2 provides a schematic representation. Denote by $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ the experiment generated from X, with 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) beliefs, knowing that either $\boldsymbol{s}_i$ or $\boldsymbol{s}_j$ has occurred, are denoted by $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{p}_{ij}$. Note that if $|S_X| = 2$, then this simply reduces to X, $\boldsymbol{\mu}$, and $\boldsymbol{p}$ respectively.

Deterministic moderation

We begin with the simplest moderation policy - a deterministic policy that replaces one (or several) signal(s) with another from S_X. If, for instance, $\boldsymbol{s}_i$ is replaced with $\boldsymbol{s}_j$, then after observing $\boldsymbol{s}_j$, an unbiased observer can only conclude that either of the signals has occurred, which leads to a posterior $\boldsymbol{\mu}_{ij}$. 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 takes a_j after both $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. This removal of information

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

[13] Formally, an information experiment $X \in \mathbb{X}$ is more informative than an information experiment $Y \in \mathbb{X}$, with $|S_X| = |S_Y| = |S|$, if there exists a $|S| \times |S|$ row-stochastic matrix $G \neq I$ such that $Y = XG$. In this case, we say the two experiments are Blackwell-ordered.

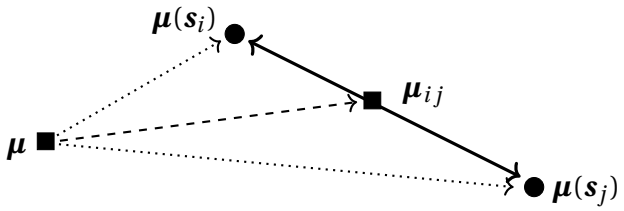


Figure 2: Starting from a prior $\boldsymbol{\mu}$, the experiment X results in a posterior $\boldsymbol{\mu}(\boldsymbol{s}_i)$ if signal $\boldsymbol{s}_i$ occurs (equivalently for $\boldsymbol{s}_j$). The solid line illustrates the 'relative information' between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Starting from $\boldsymbol{\mu}_{ij}$, experiment $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ leads to a posterior $\boldsymbol{\mu}(\boldsymbol{s}_i)$ or $\boldsymbol{\mu}(\boldsymbol{s}_j)$.

is beneficial if the conditional gain from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is negative at $\boldsymbol{\mu}_{ij}$, i.e., in the case where the DM's failure to effectively use the relative information leads to a choice sufficiently far from the optimum. In contrast, a sophisticated DM is aware of the removal of information and thus chooses the optimal action at $\boldsymbol{p}_{ij}$ (the conditional belief knowing that either $\boldsymbol{s}_i^d$ or $\boldsymbol{s}_j^d$ has realized). This is beneficial if the gain from $X_\triangle^m(\boldsymbol{s}_i, \boldsymbol{s}_j)$ 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 $\boldsymbol{\mu}_{ij}$.

Proposition 1. *A beneficial deterministic moderation policy exists*

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

for some signals $\boldsymbol{s}_i, \boldsymbol{s}_j \in S_X$.

While the optimal deterministic policy might affect more than two signals, a loss from a binary experiment at an intermediate belief is nevertheless necessary and sufficient for beneficial (deterministic) moderation to exist for a naive, and sufficient for a sophisticated DM. Proposition 1 thus links Lemma 1 to the relevant informational aspects of the experiment. A deterministic moderation policy is, of course, extreme in its effect on the information content of signals. Nevertheless, as Corollary 1.1 shows, it is optimal for a naive decision maker.

Corollary 1.1. *The optimal moderation policy for a naive DM is deterministic.*

Non-deterministic moderation

Even for a sophisticated decision maker, a deterministic moderation policy can be optimal. Figure 3 (**a**), (**b**) schematically illustrates this case. Panel (**a**) highlights which choices are optimal from the DM's perspective, and (**b**) what the moderator considers optimal. After signal $\boldsymbol{s}_i$, the DM and moderator disagree about the optimal action: the moderator prefers action a_2 over a_1. Moreover, a_2 is also the preferred action by both moderator and DM at $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{p}_{ij}$ respectively. From the moderator's point of view, the information from $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is not valuable. As the DM takes the 'correct' action at $\boldsymbol{p}_{ij}$, 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 the two signals. Aware of this garbling, the DM can no longer distinguish which of the two signals signals has occurred and hence takes the moderator's preferred action a_2. 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. The DM correctly interprets the (perceived) signals as less informative and reacts accordingly.

As was already foreshadowed by the discussion of Lemma 1, a deterministic moderation policy might neither be the only, nor the optimal way to influence a sophisticated DM. Figure 3 (**c**), (**d**) depicts a scenario where a non-deterministic policy is uniquely optimal for a sophisticated DM and, in fact, achieves higher expected utility than any possible moderation policy for a naive DM. By garbling some $\boldsymbol{s}_j$ signals into $\boldsymbol{s}_i$, the moderator moves the posterior belief close enough to $\boldsymbol{p}_{ij}$ so that a_1 becomes optimal for the DM. The posterior belief after observing $\boldsymbol{s}_j$ remains unaffected (even though the belief itself becomes less likely).[14]

Finally, 3 (**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 preferred a_2 to a_0 at $\boldsymbol{\mu}(\boldsymbol{s}_i)$, a naive DM would do strictly better under their optimal (deterministic) policy $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$ than the sophisticated DM, whose optimal moderation policy is non-deterministic.

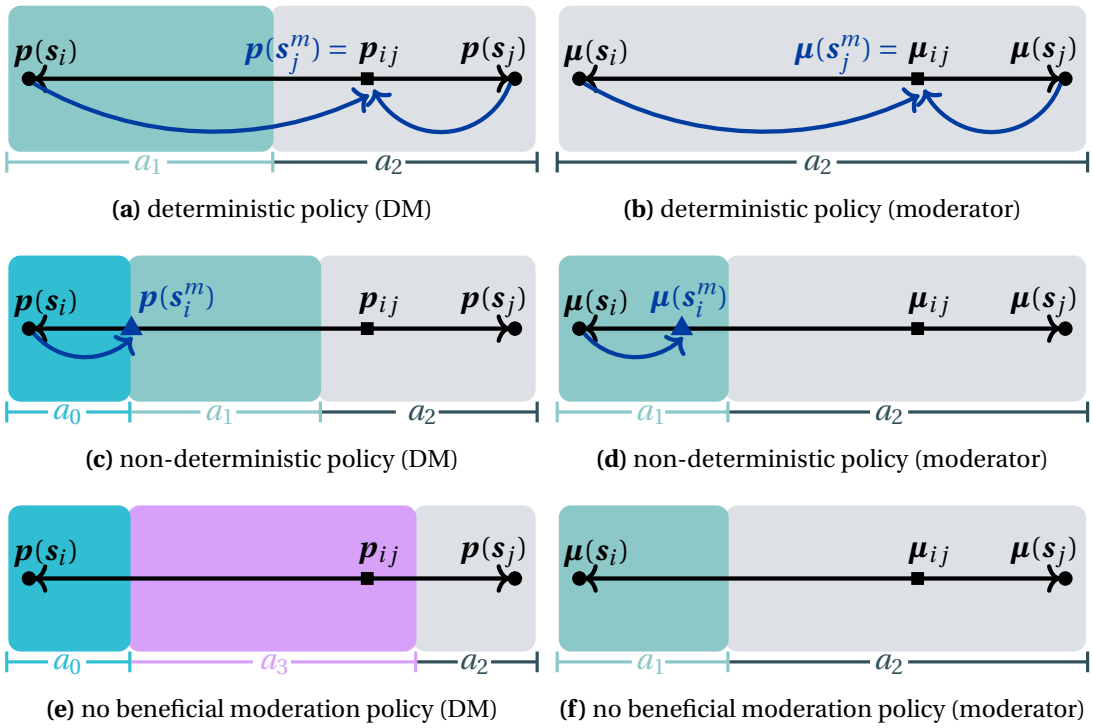
This discussion suggests that the optimal moderation policy sharply differs from what we established for a naive DM. In contrast to Corollary 1.1, Proposition 2 shows that if a decision maker is sophisticated, a non-deterministic moderation policy is always weakly better and uniquely optimal in some cases.

Proposition 2. *Suppose the DM is sophisticated. Then for every beneficial deterministic policy, there exists a non-deterministic policy that achieves (weakly) higher expected utility. The set of optimal policies might not include a deterministic policy.*

Example 1.2 illustrates the differences in optimal moderation policies in a specific case and highlights the ambiguous effect of sophistication, particularly given a bias in prior.

Example 1.2. Consider a slightly modified version of Example 1.1. Suppose now $u(a_M|\omega_H) = u(a_M|\omega_L) = 3.75$, with all other payoffs as before. Signals are such that $s_H = 0.75 = t_L$, and there is no misperception. The patient can benefit from the diagnostic test by choosing a profile (a_L, a_M) for low/intermediate priors, and (a_M, a_H) for higher priors

[14] 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 case, a sophisticated DM benefits more from moderation than a naive one.



(a) deterministic policy (DM)

(b) deterministic policy (moderator)

(c) non-deterministic policy (DM)

(d) non-deterministic policy (moderator)

(e) no beneficial moderation policy (DM)

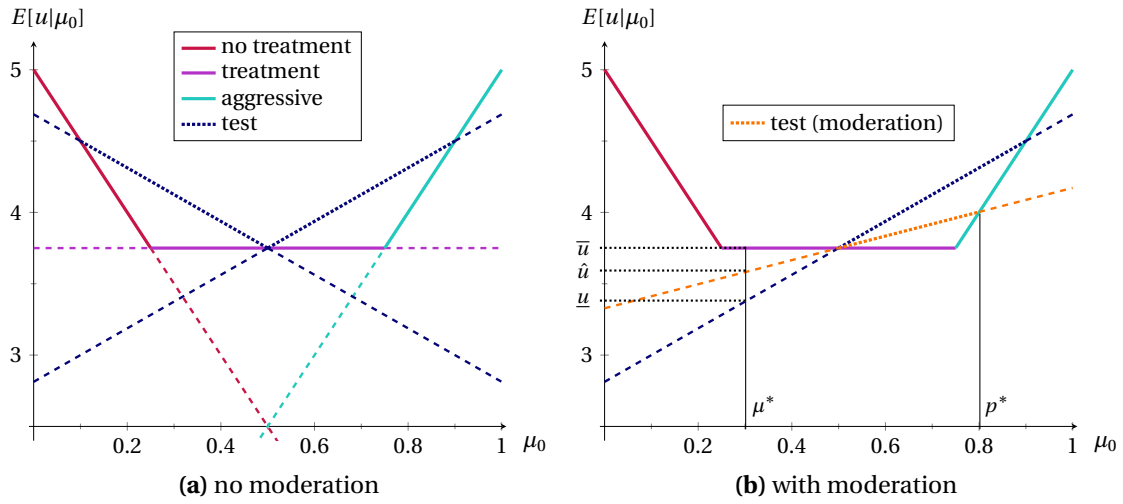
(f) no beneficial moderation policy (moderator)

Figure 3: (In-)feasibility of beneficial moderation with a sophisticated DM

(see Figure 4 (**a**)). Suppose the doctor concludes $\mu_0 \in (0.25, 0.5)$, while the patient believes $p_0 \in (0.5, 0.75)$. 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_H) 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 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 ($p_0 \in (0.75, 0.9)$), 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 4 (**b**)). It turns positive into negative test results with a just high enough probability, such that the patient is indifferent between (a_H, a_H) and (a_M, a_H). For priors μ^* and p^*, the optimal moderation policy yields $\hat{u}$ instead of $\underline{u}$. 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 is strictly better off. ◇

An important aspect in determining whether sophistication poses a (dis-)advantage is



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

Figure 4: Expected utility of action profiles (Example 1.2)

whether or not the moderator and the DM agree about the default action, i.e., the action taken when all relative information is removed. In Figure 3 (**e**) and (**f**), as well as in Example 1.2 / Figure 4 (**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 3 (**c**) and (**d**) illustrates a case where the moderator and DM agree on the default action. Since 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 2.1 formalizes this observation.

Corollary 2.1. *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 DM's adjustments in choices 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 it 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, 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 becomes 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 conditional beliefs. 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

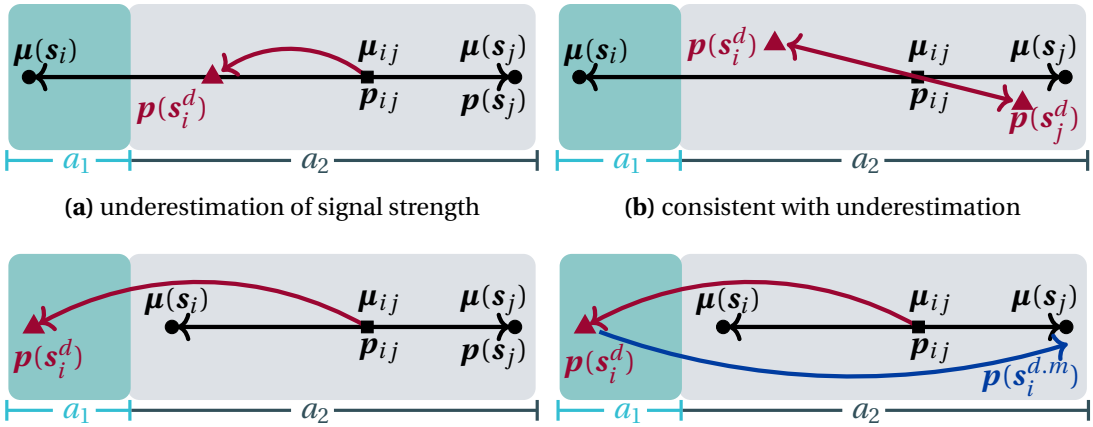
If there are only two states of the world, any distortion can be described as either an over- or underestimation of signal strength. Furthermore, correlations between signals and states are either retained or reversed. For each $\boldsymbol{s}$, the DM updates 'too much' ($|\boldsymbol{\mu}(\boldsymbol{s}^d) - \boldsymbol{\mu}| > |\boldsymbol{\mu}(\boldsymbol{s}) - \boldsymbol{\mu}|$), 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 $\boldsymbol{a} = (a_1, ..., a_k)$, 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 chosen 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 choice can be rationalized with a $X_{\triangle}(\boldsymbol{s}_i^d, \boldsymbol{s}_j^d)$ that is less Blackwell-informative than $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$. The definition further requires that the correlation between states and signals is not reversed. Any choice inconsistent with such a relative underestimation implies that the DM wrongly believes that 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 3 establishes that the latter is a requirement for beneficial moderation to be feasible. It is subsequently demonstrated that the 'relative' perspective is crucial.

Proposition 3. *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 5 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.



(a) underestimation of signal strength

(b) consistent with underestimation

(c) overestimation of signal strength

(d) beneficial moderation (naive DM)

Figure 5: (In-)feasibility of beneficial moderation

As an immediate implication, in a binary setting with only two states and two signals, any distortion that makes a signal appear less precise to the DM (without reversing the correlation between signals and states, as captured by the the second condition in Corollary 3.1), does not allow for beneficial moderation.

Corollary 3.1. *Suppose $|\Omega| = |S_X| = 2$ and $\boldsymbol{\mu} = \boldsymbol{p}$. Then there exists a beneficial moderation policy only if for some $\boldsymbol{s}_i \in S_X$:*

$$\frac{s^d_{i.\omega}}{s^d_{i.\omega'}} > \frac{s_{i.\omega}}{s_{i.\omega'}} > 1 \quad \textit{or} \quad \frac{s_{i.\omega}}{s_{i.\omega'}} > 1 > \frac{s^d_{i.\omega}}{s^d_{i.\omega'}}.$$

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 straightforward: suppose the distortion is

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 as the distortion generally causes the DM's intermediate belief $\boldsymbol{p}_{il}$ to differ from $\boldsymbol{\mu}_{il}$ for a signal $\boldsymbol{s}_l \neq \boldsymbol{s}_i, \boldsymbol{s}_j$. At $\boldsymbol{\mu}_{il}$, 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 3 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 being equally likely ex ante. There are two actions, a_1 and a_2, resulting in payoffs $u(a_1, \omega_1) \geq u(a_1, \omega_2) > u(a_1, \omega_3)$, and $u(a_2, \omega_1) = u(a_2, \omega_3) > u(a_2, \omega_2)$. There are three symmetric signals $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$, 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 5, a_1 is optimal after signals $\boldsymbol{s}_1$, and $\boldsymbol{s}_2$, while a_3 is optimal after $\boldsymbol{s}_3$.[15]

The (naive) DM instead perceives a distorted experiment, where 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 of the entire experiment.[16] 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 to 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 were mostly concerned with moderation causing a strict reduction in informativeness of signals. For a naive DM, optimal moderation in those cases is rather 'heavy handed'. As the optimal moderation policy is deterministic, the relative information between all signals that are garbled into each other is destroyed. Furthermore, they all lead a naive DM to the same posterior, which corresponds to one of the posteriors in the absence of moderation. The DM is being (completely) misinformed in at least some cases. Optimal interventions for a sophisticated DM are (weakly) less aggressive. A moderator only destroys as much information as necessary to induce superior choices. As a sophisticated DM takes into account the reduced informativeness, posteriors become less distorted. Nevertheless,

[15] The graph is based on the following values: $u(a_1, \omega_1) = 10, u(a_1, \omega_2) = 5, u(a_1, \omega_3) = 0$, 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 all other states.

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

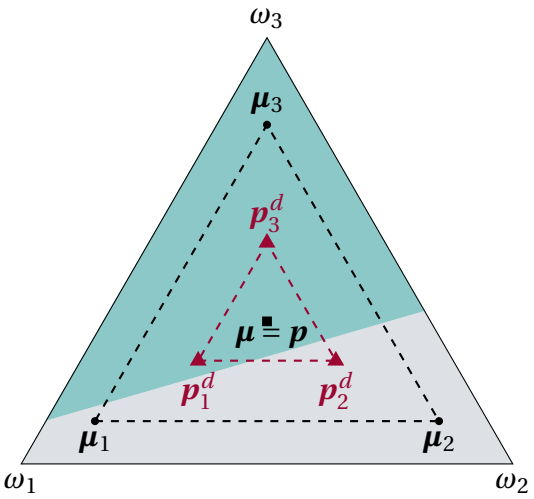


Figure 6: Posteriors of three distorted and undistorted signals in a belief simplex (Example 2). Posteriors after a signal $\boldsymbol{s}_i$ are denoted by $\boldsymbol{\mu}_i$ and $\boldsymbol{p}_i$. The DM underestimates the signal strength of every signal. Beneficial moderation remains possible as a_2 is not consistent with an underestimation of $\boldsymbol{s}_1$ relative to $\boldsymbol{s}_2$.

there are cases where the complete destruction of information is more beneficial than a partial one.

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 swapping each signal and thus inducing 'reversed' posteriors. From a strategic perspective, the interaction would appear to an outside observer akin to a 0-sum game. This would be hardly surprising in a sender-receiver game when preferences are opposed, but here the sender (i.e., moderator) and receiver (i.e., DM) agree about which action should be taken in which state. Of course, if the 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 even when the moderator and DM agree, at least in principle, about the information content of the signals. In fact, we show that such disagreement can arise solely as a result of a bias in prior.

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 $\boldsymbol{a}_{i\to j}$ denotes a single substitution, $\boldsymbol{a}_{i\leftrightarrow j}$ describes a symmetric one.

Definition 3. *Given a chosen action profile $\boldsymbol{a}$, the moderator and DM are in complete dis-*

agreement if there exists $\boldsymbol{a}_{i\leftrightarrow j}$ *such that*

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

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

Intuitively, complete disagreement becomes possible 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. Lemma 2 fully characterizes geometrically when an information environment allows for complete disagreement.

Lemma 2. *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.*

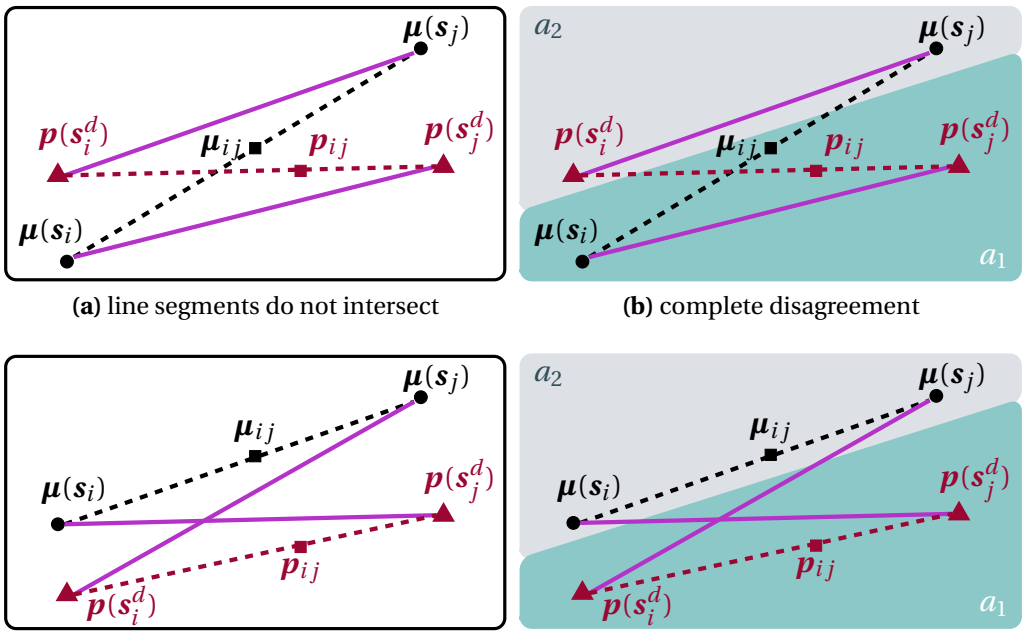
Figure 7 schematically depicts the two cases. In (**a**), the distortion and bias lead to a clockwise rotation of the posterior beliefs relative to the moderator's. The relevant line segments (solid lines) do not cross. As shown in (**b**), there are preferences such that the beliefs $\boldsymbol{\mu}(\boldsymbol{s}_i)$ 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. There are no preferences that lead to complete disagreement; (**d**) depicts a particular example.

While the geometric characterization in Lemma 2 is complete, it 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 between a DM and moderator differs is then reflected by differences between the corresponding 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 prior* $\boldsymbol{p}$ *and distortion* d*, is the linear space spanned by the vectors* $\{\boldsymbol{v}_1,...,\boldsymbol{v}_k,\boldsymbol{w}_1,...,\boldsymbol{w}_k\}$*, where* $\boldsymbol{v}_i = \boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}$ *and* $\boldsymbol{w}_i = \boldsymbol{p}(\boldsymbol{s}_i^d) - \boldsymbol{\mu}$*.*

These vectors then 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



(a) line segments do not intersect

(b) complete disagreement

(c) line segments intersect

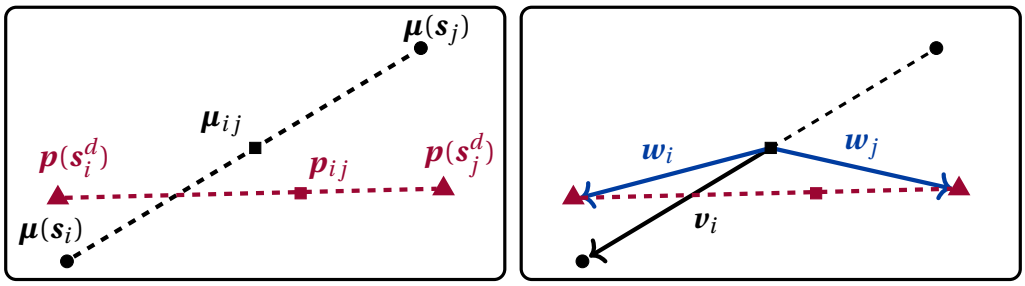
(d) no complete disagreement

Figure 7: Possibility of complete disagreement

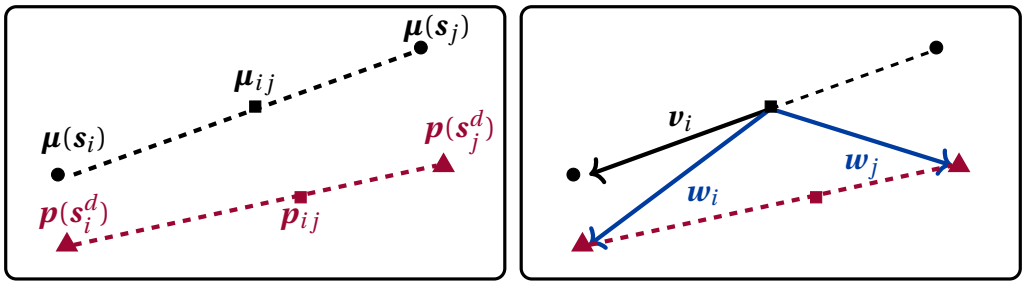
any $\alpha > 0$ and all $i \in \{1, ..., k\}$, i.e., the direction of the update is not fully reversed. Naturally, the definitions for belief space and non-reversal can be applied to any $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$. The dimensions of the belief space for a given $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ has direct implications for the possibility for complete disagreement. For a full characterization, we require one additional property:

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 are a basis 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 experiment $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$. The belief update from some $\boldsymbol{\mu}_{ij}$ to $\boldsymbol{\mu}(\boldsymbol{s}_i)$, as shown in Figure 8 (**a**), can be described by the vector $\boldsymbol{v}_i = \boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}_{ij}$. For the DM, the same signal (perceived distortedly) leads to the posterior $\boldsymbol{p}(\boldsymbol{s}_i^d)$, captured by $\boldsymbol{w}_i = \boldsymbol{p}(\boldsymbol{s}_i^d) - \boldsymbol{\mu}_{ij}$ in the belief space. The equivalent is true for $\boldsymbol{s}_j$. If $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ satisfy opposing orientation relative to $\boldsymbol{v}_i$, then - loosely speaking - one points to the left of $\boldsymbol{v}_i$ and the other to the right. If we think of a hypothetical line through $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and $\boldsymbol{\mu}(\boldsymbol{s}_j)$ (and thus also $\boldsymbol{\mu}_{ij}$), then $\boldsymbol{p}(\boldsymbol{s}_i^d)$ must be on one side, and $\boldsymbol{p}(\boldsymbol{s}_j^d)$ on the other. This is the case in (**b**). In a sense, the distortion introduces perceived information that is not just orthogonal to the experiment $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$, but that acts in opposing directions on the posteriors after $\boldsymbol{s}_i^d$ and $\boldsymbol{s}_j^d$. This creates a rotation in the posterior beliefs relative to $\boldsymbol{\mu}_{ij}$. Panel (**c**) shows a similar case, but here the rotation is small compared to the shift in beliefs. Vectors $\boldsymbol{w}_i$ and



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



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

Figure 8: Opposing orientation illustrated

$\boldsymbol{w}_j$ have the same orientation relative to $\boldsymbol{v}_i$. Note that Figure 8 depicts the two cases already illustrated in Figure 7. The constellation of posterior beliefs in (**a**/**b**) allows for complete disagreement, while no such disagreement is possible in (**c**/**d**). This points towards the key role of orientation for complete disagreement.

Theorem 2 formalizes this relation and fully characterizes (non-trivial) cases of complete disagreement. For a belief space of $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ relative to $\boldsymbol{\mu}_{ij}$, also let $\boldsymbol{w}_0 = \boldsymbol{p}_{ij} - \boldsymbol{\mu}_{ij}$.

Theorem 2. *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 $\boldsymbol{a} = (a_1, ..., a_k)$ 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 the opposing orientation property is satisfied by one of the following:*
  - *(i) $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ relative to $\boldsymbol{v}_i$, or*
  - *(ii) $\boldsymbol{v}_i - \boldsymbol{w}_0$ and $\boldsymbol{v}_j - \boldsymbol{w}_0$ relative to $\boldsymbol{w}_i - \boldsymbol{w}_0$, or*
  - *(iii) $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ relative to $\boldsymbol{w}_0$.*
- *If $z = 3$, there always exist preferences such that there is complete disagreement.*

If the belief space of some $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is 1-dimensional, then beliefs cannot be rotated relative to each other. Hence, there cannot be complete disagreement. Except, of course, in the trivial case where a distortion completely reverses the correlation between signals and states. This is ruled out by non-reversal. It follows that non-trivial cases of complete disagreement require Ω to contain at least three states.

Corollary 3.2. *Under non-reversal, there can be complete disagreement only if* $|\Omega| \geq 3$.

With only two states, the belief space of any experiment is 1-dimensional, ruling out (non-trivial) rotations. In a 2-dimensional 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. To easily verify opposing orientation, Appendix A (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).[17] If this is the case, 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. This echoes the observation by Alonso and Câmara (2016), that a richer state space, in particular $|\Omega| \geq 3$, creates additional possibilities for persuasion under heterogeneous priors.

Moderating complete disagreement

Since Bernoulli utilities are fully aligned between the decision maker and moderator, complete disagreement is based on a different understanding how information is to be interpreted: it is as if correlations between states and some signals $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are completely reversed. For $|\Omega| \geq 3$, an actual reversal is not required, however. In fact, either a small distortion that leads to a rotation of beliefs or a small difference in prior beliefs is sufficient. Since complete disagreement implies a negative (conditional) gain from some experiment $X_{\triangle}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ at $\boldsymbol{\mu}_{ij}$ (Lemma 3 in Appendix A.1), beneficial moderation is generically possible for naive and sophisticated DMs.

Proposition 4. *Suppose that given a chosen action profile* $\boldsymbol{a}$*, the moderator and DM are in complete disagreement over some* a_i *and* a_j*. Then a beneficial moderation policy for both a naive and a sophisticated DM generically exists.*

If there is complete disagreement regarding some action profile $\boldsymbol{a}$, an alternative action profile can be constructed from the choices in $\boldsymbol{a}$ that achieves strictly positive conditional gain. In principle, 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 experiments X and X^m are equally (Blackwell) informative. This constitutes a strictly beneficial moderation policy and yet the naive DM ends up completely misinformed in terms of posterior beliefs. For a sophisticated DM, 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,

[17] Note that the vectors $\boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}_{ij}$ and $\boldsymbol{\mu}(s_j) - \boldsymbol{\mu}_{ij}$ are necessarily linearly dependent.

which potentially leaves a naive DM strictly better-off. Corollary 4.1 formalizes this for a particular case, i.e., $\boldsymbol{a}_{i\leftrightarrow j}$ is the (unconstrained) optimal action profile and a_i or a_j is unique in $\boldsymbol{a}$.[18] While sophistication makes a decision maker harder to manipulate, it can also limit beneficial interventions.

Corollary 4.1. *Suppose given a chosen action profile $\boldsymbol{a} = (a_1, ..., a_k)$, we have $V(X|\boldsymbol{\mu}) = V(X|\boldsymbol{a}_{i\leftrightarrow j}, \boldsymbol{\mu})$, and either a_i or a_j are not equal to any other $a \in \{a_1, ..., a_k\}$. Then the optimal moderation policy for a naive DM generically achieves strictly higher expected utility than the optimal policy for a sophisticated DM.*

To illustrate this further, suppose there is complete disagreement over some actions a_i and a_j. 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 strategic response by a sophisticated DM 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 (symmetrically) swapping signals achieves the first-best. Such a swap can be directly implemented for a naive DM. As sophistication prevents a relabeling, for a sophisticated DM this swap can be only achieved if garbling $\boldsymbol{s}_i$ with some $\boldsymbol{s}_l$ induces a_j, and equivalently for $\boldsymbol{s}_j$. If, however, the information provided by at least one of the corresponding signals is strictly valuable relative to any other signal,[19] meaning $V(X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_l)|\boldsymbol{\mu}_{il}) > \max_{a\in\mathbb{A}} E[u(a|\omega)|\boldsymbol{\mu}_{il}]$ for all $\boldsymbol{s}_l \neq \boldsymbol{s}_i$, then even if the swap can be achieved, it causes a strict utility loss. A naive DM is then strictly better-off. This also provides a complementary perspective to Corollary 2.1, 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 4.2. *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 without distortions

Complete disagreement follows from a sufficiently distinct interpretation of signals between the moderator and DM. Proposition 5 shows that it does not require any distortion at all, but can arise from a bias in prior alone. Moreover, if $|\Omega| \geq 3$ and signals have distinct probability ratios across states, then a bias in prior that leads to complete disagreement necessarily

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

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

exists. Furthermore, we can find such a bias even when difference in beliefs are only ϵ-small (Lemma 4 in Appendix A.1).

Proposition 5. *Let $|\Omega| \geq 3$ 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 illustrates a case where complete disagreement arises only from a difference in priors and, in line with Corollary 4.2, a naive DM benefits strictly more from moderation.

Example 3.1. A firm is considering an applicant for a position and can either hire them (a_H) or not (a_N). The firm employs an outside HR consultancy to conduct an assessment and provide a recommendation. On the assessment test, 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 ($\boldsymbol{\mu}$), with priors of $\boldsymbol{p} = (0.1, 0.6, 0.3)^T$ and $\boldsymbol{\mu} = (0.6, 0.1, 0.3)^T$.

Figure 9 visualizes the posterior beliefs and payoff-maximizing actions in the belief simplex. The 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 candidate. The consultancy, however, takes a negative result at face value and sees it as evidence for an unskilled applicant. The reactions to a positive test result are 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

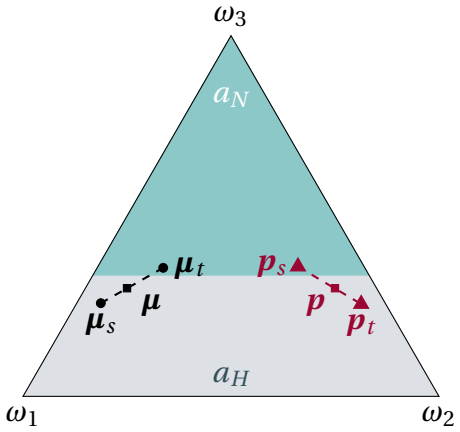


Figure 9: Prior and posterior beliefs of HR consultancy (moderator) and employer (DM). The assessment test induces complete disagreement.

performance. Reversing the test results ($m(s) = t$, $m(t) = s$) 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 1.2, 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 Discussion

Welfare. The Welfare analysis in this paper is conducted from the perspective that the moderator holds a more accurate prior belief and has a better understanding of the information that signals reveal about the state of the world. We believe in many contexts, the assumption that an expert suffers from fewer perception issues appears at least reasonable. In these settings, given that information is purely instrumental, the moderation policy implemented by a benevolent moderator will result in the DM making better choices on average. One might plausibly interject that not everybody in the position to moderate information has necessarily a more accurate view of the information environment. After all, a firm might have superior information about which skills make an applicant a truly good fit. In settings where the moderator has a different, but not more accurate view, the welfare interpretation 'flips'. In particular, as follows from Blackwell (1951), if the DM's view is more accurate, any non-trivial moderation policy that either cannot be undone, or is

not being undone (i.e., the case of a naive DM with complete disagreement), has a strictly negative welfare effect. Possibilities for beneficial moderation turn into instances where a DM's utility is reduced, despite aligned preferences. Similarly for the comparison across types, rather than sophistication allowing for more instances where a DM can be helped, it facilitates more situations where a well-intentioned moderation policy proves harmful. And more beneficial interventions for naive types become more destructive. Nevertheless, the analysis presented here might prove useful to distinguish these cases.

State dependent misperception. Misperception is described throughout as an inaccurate view of the information experiment, meaning the DM is under the impression they face an experiment X^d, rather than X. While this can arise from a subjective probability assessment, it might also be directly related to how a realized state affects perception. For example, a patient might conduct a medical test perfectly if healthy, but fails to accurately follow instructions when sick, possibly due to reduced cognitive performance.[20] Misperception becomes state-dependent. Since no restriction was placed on d that would rule-out such forms of misperception, our analysis includes this case. Even though Blackwell garbling - and thus by extension a moderation policy - is state independent, it can be employed to improve upon state-dependent errors. For more specific applications, modelling misperception explicitly might provide additional structure that allows for further insight. For example, as in the medical example, it is entirely plausible that in some contexts X^d and X fully coincide for all but one state. The moderator might then want to garble signals if a particular state is sufficiently likely. This could potentially be exploited to derive results more specific to the application.

Preferences vs. beliefs. We identified and characterized settings in which a moderator would want to misinform a decision maker. Even small differences in prior and/or the perception of an information experiment can lead to complete disagreement, meaning 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 two parties have opposing interests. But as shown, this can be caused by differences in how new information is interpreted. As demonstrated in Example 1.1, interactions may appear/become strategic despite identical preferences. This raises the question whether a distinction between the underlying causes is even necessary. We want to highlight, though, that despite the observationally similar consequences, policy implications can differ. A policy maker interested in the Welfare of a (rational) decision maker would want to implement mechanisms that maximize the amount of information that is released. The role of an expert is merely to administer the test. In the presence of biases and misperception, on the other hand, full disclosure is not necessarily the optimal policy and discretion should be left to the expert. Furthermore, despite the

[20] We thank an anonymous referee for this suggestion.

overlap, the settings are not entirely equivalent. If beliefs are observed, the conditions for complete disagreement allow for the possibility to distinguish when a specific moderation policy could not possibly be motivated by differences in perception alone.

From a more general perspective, this analysis also points to the danger of judging individuals' apparent information misperception. What might appear to one person as an irrational choice could simply be based on slight heterogeneity in how information is perceived, especially in more complex settings with more than two states. From a modelling perspective, reducing problems to a binary setting can (potentially) lead to incorrect inference, even if some states are payoff equivalent, as in Example 3.1.

Interaction of biases and misperception. Outside of the strategic interpretation, this analysis also sheds light on how biases and perception mistakes interact in non-monotone ways. A bias in prior pushes the decision maker to an alternative course of action. While suboptimal in a perfect world, this can mitigate some of the negative impact of misperception; either because the alternative is less sensitive to (incorrectly perceived) information, or because misperception acts as beneficial moderation. Reconsider Example 3.1, which demonstrates how a biased prior can lead to the choice of an action profile that, evaluated at the true prior, is strictly worse than choosing either of its actions independent of the signal. Adding any misperception that sufficiently reduces the signal strength has a positive impact.[21] Alternatively, imagine a decision maker who sometimes erroneously perceives one signal for another, possibly through incorrect recall, or a perception error as in Rabin and Schrag (1999), where a conformation bias is modelled in this way. Such an error effectively acts as a form of moderation. So if there exists a beneficial moderation policy, there also exist a beneficial, signal-swapping error. Furthermore, under complete disagreement, *any* such error is strictly beneficial, even if it just amounts to additional white-noise. And a DM unaware of such errors can be better-off than one who factors them in. Section C.2 in the Online Appendix discusses this more formally. In this context, rather than being a more knowledgeable expert, the moderator could be seen as a metaphor for a cognitive error that alleviates existing biases and distortions.

6 Conclusion

We analyzed the effects of two fundamental mistakes in information processing and how they can be mitigated by a moderator who 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 fully aligned, a decision maker can be

[21] Indeed, an earlier version of this paper focused on these topics in more detail. Example 5.1 in the Online Appendix, which documents how biases and distortions introduce non-convexities, is the typical case where a bias in prior mitigates the downsides from distorted perception.

better-off if information is garbled and destroyed. Furthermore, the knowledge about such garbling can have an heterogeneous impact on the DM's welfare. Even though the analysis is phrased as a strategic communication problem, it equally highlights non-monotonicities in the interaction of biases in prior with perception mistakes. The results thus also characterize when adding or intensifying biases and misperception can have a positive impact rather than a 'double whammy' effect.

A Additional Results

A.1 Complete disagreement

Lemma 3. *Suppose given an action profile $\boldsymbol{a}$, the 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 4. *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 even though (ii) implies (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

$$\begin{aligned} \boldsymbol{s}_\omega &> \boldsymbol{\mu}_\omega \boldsymbol{s}_\omega + \boldsymbol{\mu}_h \boldsymbol{s}_h + \boldsymbol{\mu}_l \boldsymbol{s}_l + \sum_{\omega' \in \Omega^- \setminus \omega} \boldsymbol{\mu}_{\omega'} \boldsymbol{s}_{\omega'} \\ 0 &> \boldsymbol{\mu}_h \cdot (\boldsymbol{s}_h - \boldsymbol{s}_\omega) - \boldsymbol{\mu}_l \cdot (\boldsymbol{s}_\omega - \boldsymbol{s}_l) + \sum_{\omega' \in \Omega^- \setminus \omega} \boldsymbol{\mu}_{\omega'} \cdot (\boldsymbol{s}_{\omega'} - \boldsymbol{s}_\omega) \end{aligned} \tag{9}$$

Setting $\boldsymbol{\mu}_{\omega'} = 0$ 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

$$\frac{\boldsymbol{\mu}_l}{\boldsymbol{\mu}_h} > \frac{\boldsymbol{s}_h - \boldsymbol{s}_\omega}{\boldsymbol{s}_\omega - \boldsymbol{s}_l} \tag{10}$$

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.[22]

(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 (9) 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 $\boldsymbol{\mu}_\omega = \boldsymbol{p}_\omega$.

□

A.2 Belief space & orientation

Lemma 5 provides an equivalent but potentially more intuitive definition of 'opposing orientation', when the vector space is $\mathbb{R}^2$. This only requires the comparison of the sign of the determinant of the relevant matrices constructed from the vectors in coordinate form.

Lemma 5. *Let $\boldsymbol{v}$, $\boldsymbol{w}$, and $\boldsymbol{x}$ be vectors in $\mathbb{R}^2$. Then $\boldsymbol{v}$ and $\boldsymbol{w}$ have opposing orientation relative to $\boldsymbol{x}$ if and only if the matrices*

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

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

[22] 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).

Proof. Opposing orientation requires that the sets $\{\boldsymbol{x}, \boldsymbol{v}\}$ and $\{\boldsymbol{x}, \boldsymbol{w}\}$ both are a basis and there exists a matrix L with $\{\boldsymbol{x}, \boldsymbol{w}\} = \{L\boldsymbol{x}, L\boldsymbol{v}\}$ and $\det(L) < 0$. As a basis, they span $\mathbb{R}^2$ and $\det(A) \neq 0$ and $\det(B) \neq 0$. 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. □

Using this, we establish the necessity and sufficiency of the opposing orientation property of three sets of vectors for the possibility of complete disagreement in $\mathbb{R}^2$ (i.e., when the belief space is 2-dimensional). The three different comparisons are needed as they correspond to different possibilities of lines separating the beliefs $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and $\boldsymbol{p}(\boldsymbol{s}_j^d)$ from $\boldsymbol{\mu}(\boldsymbol{s}_j)$ and $\boldsymbol{p}(\boldsymbol{s}_i^d)$. If such a separating line exists, then there are preferences with an indifference curve coinciding with that line that yield complete disagreement. Intuitively, misperception and/or a bias in prior needs to induce a belief update that is orthogonal and in an opposing direction for both signals in order to allow for complete disagreement. To stress this, Proposition 6 phrases the opposing orientation property in terms of an orthogonal vector, and subsequently shows the equivalence of this formulation with the notion in Definition 5.

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. Furthermore, suppose without loss that $\boldsymbol{v}_0 = \boldsymbol{0}$. If that wasn't the case, we could apply a translation vector $-\boldsymbol{v}_0$. Further define $\Delta\boldsymbol{v}_1 \equiv \boldsymbol{v}_1 - \boldsymbol{w}_0$, $\Delta\boldsymbol{v}_2 \equiv \boldsymbol{v}_1 - \boldsymbol{w}_0$, and $\Delta\boldsymbol{w}_1 \equiv \boldsymbol{w}_1 - \boldsymbol{w}_0$, as well as $\Delta\boldsymbol{w}_2 \equiv \boldsymbol{w}_1 - \boldsymbol{w}_0$.

Proposition 6 (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) $\langle \boldsymbol{u}, \boldsymbol{v}_1 \rangle = 0$, *while* $\boldsymbol{w}_1 = a_1\boldsymbol{v}_1 + b_1\boldsymbol{u}$, *and* $\boldsymbol{w}_2 = a_2\boldsymbol{v}_1 + b_2\boldsymbol{u}$, *or*

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

(iii) $\langle \boldsymbol{u}, \boldsymbol{w}_0 \rangle = 0$, *while* $\boldsymbol{w}_1 = a_1\boldsymbol{w}_0 + b_1\boldsymbol{u}$, *and* $\boldsymbol{v}_1 = a_2\boldsymbol{w}_0 + b_2\boldsymbol{u}$.

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

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

where

$$\boldsymbol{v}_1 = \begin{pmatrix} v_{1,1} \\ v_{2,1} \end{pmatrix} \qquad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \qquad \boldsymbol{w}_i = \begin{pmatrix} a_i v_{1,1} + b_i u_1 \\ a_i v_{2,1} + b_i u_2 \end{pmatrix}.$$

We can compute $\det(A) = b_1(v_{1,1}u_2 - v_{2,1}u_1)$, and $\det(B) = b_2(v_{1,1}u_2 - v_{2,1}u_1)$. As $b_1 < 0 < b_2$, $\det(A)$ and $\det(B)$ are of opposing sign. They thus describe linear maps of different orientation (Lemma 5). Consequently, the cross products $\boldsymbol{v}_1 \times \boldsymbol{w}_1$ and $\boldsymbol{v}_1 \times \boldsymbol{w}_2$ have opposite signs (i.e., point to opposite sides relative to $\boldsymbol{v}_1$). Since V_0 lies on a line between V_1 and V_2, and as $\boldsymbol{v}_0 = \boldsymbol{0}$, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly dependent and we can write $\boldsymbol{v}_2 = -\kappa \boldsymbol{v}_1$ for some $\kappa \in \mathbb{R}^+$ and accordingly $V_2 = V_1 - \kappa \boldsymbol{v}_1$. Note further that $\boldsymbol{v}_1 \times (\boldsymbol{w}_2 - \boldsymbol{v}_1) = \boldsymbol{v}_1 \times \boldsymbol{w}_2$ (i.e., translations along the direction of $\boldsymbol{v}_1$ cannot affect the sign). Furthermore, $\boldsymbol{v}_1 \times (\boldsymbol{w}_1 - \boldsymbol{v}_2) = \boldsymbol{v}_1 \times (\boldsymbol{w}_1 + \kappa \boldsymbol{v}_1)$, which then again must have the same sign as $\boldsymbol{v}_1 \times \boldsymbol{w}_1$. We can further write $\mathbb{P}_2$ as $\{V_1 - \kappa \boldsymbol{v}_1 + \lambda(\boldsymbol{w_1} - \boldsymbol{v_2}) | \lambda \in [0,1]\}$. As $V_1 \neq V_2$, and hence $\kappa \neq 0$, and as $(\boldsymbol{w_1} - \boldsymbol{v_2})$ and $(\boldsymbol{w_2} - \boldsymbol{v_1})$ point in opposite directions relative to $\boldsymbol{v}_1$, the sets $\mathbb{P}_1$ and $\mathbb{P}_2$ describing the line segments $\overline{V_1 W_2}$ and $\overline{V_2 W_1}$ 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. Again note that we can write $\boldsymbol{v}_2 = -\kappa \boldsymbol{v}_1$ for some $\kappa > 0$. We can thus write $\boldsymbol{v}_2 = a_3 \boldsymbol{w}_0 + b_3 \boldsymbol{u}$, where $b_3 < 0$ (i.e., the same sign as b_1). Construct the matrices

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

Computing the determinants, we can conclude that $\text{sign}(\det(C)) = \text{sign}(\det(D))$. But then the cross products $\boldsymbol{w}_0 \times \boldsymbol{w}_1$ and $\boldsymbol{w}_0 \times \boldsymbol{v}_2$ have the same sign (i.e., point to the same side relative to $\boldsymbol{w}_0$). As W_0 lies on a line strictly between W_1 and W_2, the vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ must have the opposing orientation relative to $\boldsymbol{w}_0$. It follows that the cross products $\boldsymbol{w}_0 \times \boldsymbol{w}_2$ must have the opposing sign of $\boldsymbol{w}_0 \times \boldsymbol{w}_1$. Furthermore, using again $\boldsymbol{v}_2 = -\kappa \boldsymbol{v}_1$, we can establish that $\boldsymbol{w}_0 \times \boldsymbol{v}_1$ must have the opposing sign of $\boldsymbol{w}_0 \times \boldsymbol{v}_2$. It follows that $\boldsymbol{w}_0 \times \boldsymbol{v}_1$ and $\boldsymbol{w}_0 \times \boldsymbol{w}_2$ also have the same sign, but the opposite sign compared to the previous two cross products. Furthermore, $\boldsymbol{w}_0 \times \boldsymbol{w}_2 = \boldsymbol{w}_0 \times (\boldsymbol{w}_2 - \boldsymbol{w}_0)$ and $\boldsymbol{w}_0 \times \boldsymbol{w}_1 = \boldsymbol{w}_0 \times (\boldsymbol{w}_1 - \boldsymbol{w}_0)$ (i.e., a translation along $\boldsymbol{w}_0$ cannot change the sign of the cross product). This implies V_1 and W_2 lie on the same side of the line segment $\overline{V_0 W_0}$, while V_2 and W_1 both lie on the opposite side. Sets $\mathbb{P}_1$ and $\mathbb{P}_2$ are disjoint as required.

Necessity:

We prove the contrapositive. Suppose (i), (ii), and (iii) are not satisfied. Using the previous argument, it follows from the violation of (iii) that V_1 and W_1 must lie on the same side of the line segment $\overline{V_0 W_0}$, with V_2 and W_2 both on the opposite side. Then either (a) the line segments $\overline{V_1 W_1}$, $\overline{W_1 W_2}$, $\overline{V_2 W_2}$, and $\overline{V_1 V_2}$ 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 (see Figure 10 for illustration). 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 $\boldsymbol{w}_1 - \boldsymbol{v}_2 = a_1\boldsymbol{v}_1 + b_1\boldsymbol{u}$ and $\boldsymbol{w}_2 - \boldsymbol{v}_1 = a_2\boldsymbol{v}_1 + b_2\boldsymbol{u}$, 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 $\Delta\boldsymbol{w}_1$, 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. □

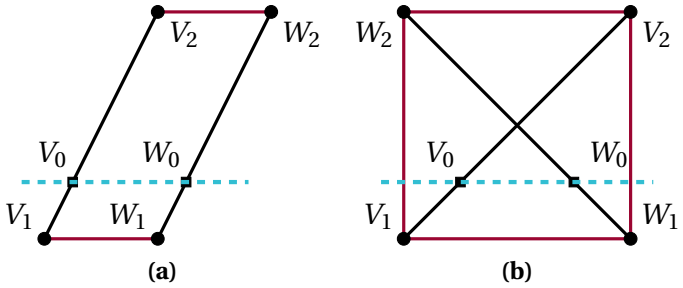


Figure 10: This illustrates the relevant quadrilaterals. In (**a**), the diagonals $\overline{V_1W_2}$ and $\overline{V_2W_1}$ cross. In (**b**), the vectors corresponding to the line segments $\overline{V_1W_1}$ and $\overline{V_1W_2}$ have opposing orientation relative to the vector corresponding to $\overline{V_1V_0}$, which leads to a contradiction.

A.3 Complete Disagreement & Blackwell-ordered Misperception

Complete disagreement does not just arise if actual and perceived experiments are difficult to compare, but even if the DM's distorted view is simply a noisier, garbled version of the original experiment, meaning X and X^d are strictly Blackwell-ordered. Furthermore, with $\text{rank}(X) \geq 3$, complete disagreement for such experiments is possible even without a bias in prior. Example 4.1 demonstrates a specific instance, and Proposition 7 provides a formal treatment.

Example 4.1. Let there be three states, $\{\omega_1, \omega_2, \omega_3\}$, and three signals, $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$, with the following signal likelihoods:

$$\boldsymbol{s}_1 = \begin{pmatrix} 0.8 \\ 0 \\ 0.2 \end{pmatrix} \quad \boldsymbol{s}_2 = \begin{pmatrix} 0.2 \\ 0.8 \\ 0 \end{pmatrix} \quad \boldsymbol{s}_3 = \begin{pmatrix} 0 \\ 0.2 \\ 0.8 \end{pmatrix} \quad X = \begin{pmatrix} | & | & | \\ \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 \\ | & | & | \end{pmatrix}.$$

The DM and moderator have the same prior belief, which views each state as equally likely. The DM's distorted view of the experiment, X^d, is a garbled version of the original experi-

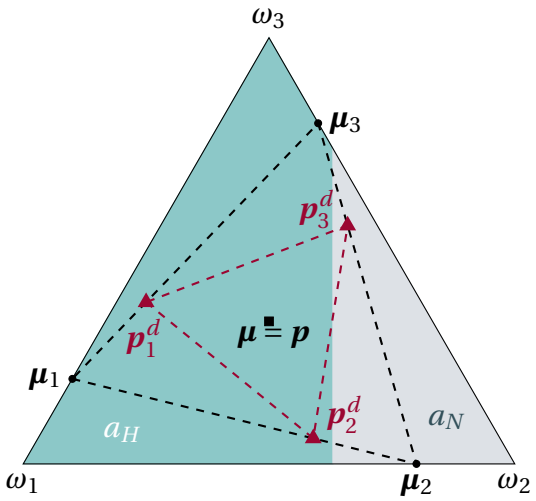


Figure 11: Complete disagreement for a distorted experiment X^d that is strictly nosier than X.

ment, i.e., $X^d = XG$. In particular

$$X^d = \begin{pmatrix} 0.56 & 0.38 & 0.06 \\ 0.06 & 0.56 & 0.38 \\ 0.38 & 0.06 & 0.56 \end{pmatrix}, \quad G = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0 & 0.7 & 0.3 \\ 0.3 & 0 & 0.7 \end{pmatrix}.$$

Finally, there are two actions, a_H and a_N, with payoffs $u(a_H|\omega_1) = 1.7$, $u(a_H|\omega_2) = 0$, $u(a_H|\omega_3) = 0.35$, and $u(a_N|\omega_1) = u(a_N|\omega_3) = 0$, $u(a_N|\omega_2) = 1$. ◇

Figure 11 depicts the example graphically. Due to the garbling of X, the DM's posteriors are less extreme. Moreover, the pairwise garbling of G results in posteriors for the DM that are a convex combination of the two respective posteriors of the moderator. Since a_H is preferred at $\boldsymbol{\mu}_3$ and $\boldsymbol{p}_2^d$ while a_N is preferred at $\boldsymbol{\mu}_2$ and $\boldsymbol{p}_3^d$, there is complete disagreement between the DM and moderator after signal $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$.

Proposition 7. *Suppose X is an $n \times k$ matrix with full rank, $\boldsymbol{p} = \boldsymbol{\mu}$, and beliefs satisfy non-reversal.*

- *If $k = 2$, there exists no garbling matrix G with $X^d = XG$, such that there is complete disagreement.*
- *If $k \geq 3$, there exists preferences and a garbling G with $X^d = XG$ such that there is complete disagreement.*

Proof. Suppose $k = 2$. Full rank of X implies $n = 2$. It follows from Proposition 2 that there can be no disagreement as beliefs satisfy non-reversal.
Now suppose $k \geq 3$. Consider the (sub)-set of signals $\hat{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\} \subseteq S_X$. Let $\boldsymbol{\mu}(\hat{S})$ be the intermediate belief, given that one of the three signals has occurred. Let $\boldsymbol{v}_i$ denote the vector in coordinate form corresponding to $\boldsymbol{\mu}(\boldsymbol{s}_i) - \boldsymbol{\mu}(\hat{S})$, given a basis $\{\boldsymbol{\mu}(\boldsymbol{s}_1) - \boldsymbol{\mu}(\hat{S}), \boldsymbol{\mu}(\boldsymbol{s}_2) - \boldsymbol{\mu}(\hat{S})\}$.

The full-rank assumption ensures that $\boldsymbol{\mu}(\boldsymbol{s}_3) - \boldsymbol{\mu}(\hat{S})$ can be expressed as a coordinate vector relative to this basis, while Bayes' consistency implies that $\boldsymbol{v}_3 < \mathbf{0}$.

Construct a row-stochastic matrix G as follows: G equals the identity matrix except $g_{3,2} > 0$ (and $g_{3,3} = 1 - g_{3,2}$), as well as $g_{2,1} > 0$ (and $g_{2,2} = 1 - g_{2,1}$). Denote the garbled signals by $\boldsymbol{s}_1^d, \boldsymbol{s}_2^d$, and $\boldsymbol{s}_3^d$ respectively. As X has full rank and X^d is a garbling of X with only signals in $\hat{S}$ affected, we can conclude that $\boldsymbol{\mu}(\hat{S}) = \boldsymbol{p}(\hat{S})$. It follows further that the vectors $\boldsymbol{\mu}(\boldsymbol{s}_1) - \boldsymbol{\mu}(\hat{S})$ and $\boldsymbol{\mu}(\boldsymbol{s}_2) - \boldsymbol{\mu}(\hat{S})$ span the corresponding belief space, which is two-dimensional. We can express the vectors $\boldsymbol{p}(\boldsymbol{s}_i^d) - \boldsymbol{\mu}(\hat{S})$ as coordinate vectors $\{\boldsymbol{w}_i\}_{i=1}^3$ relative to the same basis $\{\boldsymbol{\mu}(\boldsymbol{s}_1) - \boldsymbol{\mu}(\hat{S}), \boldsymbol{\mu}(\boldsymbol{s}_2) - \boldsymbol{\mu}(\hat{S})\}$.

Observe that by construction $\boldsymbol{w}_1 > \mathbf{0}$. Furthermore, as $\boldsymbol{v}_3 < \mathbf{0}$ and $\boldsymbol{w}_2$ is a convex combination of $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$, we have that $\boldsymbol{w}_2 < \boldsymbol{v}_2$, with the first coordinate strictly negative. Finally, $\boldsymbol{w}_3 = \boldsymbol{v}_3$ as $\boldsymbol{\mu}(\boldsymbol{s}_3) = \boldsymbol{p}(\boldsymbol{s}_3^d)$. Again by construction, $\boldsymbol{p}(\boldsymbol{s}_1^d)$ and $\boldsymbol{p}(\boldsymbol{s}_2^d)$ lie on the boundary of the convex hull of $\{\boldsymbol{\mu}(\boldsymbol{s}_i)\}_{i=1}^3$. However, as $\boldsymbol{w}_1 > \mathbf{0}$ while $\boldsymbol{w}_2 < \boldsymbol{v}_2$, any convex combination of the two must lie strictly inside the convex hull. Let $\boldsymbol{p}_{1,2}^d$ be the intermediate belief knowing that either $\boldsymbol{s}_1^d$ or $\boldsymbol{s}_2^d$ has occurred. This is such a convex combination and thus lies strictly inside the convex hull of $\{\boldsymbol{\mu}(\boldsymbol{s}_i)\}_{i=1}^3$. Denote the corresponding vector by $\boldsymbol{w}_0$. Let $\Delta\boldsymbol{v}_1 \equiv \boldsymbol{v}_1 - \boldsymbol{w}_0$, $\Delta\boldsymbol{v}_2 \equiv \boldsymbol{v}_2 - \boldsymbol{w}_0$ and $\Delta\boldsymbol{w}_1 \equiv \boldsymbol{w}_1 - \boldsymbol{w}_0$. By construction, $\Delta\boldsymbol{w}_1$ is a (non-degenerate) convex combination between $\Delta\boldsymbol{v}_1$ and $\Delta\boldsymbol{v}_2$. This implies that $\Delta\boldsymbol{v}_1$ and $\Delta\boldsymbol{v}_2$ must have the opposing orientation property relative to $\Delta\boldsymbol{w}_1$. The result then follows from Proposition 6. □

B Proofs

B.1 Beneficial moderation

Proof of Theorem 1. Any moderation policy for an experiment X can be expressed by a $k \times k$ garbling matrix M. By definition, M is row-stochastic. The moderated signals correspond to the experiment $X^m = XM$. Let $\boldsymbol{s}_i^m$ denoted the i-th column of this experiment. Applying Bayes' rule, we can verify that this moderated signal leads to a posterior $\boldsymbol{q}_i = \frac{\sum_j^k \boldsymbol{\mu}(\boldsymbol{s}_j) m_{ji} \langle \boldsymbol{\mu}, \boldsymbol{s}_j \rangle}{\sum_{j=1}^k m_{ji} \langle \boldsymbol{\mu}, \boldsymbol{s}_j \rangle}$. It follows from Bayes' consistency that the (ex-ante) probability of receiving such a signal $\boldsymbol{s}_i^m$ equals $\pi_i = \sum_{j=1}^k m_{ji} \langle \boldsymbol{\mu}, \boldsymbol{s}_j \rangle$. Now note that the belief $\boldsymbol{q}_i$ can be equally expressed as a convex combination of the posteriors $\{\boldsymbol{\mu}(\boldsymbol{s}_j)\}_{j=1}^k$. This yields convex weights that can be denoted by a column vector $\boldsymbol{w}_i = (w_{1i}, ..., w_{ki})$. By inspection, this is only consistent with a garbling of signals and the corresponding ex-ante probability $\boldsymbol{\pi}_i$ if $\boldsymbol{w}_{ji} = \frac{m_{ji} \langle \boldsymbol{\mu}, \boldsymbol{s}_j \rangle}{\sum_{j=1}^k m_{ji} \langle \boldsymbol{\mu}, \boldsymbol{s}_j \rangle}$. Note that these relations hold as an identity. Transferring this argument to the perspective of the DM yields the corresponding weights $\boldsymbol{w}_i^d = (w_{1i}^d, ..., w_{ki}^d)$ with $w_{ji}^d = \frac{m_{ji} \langle \boldsymbol{p}, \boldsymbol{s}_j^d \rangle}{\sum_{j=1}^k m_{ji} \langle \boldsymbol{p}, \boldsymbol{s}_j^d \rangle}$.

For the remaining result, we distinguish between naive and sophisticated DMs.

Naive: Since a naive DM does not adjust choices, signal $\boldsymbol{s}_i^m$ is followed by action a_i. The

expected utility from the perspective of the moderator of any moderation policy equals $\sum_{i=1}^{k}\sum_{j=1}^{k} m_{ji}\langle \boldsymbol{\mu}, \boldsymbol{s}_i\rangle \cdot U(a_i|\boldsymbol{\mu}(\boldsymbol{s}_j))$. It follows from linearity as well as the previous argument that this can be equivalently expressed as $\sum_{i=1}^{k} \pi_i \cdot U(a_i|\boldsymbol{q}_i)$, where q_i is the posterior belief corresponding to $\boldsymbol{s}_i^m$. Clearly, this is only beneficial if it is larger than $V(X|\boldsymbol{a},\boldsymbol{\mu})$. Necessity follows immediately, noting that any moderation policy must be expressible as a garbling.
Sophisticated: The argument follows analogous to the naive DM, except that after signal $\boldsymbol{s}_i^{d.m}$ (i.e., the i-th column of the moderated, distorted experiment $X^d M$), a sophisticated DM chooses an action $\hat{a}$ such that $\hat{a} = \arg\max_{a\in A}\{U(a|\boldsymbol{p}(\boldsymbol{s}_i^{d.m})\}$. Equations 7 and 8, and the definition of sender-preferred equilibrium, yield an expression for the expected utility (from the perspective of the moderator) of $\sum_{i=1}^{k} \pi_i \cdot \overline{U}(C(\boldsymbol{w}_i^{\boldsymbol{d}})|\boldsymbol{q}_i)$. The remaining argument corresponds to the one for the naive. □

Proposition 8. *For every set of beliefs* $\{\boldsymbol{q}_i\}_{i=1}^{k} \subset Q$ *and probability vector* $\boldsymbol{\pi}_i = (\pi_1,...,\pi_k)$ *with* $\sum_i^k \pi_i \boldsymbol{q}_i = \boldsymbol{\mu}$*, there exists a row-stochastic matrix* $M = \left(m_{ij}\right)_{1\le i,j\le k}$ *and an action profile* $\boldsymbol{a} = (a_1,...a_k)$ *such that* $\boldsymbol{q}_i = \frac{\sum_j^k \boldsymbol{\mu}(\boldsymbol{s}_j) m_{ji}\langle \boldsymbol{\mu}, \boldsymbol{s}_j\rangle}{\sum_{j=1}^k m_{ji}\langle \boldsymbol{\mu}, \boldsymbol{s}_j\rangle}$*, and* $a_i \in \arg\max_{a\in A} U(a|\boldsymbol{p}_i)$ *with* $\boldsymbol{p}_i = \frac{\sum_j^k \boldsymbol{p}(\boldsymbol{s}_j^d) m_{ji}\langle \boldsymbol{p}, \boldsymbol{s}_j^d\rangle}{\sum_{j=1}^k m_{ji}\langle \boldsymbol{p}, \boldsymbol{s}_j^d\rangle}$.

Proof. Let $\hat{X}$ be the experiment corresponding to the posterior beliefs $\{\boldsymbol{q}_i\}_{i=1}^{k}$. Any $\boldsymbol{q}_i \in Q$ is by definition a convex combination of posterior beliefs $\{\boldsymbol{\mu}(\boldsymbol{s}_j)\}_{j=1}^{k}$. It follows that $\hat{X}$ is (weakly) less Blackwell-informative than X and there exists a garbling matrix M such that $\hat{X} = XM$ (see Theorem 12.2.2., Blackwell and Girshick (1954)). It then follows from Theorem 1 that the posterior beliefs can be written as proposed. The chosen action follows directly from the DM's utility maximization problem. □

Proof of Lemma 1. *Naive DM - sufficiency*: For a naive DM, the policy $m(\boldsymbol{s}_j) = \boldsymbol{s}_i$, and m equal to the identity mapping for all other $\boldsymbol{s}_l \in S_X$, achieves an expected utility of

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

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 expectations and Bayes' consistency that

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

As $\boldsymbol{a}_{i\to j}$ is preferred to $\boldsymbol{a}$, we 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 $\boldsymbol{s}_j \in S_X$ with $m(\boldsymbol{s}_j) = \sum_{l=1}^{k} p_l \boldsymbol{s}_l$, where all $p_l \in [0,1]$ and $\sum_{l=1}^{k} p_l = 1$. 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 $i \neq j$ 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: Consider a 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. This achieves expected utility equal to:

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

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 expected utility,

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

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. □

B.2 Moderation & the gain from information

Lemma 6 briefly justifies the relevance of the binary perspective in the context of utility maximization.

Lemma 6. *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 $\boldsymbol{\mu}$.*

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:

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

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 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 1. Let $\boldsymbol{a}$ be the chosen action profile. We start with the *naive* DM.
Sufficiency: Suppose such signals $\boldsymbol{s}_i, \boldsymbol{s}_j$ exist. Then, $V(X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)|\boldsymbol{a}, \boldsymbol{\mu}_{ij}) - E[u(a_j|\omega)|\boldsymbol{\mu}_{ij}] < 0$, where we assumed wlog that a_j is the preferred action at $\boldsymbol{\mu}_{ij}$. It follows from Bayes'

consistency and linearity that $V(X_\triangle(\boldsymbol{s}_i,\boldsymbol{s}_j)|\boldsymbol{a}_{j\to i},\boldsymbol{\mu}_{ij}) = E[u(a_j|\omega)|\boldsymbol{\mu}_{ij}]$, Using the argument from Lemma 6, 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 that are part of $\boldsymbol{a}$ 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 6, $\boldsymbol{a}$ 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, $\boldsymbol{p}(\boldsymbol{s}_j^{d.m}) = \boldsymbol{p}_{ij}$. As a^* maximizes expected utility at $\boldsymbol{p}_{ij}$, a sophisticated DM chooses a^* after observing $\boldsymbol{s}_j^{d.m}$. A beneficial moderation policy exists. □

Proof of Corollary 1.1. This follows from the necessity argument in Proposition 1. For beneficial moderation to exist, there must be some actions a_i, a_j that are part of $\boldsymbol{a} = (a_1,...,a_k)$, 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 $m(\boldsymbol{s}_i) = \boldsymbol{s}_j$ achieves higher expected utility than any non-deterministic moderation policy, unless there is another a_l that dominates a_j at $\boldsymbol{\mu}(\boldsymbol{s}_i)$. But then the optimal policy has $m(\boldsymbol{s}_i) = \boldsymbol{s}_l$. With countable actions, the maximum is generically unique. Iterating this over all a_i gives the desired result. □

Proof of Proposition 2. Let m be a deterministic moderation policy given an experiment X with signals $S_X = \{\boldsymbol{s}_1,...,\boldsymbol{s}_k\}$. 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 $\boldsymbol{s}_l \in S_X$ with $m(\boldsymbol{s}_l) = \boldsymbol{s}_i$). As m is deterministic, $\boldsymbol{s}_i^m$ 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 $\boldsymbol{s}_l^m$ with $\boldsymbol{s}_l \in S_X^i$ (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 for all $\boldsymbol{s}_i \in S_X^i$, $\hat{m}(\boldsymbol{s}_i) = \frac{1}{K}\sum_{\boldsymbol{s}\in S_X^i} \boldsymbol{s}$, 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, $\boldsymbol{\mu}(\boldsymbol{s}_i^{\hat{m}}) = \boldsymbol{\mu}(\boldsymbol{s}_l^{\hat{m}})$ for all $\boldsymbol{s}_l \in S_X^i$. Each of these posteriors is generated with (ex-ante) probability $\langle \boldsymbol{s}_l^{\hat{m}}, \boldsymbol{\mu} \rangle = \langle \boldsymbol{s}_i^{\hat{m}}, \boldsymbol{\mu} \rangle = \frac{1}{K} \langle \boldsymbol{s}_i^{m}, \boldsymbol{\mu} \rangle$. 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 1.2 provides a specific example where a non-deterministic policy achieves strictly higher expected utility. This completes the proof. □

Proof of Corollary 2.1. The optimal moderation policy for a naive DM must be deterministic (Corollary 1.1). 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 1). 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 3. Trivially, a_i needs to be suboptimal for there to exist a beneficial moderation policy. Suppose now to the contrary that 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 consistent with an 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 1 and Corollary 1.1, there cannot be a beneficial moderation policy for a naive DM.
Sophisticated DM: For beneficial moderation, there needs to exist an action $\hat{a} \in \mathbb{A}$ such that $E[u(\hat{a}|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)] > E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_i)]$, and $\hat{a} = \arg\max_{a \in \mathbb{A}} E[u(a|\omega)|\boldsymbol{p}(\boldsymbol{s}_i^{d,m})]$, 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 (not necessarily unique) set $\{\boldsymbol{\mu}_{\alpha_j} | j \neq i, j \in \{1, ..., k\}\}$ of such beliefs that all lie on a hyperplane in the belief simplex Δ^{n-1}. By definition of $\boldsymbol{\mu}_\alpha$, this hyperplane separates Δ^{n-1} in two subsets: one containing $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and the other containing all other posteriors for signals in S_X. Denote the former by B_i and its complement by B_i^C. By construction, both must be convex. For moderation to

be beneficial, $\hat{a}$ must be optimal for some belief $\hat{\boldsymbol{\mu}} \in B_i$, otherwise it would never be chosen for any belief. By linearity, the set of beliefs for which $\hat{a}$ achieves higher expected utility than a_i must be a convex subset of B_i. Denote this by $\hat{B}_i$. Now note that for all $\boldsymbol{s}_j^d \neq \boldsymbol{s}_i^d$, by construction we have $\boldsymbol{p}(\boldsymbol{s}_j^d) \notin B_i$. This follows from the premise that all a_j must be consistent with an underestimation of signal strength at each $\boldsymbol{\mu}_{ji}$. Linearity 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 posterior beliefs of a sophisticated DM must be in the convex hull of the set of distorted posteriors $\{\boldsymbol{p}(\boldsymbol{s}_l^d)|\boldsymbol{s}_l^d \in S_{X^d}\}$, which is denoted by P. Since $\{\boldsymbol{p}(\boldsymbol{s}_l^d)|\boldsymbol{s}_l^d \in S_{X^d}\} \subset B_i^C$ and as B_i^C is a convex set, we can conclude that $\hat{B}_i \cap P = \emptyset$. No moderation policy can generate a belief $\boldsymbol{p}(\boldsymbol{s}_i^{d,m}) \in \hat{B}_i$ that induces a choice $\hat{a}$. No beneficial moderation policy exists. □

Proof of Corollary 3.1. Wlog, suppose that $\frac{s_{i,\omega}}{s_{i,\omega'}} > 1$, noting that we assumed that at least one signal must be informative. If the distortion is such that $\frac{s_{i,\omega}}{s_{i,\omega'}} \geq \frac{s_{i,\omega}^d}{s_{i,\omega'}^d} \geq 1$, then $|\mu_\omega(\boldsymbol{s}_i) - \mu_\omega| \geq |\mu_\omega(\boldsymbol{s}_i^d) - \mu_\omega|$. As $|\Omega| = 2$, we can then write $\boldsymbol{\mu}(\boldsymbol{s}_i^d) = \alpha\boldsymbol{\mu}(\boldsymbol{s}_i) + (1-\alpha)\boldsymbol{\mu}$ for some $\alpha \in [0,1]$. As $\boldsymbol{\mu} = \boldsymbol{p}$ and hence $\boldsymbol{\mu}(\boldsymbol{s}_i^d) = \boldsymbol{p}(\boldsymbol{s}_i^d)$, the result follows from Proposition 3. □

B.3 Complete disagreement

Proof of Lemma 2. 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.,

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

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

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

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 2. **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 $\boldsymbol{p}(\boldsymbol{s}_j^d)$ cannot lie between $\boldsymbol{\mu}(\boldsymbol{s}_i)$ and $\boldsymbol{p}(\boldsymbol{s}_i^d)$, and $\boldsymbol{p}(\boldsymbol{s}_i^d)$ cannot lie between $\boldsymbol{\mu}(\boldsymbol{s}_j)$ and $\boldsymbol{p}(\boldsymbol{s}_j^d)$. It follows from Lemma 2 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 6 (Vector Orientation). 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 then follows from Proposition 6.
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 2 that there exist preferences such that complete disagreement is possible. □

Proof of Corollary 3.2. Following Theorem 2, 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 4. It follows from Lemma 3 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 4.1. For a naive DM, the moderation policy $m(s_i) = s_j$ and $m(s_j) = s_i$ achieves an expected utility of

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

By definition, a sophisticated DM could at most achieve equal expected utility. Suppose the sophisticated DM attains the same expected utility as the naive. Following Proposition

2, if beneficial moderation is possible, there exists an optimal non-deterministic policy for a sophisticated DM. Given the assumption on either a_i or a_j, it is wlog to assume that a_i is unique in $\{a_1,...,a_k\}$. Suppose further that $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)] < E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$ for all $j \in 1,...,k$, $k \neq i$. This is generically true if a_i is unique in $\{a_1,...,a_k\}$. If $\boldsymbol{s}_j^m \neq \mathbf{0}$, then optimality requires that

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

as well as $\boldsymbol{s}_j^m = \boldsymbol{s}_j$. This 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, since m would have to be uniquely deterministic.

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 $\boldsymbol{s}_k \neq \boldsymbol{s}_j$. But this is again a contradiction. The result follows. □

Proof of Corollary 4.2. 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 4.1. □

Proof of Proposition 5. 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., they are the signals associated with $X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)$. From Lemma 4, it follows that there exist beliefs $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{p}_{ij}$ (strictly inside $\Delta(\Omega)$), such that $\text{sign}\big(\mu_\omega(\tilde{\boldsymbol{s}}_i) - \mu_\omega\big) = -\text{sign}\big(p_\omega(\tilde{\boldsymbol{s}}_i) - p_\omega\big)$, for some state $\omega \in \Omega$ with $\mu_{\omega.ij} = p_{\omega.ij}$. 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 (Theorem 2).

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 6, 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.

Stephen Ansolabehere, Marc Meredith, and Erik Snowberg. Asking about numbers: Why and how. *Political Analysis*, 21(1):48–69, 2013.

Olivier Armantier, Scott Nelson, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar. The price is right: Updating inflation expectations in a randomized price information experiment. *Review of Economics and Statistics*, 98(3):503–523, 2016.

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

Jean-Pierre Benoit and Juan Dubra. Apparent overconfidence. *Econometrica*, 79(5):1591–1625, 2011.

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.

David Blackwell and Meyer Girshick. *Theory of Games and Statistical Decisions*. John Wiley Sons, New York, NY, 1954.

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.

Hector Chade and Edward E. Schlee. Another look at the radner-stiglitz nonconcavity in the value of information. *Journal of Economic Theory*, 107(2):421–452, 2002.

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

Olivier Coibion, Yuriy Gorodnichenko, and Saten Kumar. How do firms form their expectations? new survey evidence. *American Economic Review*, 108(9):2671–2713, 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.

Juan Dubra. Optimism and overconfidence in search. *Review of Economic Dynamics*, 7(1): 198–218, 2004.

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.

Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic Perspectives*, 10(3):103–118, 1996.

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.

Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102(4):684–702, 1995.

Gerd Gigerenzer, Wolfgang Hell, and Hartmut Blank. Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):513–525, 1988.

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

Jerry Green and Nancy Stokey. A two-person game of information transmission. *Journal of Economic Theory*, 135(1):90–104, 2007.

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.

Ingar Haaland, Christopher Roth, and Johannes Wohlfart. Designing information provision experiments. *Journal Of Economic Literature*, 2023. forthcoming.

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.

Navin Kartik, Marco Ottaviani, and Francesco Squintani. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1):93–116, 2007.

Botond Kőszegi. Emotional agency. *Quarterly Journal of Economics*, 121(1):121–155, 2006.

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.

Andreas I. Mueller and Johannes Spinnewijn. Expectations data, labor market and job search. *Working Paper*, 2022.

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.

Tristan Potter. Learning and job search dynamics during the great recession. *Journal of Monetary Economics*, 117:706–722, 2021.

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.

Roey Radner and Joseph E. Stiglitz. A nonconcavity in the value of information. In Marcel Boyer and Richard E. Kihlstrom, editors, *Bayesian Models of Economic Theory*, pages 33–52. Elsevier, Amsterdam, 1984.

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.

Eran Shmaya and Leeat Yariv. Experiments on decisions under uncertainty: A theoretical framework. *American Economic Review*, 106(7):1775–1801, 2016.

Johannes Spinnewijn. Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association*, 13(1):130–167, 2015.

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.

Elias Tsakas and Nikolas Tsakas. Noisy persuasion. *Games and Economic Behavior*, 130: 44–61, 2021.

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

C Online Appendix

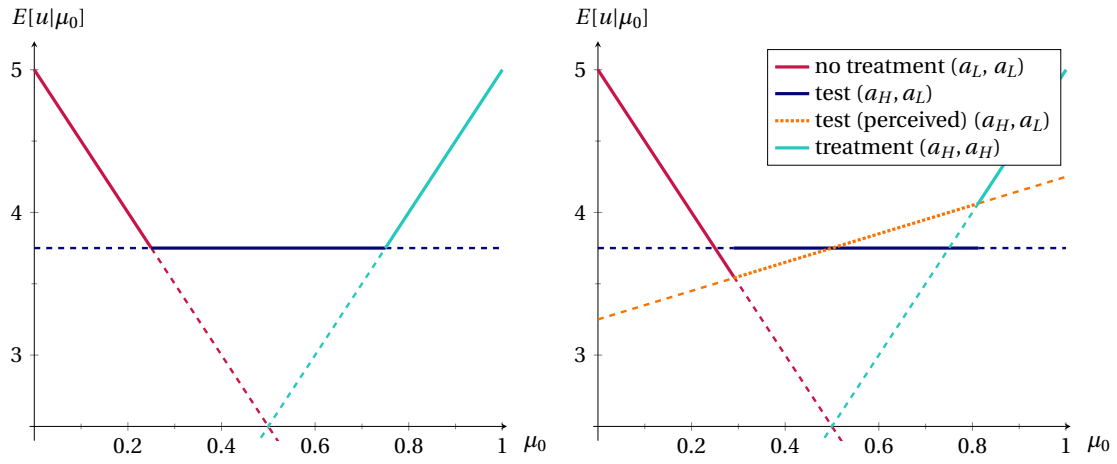
C.1 Distortions, biases and their implications

This section provides some additional depth to the observations made in Section 4.1, namely that unbeknownst to the decision maker, biases and distortions introduce non-convexities (in beliefs) in $V(X|\boldsymbol{a}^*, \boldsymbol{\mu})$ and thus alter the value and gain from experimentation. Consequently, the DM might fail to realise the full gain from an experiment either by relying too much or too little on the outcome. A DM with (only) a bias in prior judges the informativeness of signals correctly, but weighs them according to a different prior. This can lead to equally distorted posteriors and thus have similar effects. Examples 5.1 and 5.2 illustrate this and Result A.1 offers a formal summary.

Example 5.1. Analogous to Example 1.1, suppose a patient was potentially exposed to an infectious disease. The patient is either infected (ω_H) or not (ω_L), and can immediately seek treatment (a_H), or continue as usual (a_L). To check for infection, the patient can perform a diagnostic test, and react based on the outcome, i.e., $\boldsymbol{a} = (a_H, a_L)$. 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 $s_H = 0.75 = t_L$. Let $u(a_H|\omega_H) = 5 = u(a_L|\omega_L)$, and 0 otherwise. It follows from the symmetry of payoffs that, if the test is taken, the patient performs action a_H or a_L depending on whether the posterior belief after observing the result is greater or smaller than $\frac{1}{2}$.

Figure 12 (a) illustrates the expected utility outcomes as a function of beliefs. Profiles (a_H, a_H) and (a_L, a_L) 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 Result A.1, the maximum expected utility (bold line segments) is convex in μ_0.

Suppose now the patient misjudges the accuracy of the test by underestimating the chance of a false negative result and overestimating the probability of a false positive. In particular, $s_H = 0.75 < s_H^d = 0.85$ and accordingly $t_H = 0.25 < t_H^d = 0.15$. Furthermore, $t_L = 0.75 > t_L^d = 0.65$, and thus $s_L = 0.25 < s_L^d = 0.35$. 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. Conse-



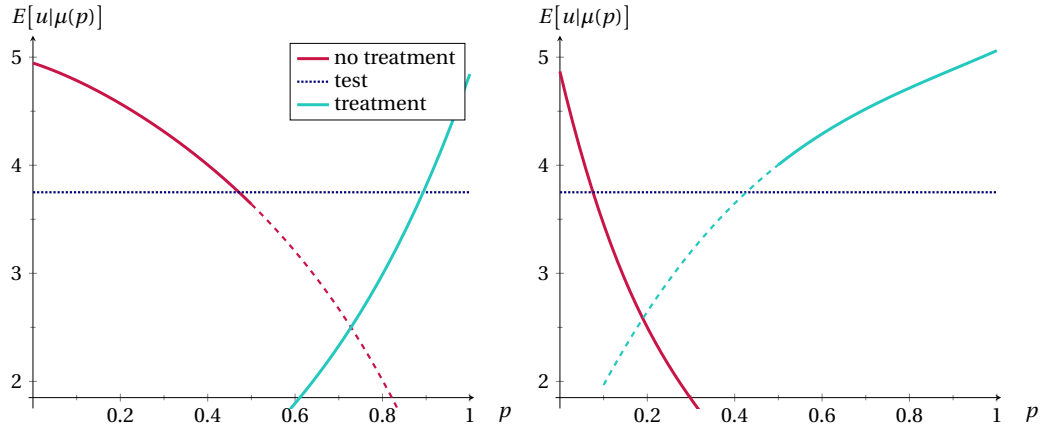
(a) no distortion **(b)** distorted accuracy

Figure 12: Expected utility of action profiles (Example 5.1)

quently, convexity of the true expected utility fails in regions where the information is only perceived to be valuable (in the neighbourhood of $\mu_0 = 0.8$). 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 possible infection. There is a range of prior beliefs where the test is not taken, despite being valuable (in the neighbourhood of $\mu_0 = 0.29$). 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. 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 5.2. 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 ($1/5 = \mu_0 < p_0 = 2/5$). 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 beliefs, e.g., $p(\boldsymbol{s}) \in (0.5, 0.75)$ in Figure 13 (a), the observer disagrees with the patient over which action should be taken. For an experiment



(a) prior beliefs: $\mu_0 = \frac{1}{5}$, $p_0 = \frac{2}{5}$ **(b)** prior beliefs: $\mu_0 = \frac{4}{5}$, $p_0 = \frac{2}{5}$

Figure 13: Expected utility (Ex. 5.2) 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.

to be considered valuable by the observer at $p = 2/5$, it would need a higher accuracy. A positive 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 Figure 13 (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. ◇

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

Result A.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 $\boldsymbol{p}$ if one of the following holds:*

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

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-1}. The maximum value from any experiment can also be written as a linear function in $\boldsymbol{\mu}$:

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

By the properties of the maximum, the combination of hyperplanes that achieve maximum expected utility is necessarily convex in $\boldsymbol{\mu}$.

Non-convexity from distortions: We will prove non-convexity by showing that any non-trivial X^d generates a discontinuity in $\hat{V}(X^d|\boldsymbol{\mu})$ at some μ^*.

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 $V(X^d|\boldsymbol{a},\boldsymbol{\mu}) \neq V(X|\boldsymbol{a},\boldsymbol{\mu})$, i.e., the perceived value from X^d differ from the true ones. 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 for a given action profile that for any non-trivial X^d, there exists a signal sensitive profile $\boldsymbol{a}$, i.e., not all actions in the action profile are identical, and a belief $\boldsymbol{\mu}^*$ that is extreme enough so that the constant action a^* is such that $E[u(a^*|\omega)|\boldsymbol{\mu}^*] = V(X^d|\boldsymbol{\mu}^*) = V(X^d|\boldsymbol{a},\boldsymbol{\mu}^*)$, while $V(X^d|\boldsymbol{a},\boldsymbol{\mu}^*) \neq V(X|\boldsymbol{a},\boldsymbol{\mu}^*)$. In other words, $\boldsymbol{\mu}^*$ is a belief at which the DM is just indifferent between the constant action a^* and some signal sensitive profile. Note that given the assumption that no single action is optimal for all states, such a belief $\boldsymbol{\mu}^*$ necessarily exists if $X \neq X^d$. As V is continuous in beliefs, and as X is not fully informative, for any $\epsilon > 0$, we can find $\boldsymbol{\mu}_\epsilon \neq \boldsymbol{\mu}^*$ 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 $\hat{V}(X^d|\boldsymbol{\mu}_\epsilon) = V(X^d|\boldsymbol{\mu}_\epsilon)$. 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 $\lim_{\epsilon\to 0} \hat{V}(X^d|\boldsymbol{\mu}'_\epsilon) \neq \lim_{\epsilon\to 0} \hat{V}(X^d|\boldsymbol{\mu}_\epsilon)$. There is a discontinuity at $\boldsymbol{\mu}^*$. Convexity in $\boldsymbol{\mu}$ (and hence $\boldsymbol{p}$) 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 $\alpha_1, ..., \alpha_k$ be weights such that $\sum_{i=1}^k \alpha_i \mu_\omega(\boldsymbol{s}_i) = \boldsymbol{\mu}$. These correspond to the probabilities with which each signal in S_X is observed from the perspective of the moderator. We can write:

$$\sum_i \alpha_i p_{\omega^*}(\boldsymbol{s}_i) \langle \boldsymbol{\mu}(\boldsymbol{s}_i), \boldsymbol{p} \circ \boldsymbol{\mu}^{-1} \rangle = \frac{p_{\omega^*}}{\mu_{\omega^*}} \sum_i \alpha_i \mu_{\omega^*}(\boldsymbol{s}_i). \tag{11}$$

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 $i \in \{1, ...k\}$. Substituting this into (11), 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. □

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.

C.2 Interaction between biases and distortions

We first show that if there exists a beneficial moderation policy that improves on a bias in prior, then there also exists a distortion that yields a strict improvement. In particular, a distortion that induces the same beliefs as the actual but moderated experiment would lead to the same choices after each perceived signal. Since both types of decision makers have the same view on X^d, this effectively 'creates' a sophisticated decision maker that responds to their belief about the experiment in a way the moderator would want them to. If the garbling is beneficial, so is this distortion.

Recall that $\hat{V}(X^d|\boldsymbol{p})$ is defined as the DM's (actual) expected utility given their choices, i.e., $\hat{V}(X^d|\boldsymbol{p}) = V(X|\hat{\boldsymbol{a}}, \boldsymbol{\mu})$, where $\hat{\boldsymbol{a}}$ the DM's choice given X^d and $\boldsymbol{p}$.

Result A.2. *Suppose a DM has a biased prior $\boldsymbol{p}$ and there exists a garbling M such that*

$$V(XM|\boldsymbol{a}^*, \boldsymbol{\mu}) > V(X|\boldsymbol{a}^*, \boldsymbol{\mu}), \tag{12}$$

where $\boldsymbol{a}^*$ is the action profile consistent with $V(X|\boldsymbol{p})$. Then for naive and sophisticated DMs, there generically exists a distortion d such that

$$\hat{V}(X|\boldsymbol{p}) < \hat{V}(X^d|\boldsymbol{p}). \tag{13}$$

Proof. Inequality (12) implies that there exists a beneficial moderation policy for a naive DM. It follows from Lemma 1 that there exist $\boldsymbol{s}_i, \boldsymbol{s}_j \in S_X$ such that $E[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)] > E[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)]$. If $V(X_\triangle(\boldsymbol{s}_i, \boldsymbol{s}_j)|\boldsymbol{p}) > \max\{E[u(a|\omega)|\boldsymbol{p}] : a \in \{a_i, a_j\}\}$, which given (12) is generically true for a finite action set, there also exists a beneficial moderation policy for a both naive and sophisticated DMs. Let M^* be the optimal moderation policy for a sophisticated DM that results in an action profile $\hat{\boldsymbol{a}}$. Clearly, a distortion $X^d = XM^*$ does not violate Bayes' consistency and achieves the same choices as M^*. Applying the same distortion to a naive DM equally results in $\hat{\boldsymbol{a}}$ and achieves the same outcome, since distortions affect both types equally. This yields (13). □

We proceed by showing that in the case of complete disagreement over two actions, *any* form of noise that affects the respective signals increases expected utility. This includes the case of symmetric garbling (i.e., white noise). In other words, even if the moderator values the information from an experiment, in the sense that the moderator would condition actions on the outcome, introducing white noise, as well as any other form of noise, can be beneficial. For a naive DM, this is also true for any magnitude of the noise, while for a sophisticated DM, this might only be the case up to the point where the DM adjusts their actions (which could lower expected utility from the perspective of the moderator).

This also implies that, if a DM suffers from distortions and/or a bias in prior, a second 'layer' of mistakes that adds noise to the perception or recollection of signals (as, for instance, in Rabin and Schrag (1999)), or the execution of choices, can be beneficial. For instance, in the absence of a moderator, a bias in prior that causes complete disagreement relative to the unbiased evaluation can be improved upon by a mistake that leads the DM to swap the corresponding signals. In fact, for the case of complete disagreement, any such (random) mistake acts as beneficial moderation. And a DM unaware of such errors might be better-off than someone taking the noise into account.

Result A.3. *Suppose given an experiment X, distortion d, and priors $\boldsymbol{p}$ and $\boldsymbol{\mu}$, the DM chooses an action profile $\boldsymbol{a}$ and there is complete disagreement over some a_i and a_j. Then for all $\boldsymbol{\beta}_i, \boldsymbol{\beta}_j \in (0,1]$, the moderation policy*

$$\begin{aligned} m(\boldsymbol{s}_i) &= \beta_i \boldsymbol{s}_i + (1-\beta_i)\boldsymbol{s}_j \\ m(\boldsymbol{s}_j) &= \beta_j \boldsymbol{s}_j + (1-\beta_j)\boldsymbol{s}_i \\ m(s_l) &= \boldsymbol{s}_l \quad \forall l \in \{1,...,k\};\ l \neq i, j, \end{aligned}$$

is such that $V(XM|\boldsymbol{a},\boldsymbol{\mu}) > V(X|\boldsymbol{a},\boldsymbol{\mu})$.

Proof. Complete disagreement implies 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 $E\big[u(a_i|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big] > E\big[u(a_j|\omega)|\boldsymbol{\mu}(\boldsymbol{s}_j)\big]$ (Lemma 3). The result follows directly. □