THEMA

théorie économique, modélisation et applications

THEMA Working Paper n°2022-13
CY Cergy Paris Université, France

The perils of a coherent narrative

Alessandro Ispano

July 2022

The perils of a coherent narrative

Alessandro Ispano

July 2022

Abstract

A persuader influences a decision-maker by providing a narrative to interpret some upcoming news. The decision-maker embraces a narrative when it is coherent (likelihoods of each news conditional on a realized state must sum to unity) and compatible with the truth (the marginal distribution of news is undistorted). Even if coherence restricts the set of beliefs the persuader can induce, it may nonetheless harm also the decision-maker. As a result, both players may benefit when the persuader can provide news-contingent, overall incoherent, narratives. Likewise, both players may benefit when the persuader can privately learn the truth, or when he can design the process of news arrival.

Keywords: interpretation, consistency, misspecification, manipulation, model, cognition
JEL classifications: D82, D83, D90

1 Introduction

A recent literature highlights how persuasion often occurs by provision of an interpretation for commonly available information (e.g. Eliaz and Spiegler (2020), Eliaz et al. (2021) and Schwartzstein and Sunderam (2021)) rather than by strategic revelation of private information (e.g. Milgrom (1981) and Crawford and Sobel (1982)) or by the design of the information that becomes available (e.g. Kamenica and Gentzkow (2011)). Examples range from public policy, e.g. claims of the government that health indicators warrant lock-down in a pandemic, to finance, e.g. an advisor hinting that stock-market returns are propitious to his client's investment, and research, e.g. an empirical study documenting a significant treatment effect in the data. This

Alessandro Ispano: CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France, alessandro.ispano@gmail.com. I am thankful to Frédéric Koessler, Alfonso Montes, Egor Starkov, Peter Vida as well as seminar participants at THEMA and Paris II for useful comments. Financial support from Labex MME-DII and ANR grant 19-CE26-0010-03 is gratefully acknowledged.

short paper focuses on the timing of provision and adoption of such interpretations and, in particular, on the welfare implications of the persuader having to propose an interpretation before knowing the actual information that will become available.

We consider a framework in which a persuader influences a decision-maker by providing a narrative to interpret some upcoming, uncertain, news. The decision-maker embraces the proposed narrative when it is coherent (conditional on a state, probabilities of each possible news sum to unity) and compatible with the true (the marginal distribution of news is undistorted). Once news realizes, the decision-maker updates the prior accordingly. A natural interpretation is that the decision-maker knows the frequency of states, as specified by the prior, and of news, as specified by the true news generating process. However, she does not know, or she is not sure about, the correlation that links the two variables, for instance because only the aggregate historical frequency of states and news is available. She is hence naively willing to accept any coherent interpretation the "expert" provides that is compatible with her knowledge (for a discussion of these behavioral assumptions, see the literature on causal misspecifications initiated by Spiegler (2016) and, in particular, Eliaz et al. (2021), who consider a setting in which the persuader has an intrinsic preference for maximizing perceived correlation). As we discuss in section 4, this framework also provides a benchmark relative to additional distortions in the decision-maker's perception of reality as well as to alternative assumptions on the adoption of interpretations and, in particular, the framework of Schwartzstein and Sunderam (2021), in which the proposal and adoption of a model occur only after news realizes and the decision-maker favors ex-post more plausible explanations.

We use a canonical example in which the persuader has state-independent preferences to highlight the following points. Naturally, coherence restricts the set of beliefs that the persuader can induce, since then not all news can be favorable, and hence always weakly harms the persuader. Yet, it may harm also the decision-maker. The intuition is that when unfavorable news is more likely, if possible, the persuader's interpretation turns it into favorable one. By coherence, it necessarily also turns favorable news into unfavorable one, while an incoherent narrative could continue to present it as favorable. Thus, both players would benefit if the persuader could provide news-contingent, overall incoherent, narratives. Besides, specifically due to coherence, both players may benefit when the persuader can privately learn the truth. This observation highlights an additional merit of coherence beyond limiting manipulation, namely, it can be effective at motivating the persuader to acquire information. Finally, both players may benefit when the persuader can design the process of news arrival. Indeed, with state-independent preferences, the only reason for the persuader to provide a misleading narrative is that the probability of each news is suboptimal from his point of view relative to the optimal experiment.

Taken together, these results highlight important trade-offs arising from an interested party's flexibility in the provision of narratives when the public does not fully take into account, and discount for, the strategic motives of their provider. Flexibility pertains to the timing of provision, i.e. before or after news realizes, and also to the nature of interpretable information, i.e. exogenously given or controlled by the interested party. These insights have implications for the organization of expertise and consulting, the interviewing of politicians, and the diffusion of research findings. For example, a biased expert required to provide guidance on public policy based on the evolution of the contagion curve in a pandemic may end up always recommending the wrong policy. Likewise, procedures restricting ex-post flexibility in experimental studies, such as mandatory preregistration of tests that will be conducted, may prevent a biased researcher to always present unidirectional findings but also induce him to propose very misleading interpretations.

2 General framework

The game takes place between a sender (S, he) and a receiver (R, she). Let $\omega \in \Omega$ be the true state of the world, where Ω is finite with $|\Omega| \geq 2$. Let μ_0 be prior distribution over Ω, which has full support. Let a be R's action and $U_R(a, \omega)$ and $U_S(a, \omega)$ denote R's and S's utility, respectively. This information is common knowledge of S and R. There is a signal π, characterized by a finite realization space X with $|X| \geq 2$, and by a family of distributions which specify for each state $\omega \in \Omega$ and realization $x \in X$ a likelihood $\pi(x|\omega)$, i.e. the probability of observing x when the state is ω. Knowing the true process governing signal realizations, S may induce R to adopt a different view about the underlying model at the initial stage.

Formally, a model m specifies for each state ω and realization $x \in X$ a likelihood $\pi_m(x|\omega) \in [0, 1]$. Let $\mathcal{M}$ be the space of all possible such models. When an expression is evaluated according to the true model, we drop the subscript m. We consider the following timing. Without knowing signal realization $x \in X$, S proposes a model $m \in \mathcal{M}$ and R chooses which model to adopt, as explained here below. R then observes realization x, updates according to the chosen model, takes action a and payoffs realize. We suppose R adopts any proposed model $m \in \mathcal{M}$ which is

(i) coherent, i.e. for each ω we have that $\sum_{x \in X} \pi_m(x|\omega) = 1$;

(ii) compatible, i.e. for each $x \in X$

$$\mathbb{P}_m(x) = \mathbb{P}(x), \tag{1}$$

where $\mathbb{P}_m(x)$ and $\mathbb{P}(x)$ denote the probability of observing realization x under model m and the true model, respectively. If S does not propose any model, or his proposition does not satisfy these properties, we assume R adopts the true model.

Coherence pertains to the internal consistency of a model. It requires the model to specify proper conditional distributions of news and is very natural given the ex-ante timing of S's model proposition. For example, when trying to determine the quality of a tennis player from the outcome of his next match, R would have a hard time believing that when the player is good his probability of winning exceeds 50% and so does his probability of losing. Compatibility instead pertains to the external consistency of a model, requiring it to respect the true marginal distribution of news. The assumption on R's model adoption when S fails to propose a model that satisfies these two properties is unimportant for the analysis and, in particular, we could equivalently assume R would stick to the prior, because of the following observation.

Observation 1. *S can always induce R to adopt the true model or a model according to which news are uninformative.*

Proof. The first part is obvious. For the second part, note S can propose a model according to which, for each $\omega \in \Omega$, $\pi_m(x|\omega)$ is independent from ω and, in particular, $\pi_m(x|\omega) = \mathbb{P}(x)$ for each $x \in X$. By construction, such a model is coherent and compatible. □

Conversely, coherence and compatibility imply S cannot systematically fool R. Technically, R's causal representation is (trivially) a perfect graph (see Spiegler (2020)).

Observation 2. *To be accepted by R, a model must necessarily be such that the expected induced posterior distribution is equal to the prior.*

Proof. Let $\mu_0(\omega)$ denote the prior probability that the state is ω, $\mu_m(\omega|x)$ the posterior probability induced by coherent and compatible model m upon realization x, and $\mathbb{E}(\mu(\omega)) \equiv \sum_{x \in X} \mathbb{P}(x)\, \mu_m(\omega|x)$ its expectation taken with respect to the true model. Then, for each $\omega \in \Omega$

$$\mathbb{E}(\mu(\omega)) = \sum_{x \in X} \mathbb{P}(x) \frac{\mu_0(\omega)\pi_m(x|\omega)}{\mathbb{P}_m(x)} = \sum_{x \in X} \mathbb{P}(x) \frac{\mu_0(\omega)\pi_m(x|\omega)}{\mathbb{P}(x)} = \sum_{x \in X} \pi_m(x|\omega)\mu_0(\omega) = \mu_0(\omega).$$

□

Finally, since a coherent model specifies a proper marginal distribution of news, there is necessarily a realization for which the compatibility requirement is not binding for S.

Observation 3. *If a coherent model m is such that $\mathbb{P}_m(x) = \mathbb{P}(x)$ for $|X|-1$ realizations, then it is compatible.*

Proof. Letting $\mu_0(\omega)$ denote the prior probability that the state is ω, the result follows from the fact that, for a coherent model m,

$$\sum_{x\in X} \mathbb{P}_m(x) = \sum_{x\in X}\sum_{\omega\in\Omega} \mu_0(\omega)\pi_m(x|\omega) = \sum_{\omega\in\Omega} \mu_0(\omega) \sum_{x\in X} \pi_m(x|\omega) = \sum_{\omega\in\Omega} \mu_0(\omega) = 1.$$

□

3 Leading example

We consider the following game based on the leading example from Kamenica and Gentzkow (2011), revisited by Schwartzstein and Sunderam (2021). The state of the world can be good ($\omega = G$) or bad ($\omega = B$) and R must choose whether to invest ($a = 1$) or not ($a = 0$). S aims to persuade R to invest no matter the state, i.e. his payoff is $U_S(a) = a$. Instead, R wants to invest only if the state is good, i.e. her payoff is $U_R(a, \omega) = \mathbb{1}_{(a=1\&\omega=G)\text{ or }(a=0\&\omega=B)}$, where $\mathbb{1}$ is the indicator function. We identify a distribution μ over states with the probability that the state is good, i.e. $\mu = \mathbb{P}(\omega = G)$, and $\mu_0 \in (0, 1)$ represents the prior probability. R will hence invest only if his belief μ is at least $1/2$. The true model π is a binary and symmetric signal, i.e. $X = \{b, g\}$, $\pi(b|B) = \pi(g|G) = \rho > 1/2$ and $\pi(g|B) = \pi(b|G) = 1 - \rho$. Thus, b and g represent respectively the "bad" and the "good" realization and ρ measures the signal precision.

3.1 The role of coherence in limiting manipulation

Suppose for a moment that coherence is not required and, for concreteness, that the prior is $\mu_0 = 3/10$ and the signal has precision $\rho = 2/3$. Upon the good realization g, R's posterior under the true model is

$$\mu = \frac{\mu_0\rho}{\mu_0\rho + (1-\mu_0)(1-\rho)} = 6/13 < 1/2,$$

where $\mu_0\rho + (1 - \mu_0)(1 - \rho) = \mathbb{P}(g) = 13/30$. R would therefore not invest. By condition (ii), for a proposed model m to be accepted by R it must be that

$$\mathbb{P}_m(g) = \pi_m(g|G)\mu_0 + \pi_m(g|B)(1 - \mu_0) = 13/30 = \mathbb{P}(g). \tag{2}$$

S induces the highest μ by proposing a model such that $\pi_m(g|G) = 1$ and $\pi_m(g|B)$ is the correspondent solution to equation (2), i.e. $\pi_m(g|B) = 4/21$, inducing $\mu = \mu_0/\mathbb{P}(g) = 9/13 > 1/2$. Thus, similarly to what noted by Schwartzstein and Sunderam (2021), S can induce R to invest upon good news by making it look better than it actually is.

But in fact, absent coherence, S can induce R to invest also upon bad news by turning it into good news. Upon the bad realization b, R's posterior under the true model is

$$\mu = \frac{\mu_0(1-\rho)}{\mu_0(1-\rho) + (1-\mu_0)\rho} = 3/17 < 1/2,$$

where $\mu_0(1-\rho)+(1-\mu_0)\rho = \mathbb{P}(b) = 17/30$. If S's proposed model is again such that $\pi_m(b|G) = 1$ and $\pi_m(b|B)$ is the correspondent solution to $\mathbb{P}_m(b) = \mu_0 + \pi_m(b|B)(1-\mu_0) = 17/30$, i.e. $\pi_m(b|B) = 8/21$, S induces $\mu = \mu_0/\mathbb{P}(b) = 9/17 > 1/2$.

Coherence limits the scope for manipulation since S's proposition above entails $\pi_m(g|G) + \pi_m(b|G) > 1$ and $\pi_m(g|B) + \pi_m(b|B) < 1$. More generally, as long as $\mu_0 < 1/2$, there exists no coherent model proposition m that induces R to invest upon each realization. Indeed, as seen at observation 2, coherence and plausibility imply that S's proposition m must always necessarily be such that the expected induced posterior is equal to μ_0.

3.2 Sender optimal proposition

If $\mu_0 \geq 1/2$, S can always induce R to stick to the prior and invest by observation 1. Suppose henceforth that $\mu_0 < 1/2$, so that S cannot induce R to invest upon both realizations. Since $\mathbb{P}(b) > \mathbb{P}(g)$, if possible, S would rather induce R to invest upon the bad realization. Using the results of Schwartzstein and Sunderam (2021), upon realization $x \in \{b, g\}$, a belief of at least 1/2 can be induced by (possibly incoherent) model m if and only if

$$1/2 \leq \frac{\mu_0}{\mathbb{P}(x)},$$

namely, when the prior μ_0 is sufficiently close to 1/2 or realization x is sufficiently unlikely. Conversely, if this condition is satisfied, by observation 3 there then exists a coherent model that induces such a belief upon x and is also compatible. As $\mathbb{P}(b)$ is increasing in ρ and decreasing in μ_0, S can induce R to invest upon the bad realization if and only if μ_0 is sufficiently large or ρ sufficiently small, i.e. if and only if $\rho \leq \bar{\rho}(\mu_0) \equiv \frac{\mu_0}{1-2\mu_0}$, where cutoff $\bar{\rho}$ is the solution to $1/2 = \mu_0/\mathbb{P}(b)$ with respect to ρ. Cutoff $\bar{\rho}$ is increasing in μ_0 and the condition is always satisfied when $\mu_0 \geq 1/3$ and always violated when $\mu_0 < 1/4$. If the condition is violated, S can only hope

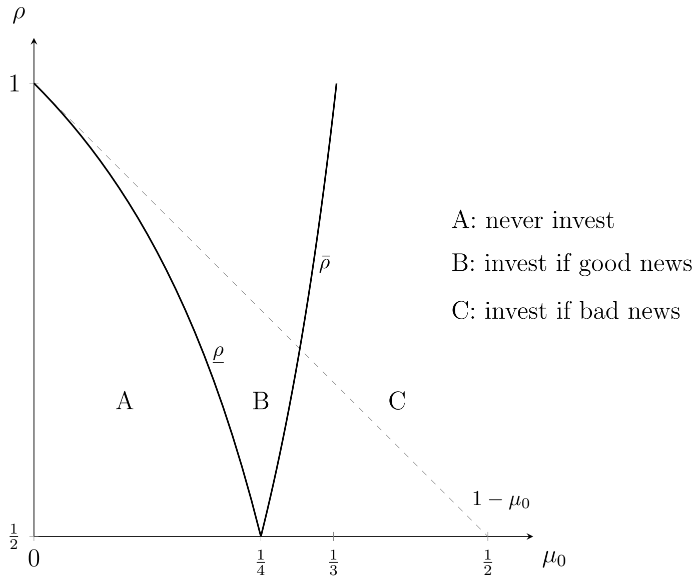


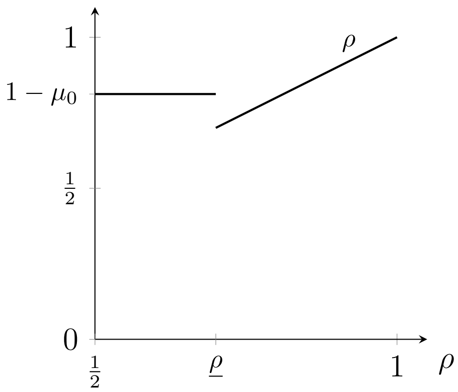
Figure 1 Receiver behavior under sender optimal proposition

to induce R to invest upon the favorable, less likely, realization. This is possible if and only if $\rho \geq \underline{\rho}(\mu_0) \equiv \frac{1-3\mu_0}{1-2\mu_0}$, where cutoff $\underline{\rho}$ is the solution to $1/2 = \mu_0/\mathbb{P}(g)$ with respect to ρ. Cutoff $\underline{\rho}$ is decreasing in μ_0 and the condition is always satisfied when $\mu_0 \geq 1/4$.

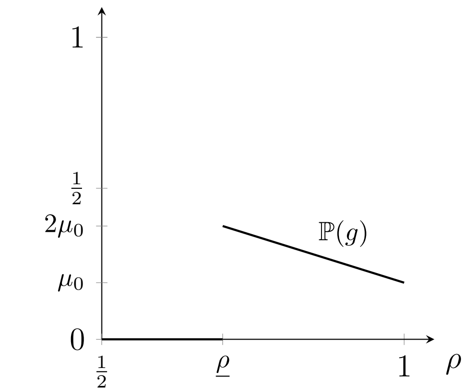
Figure 1 illustrates the three regions that partition the parameter space based on R's investment decision under S's optimal coherent model proposition. The dashed line, of equation $1-\mu_0$, represents the precision level of the signal above which R should invest upon good news under the true model, while investing upon bad news is obviously never warranted. Figure 2 represents the associated equilibrium payoff of R (left panel) and S (right panel) as a function of the precision of the signal ρ for three different values of the prior, namely, highly unfavorable ($\mu_0 < 1/4$, figures 2a and 2b), unfavorable ($\mu_0 \in (1/4, 1/3)$, figures 2c and 2d), and moderately unfavorable ($\mu_0 \in (1/3, 1/2)$, figures 2e and 2f).

3.3 Implications

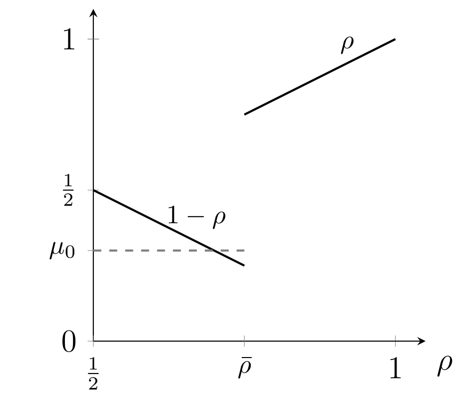
The perils of coherence Let us focus on the parameter space in which the outcome with and without the coherence requirement differ, i.e. region C in figure 1. Absent coherence, R would invest upon both realizations, while with coherence R only invests upon the bad realization. Thus, S's expected payoff is $\mathbb{P}(b)$, which is increasing in ρ since so is the probability of observing the bad realization. R's payoff is $1-\rho$, which is instead decreasing in ρ as she is taking the opposite action to the one suggested by the realization and the signal is getting more precise. As can be seen in figure 2, specifically due to coherence, R's and S's payoff may be respectively minimized and maximized when the signal is fully informative. Furthermore,



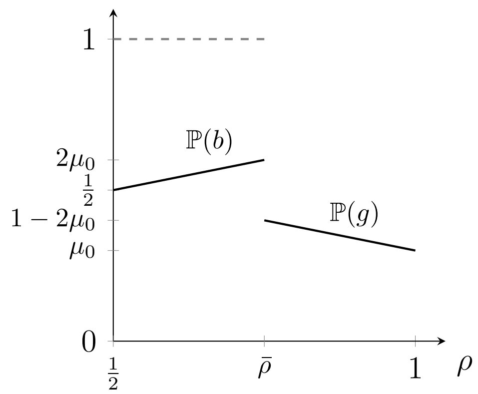
(a) Receiver payoff ($\mu_0 = 3/16$)



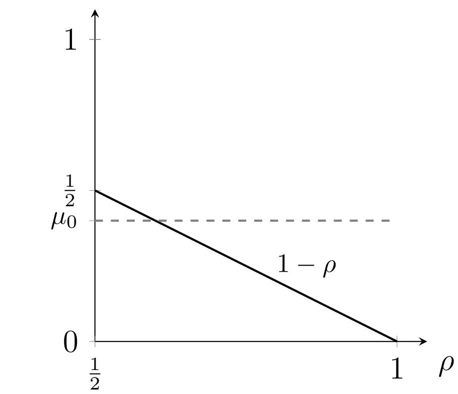
(b) Sender payoff ($\mu_0 = 3/16$)



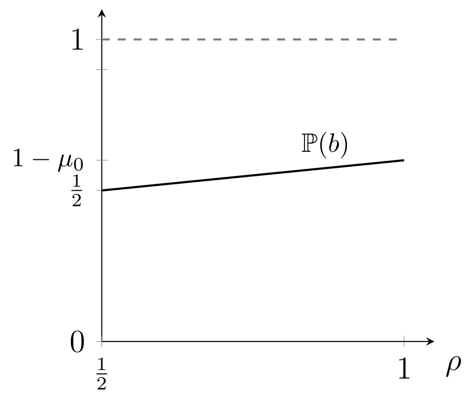
(c) Receiver payoff ($\mu_0 = 3/10$)



(d) Sender payoff ($\mu_0 = 3/10$)



(e) Receiver payoff ($\mu_0 = 4/10$)



(f) Sender payoff ($\mu_0 = 4/10$)

Figure 2 Payoffs as a function of signal precision under sender optimal proposition (dashed lines represent payoffs in the absence of coherence, whenever they differ) whenever $\rho > 1 - \mu_0$, i.e. the whole portion of the C region above the dashed line in figure 1, R would be better off if S's model need not be coherent, since in that case she would at least correctly invest upon a favorable realization. Conversely, coherence helps R when $\rho < 1 - \mu_0$, since then at least she does not wrongly invest upon the favorable realization.

Benefits from an informed sender Suppose that S can observe the state, but still not the signal realization, when proposing a model. Absent coherence, S's optimal model proposition is unaffected given that his payoff does not depend on the state. Instead, in the C region, S's optimal coherent model proposition would now induce R to invest upon the realization corresponding to the state he observed, i.e. the good realization when the state is good and the bad realization when the state is bad. Indeed, conditional on the realized state, $\pi(b|\omega) > \pi(g|\omega)$ if and only if $\omega = B$. As a result of the reduction in uncertainty about the upcoming realization, S's expected payoff would increase from $\mathbb{P}(b)$ to ρ. This scenario would clearly represent an improvement also for R, whose payoff would increase from $1-\rho$ to $\mu_0\rho+(1-\mu_0)(1-\rho) = \mathbb{P}(g)$. As $\mathbb{P}(g) > \mu_0$, R is now better off with coherence in the whole C region. Notice that if learning the state entailed a cost $k > 0$ for S, he would never acquire information absent coherence, while with coherence he has an incentive to do so as long as $k < \rho - \mathbb{P}(b)$.

Benefits from the sender designing the true model Absent coherence, it is clearly possible for S to sometimes do better under model persuasion than under the classical Bayesian persuasion problem (Kamenica and Gentzkow, 2011) in which S designs the true model and R rationally updates accordingly. This situation occurs when S induces R to always invest even though $\mu_0 < 1/2$. With coherence, this is not possible. Under Bayesian persuasion, the optimal experiment would induce R to invest with probability $2\mu_0$, in which case the induced belief would be exactly $\mu = 1/2$, and not to invest with complementary probability, in which case the induced belief would be exactly $\mu = 0$. S's payoff would hence be $2\mu_0$. For an appropriate choice of the true model, S can at best attain this payoff, i.e. when $\rho = \underline{\rho}$ for $\mu_0 < 1/4$ and when $\rho = \bar{\rho}$ for $\mu_0 \in [1/4, 1/3]$. Indeed, by construction the probability of the realization that induces R to invest, as well the associated induced beliefs, then coincide with the ones of the optimal experiment. Still, R's payoff is lower than his Bayesian persuasion one $(1 - \mu_0)$ for such values of ρ. In the region $\mu_0 < 1/4$, it is because R wrongly invests upon the good realization, which in reality is not sufficiently informative to warrant doing so. In the region $\mu_0 \in [1/4, 1/3]$, it is because R invests upon the bad realization. Finally, if $\mu_0 \in (1/3, 1/2)$, S cannot attain his Bayesian persuasion payoff since $P(b)$ converges to $1 - \mu_0$ for $\rho = 1$, which is less than $2\mu_0$ and $P(b)$ cannot further increase. Both players would be better off under Bayesian persuasion for any value of ρ. Thus, as S has state-independent preferences, his model proposition essentially amounts to a constrained Bayesian persuasion problem (realizations must have fixed probabilities) and he cannot benefit from the distortion in R's beliefs relative to the full-rationality benchmark with unrestricted information design.[1] Importantly, the model S would design would always be manipulation proof, i.e. S would have no incentive to deviate and propose an alternative model.

4 Discussion

In the framework we considered, the objective data generating process affects the decision-maker's model adoption decision only in that it determines the probabilities that the persuader's proposed interpretation must attach to each news. In some situations, it may entail additional restrictions. For example, if it is apparent that some news is more favorable than another (Milgrom, 1981), the decision-maker may only consider models which preserve this property. More fundamentally, the decision-maker may not be willing to consider models that imply an excessive distance between the true and the induced joint distribution over states and news. In the example considered, the outcome with and without coherence may then again be similar. Conversely, in other settings, the decision-maker may hold a distorted initial view of the distribution of news (and possibly also of states, i.e. a wrong prior). In this case, for an interpretation to be accepted, compatibility may be required to hold with respect to such distribution rather than the true one. Interestingly, since distorted beliefs are sometimes more immune to manipulation, the decision-maker can sometimes be better off. Matters become more complicated if the set of news the decision-maker deems possible do not coincide with the true one, which may require entering the realm of updating upon an unforeseen contingency (see Galperti (2019)).

Our assumptions on model adoption reflect a timing in which the two parties do not yet know the realized news when proposing and adopting an interpretation. In Schwartzstein and Sunderam (2021), instead, these decisions occur only ex-post once some particular news has

[1] This is not true in general. For instance, suppose that S's payoff modifies to $U_S(a,\omega) = \mathbb{1}_{(a=1\&\omega=B)\text{ or }(a=0\&\omega=G)}$, i.e. S aims to induce R to always take the "wrong" action. Consider again the C region, and in particular the area above the $1-\mu_0$ line. Since S is inducing R to take the wrong action according to the true posterior, S's optimal model proposition is unaffected, and so is R's payoff. However, S's payoff increases from $\mathbb{P}(b)$ to $\mathbb{P}(b)(1-\mu_0)+\mathbb{P}(g)(\mu_0)=\rho$. S's payoff is now higher than his Bayesian persuasion one $(1-\mu_0)$.

been observed. For the decision-maker to adopt proposed model m, they require[2]

$$\mathbb{P}_m(x) \geq \mathbb{P}(x) \tag{3}$$

to hold (with strict equality) for the particular realization x that players observe, formalizing the idea the people find plausible stories persuasive. As they note, requiring this plausibility condition alone to hold for each $x \in X$, so that equation (3) necessarily reduces to (1), does not further restrict the persuader. Thus, our framework allows to isolate the effect of the ex-ante timing and the associated coherence requirement, which the persuader could always trivially satisfy ex-post.

References

Crawford, Vincent P. and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), pp. 1431–1451.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

_ , _ , and Yair Weiss, "Cheating with models," *American Economic Review: Insights*, 2021, *3* (4), 417–34.

Galperti, Simone, "Persuasion: The art of changing worldviews," *American Economic Review*, 2019, *109* (3), 996–1031.

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

Milgrom, Paul R., "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, Autumn 1981, *12* (2), 380–391.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

Spiegler, Ran, "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

[2] To ease comparison, we are focusing on the case in which the decision-maker's default model, i.e. the model she initially thinks is true, is the true one, and she is maximally open to persuasion, i.e. the set of models she is willing to consider is $\mathcal{M}$.

– , "Can agents with causal misperceptions be systematically fooled?," *Journal of the European Economic Association*, 2020, *18* (2), 583–617.