Persuading a Motivated Thinker*

Victor Augias† Daniel M. A. Barreto‡

August 19, 2021

Abstract

We model a game of persuasion in which the receiver is a motivated thinker. Following the reception of Sender's signal, Receiver forms beliefs by trading-off the anticipatory utility any belief entails against the psychological cost of self-deluding, which results in overoptimism. We show that persuasion effectiveness depends on Receiver's material stakes: persuasion is more effective when it is aimed at encouraging behavior that is risky but can potentially yield very high returns and less effective when it is aimed at encouraging more cautious behavior. We illustrate this insight in economically relevant applications showing how financial advisors might take advantage of their clients overoptimistic beliefs and why informational interventions are often inefficient in inducing more investment in preventive health treatments. We extend the model to a binary majority voting setting in which voters hold heterogeneous partisan preferences. Optimal public persuasion induces maximum belief polarization in the electorate when voters' preferences are symmetric.

JEL classification codes: D82; D83; D91.

Keywords: Bayesian persuasion; information design; optimal beliefs; overoptimism; motivated thinking.

*This paper formerly circulated under the title "Wishful Thinking: Persuasion and Polarization." We thank Jeanne Hagenbach and Eduardo Perez-Richet for their support. We also thank S. Nageeb Ali, Michele Fioretti, Simon Gleyze, Emeric Henry, Deniz Kattwinkel, Frédéric Koessler, Laurent Mathevet, Meg Meyer, Daniel Monte, Doron Ravid, Nikhil Vellodi, Adrien Vigier and Yves Le Yaouanq for their valuable feedbacks and comments as well as seminar audiences at Sciences Po, Paris School of Economics and São Paulo School of Economics (FGV). All remaining errors are ours.

†Department of Economics, Sciences Po Paris, e-mail: victor.augias@sciencespo.fr. Victor Augias thanks the European Research Council (grant 850996 – MOREV) for financial support.

‡Department of Economics, Sciences Po Paris, e-mail: daniel.barreto@sciencespo.fr.

1 Introduction

In a variety of contexts, an agent (Sender) seeks to influence the behavior of a decision-maker (Receiver) by selectively disclosing decision-relevant information. The effectiveness of persuasion — the probability or frequency with which Sender induces her preferred behavior — depends on Receiver's prior knowledge of its environment, but also, crucially, on *how Receiver interprets the information that will be disclosed by Sender*.

It is generally assumed in strategic communication games that Receiver forms beliefs about the state of the world in a perfectly rational manner, as would a Bayesian statistician. Yet, a substantial literature in psychology and behavioral economics shows that the process by which individuals interpret information and form beliefs is not guided solely by a desire for accuracy but often depends on their motivations and material incentives, a phenomenon often termed *motivated inferences* (Kunda, 1987, 1990). In particular, an important motive that can influence information processing is *preferences*: in many situations, individuals systematically hold beliefs that are optimistically biased towards outcomes they wish to be true; a manifestation of motivated inferences commonly known as *wishful thinking* and characterized by the conjunction of *stakes-dependent* beliefs and *overoptimism* about preferred outcomes.[1] In this paper we investigate how a rational (Bayesian) persuader would optimally disclose information to a receiver forming such wishful beliefs.

Our model follows the typical structure of a persuasion game. An interested agent must decide what information to disclose to a decision-maker about an uncertain state of the world. Upon receiving this information, the decision-maker forms beliefs and acts accordingly. In particular, we assume that the agent controlling the information disclosure possesses commitment power in the same way as in Kamenica and Gentzkow (2011). This means the information policy is set ex ante while Sender is still unaware of the realization of the state of the world, and thus has no opportunity to deviate from her communication strategy at the interim stage. In contrast to Kamenica and Gentzkow (2011), however, instead of performing Bayes rule, Receiver forms motivated beliefs by optimally trading-off the anticipatory value from holding optimistic beliefs against the psychological costs of distorting subjective beliefs away from Bayesian posterior beliefs.

[1] There exists experimental evidence exhibiting such stakes-dependent overoptimistic beliefs. See in particular Bénabou and Tirole (2016), page 150 and Benjamin (2019) Section 9, as well as, e.g., Weinstein (1980), Mijović-Prelec and Prelec (2010), Mayraz (2011), Heger and Papageorge (2018), Coutts (2019), Engelmann et al. (2019) or Jiao (2020).

We characterize, in Section 3, the way a wishful Receiver optimally distorts beliefs from any Bayesian posterior induced by Sender's information policy. First, under mild assumptions on the cost of belief distortion, Receiver distorts away from the Bayesian posteriors if, and only if, the resulting action does not lead to a constant payoff across all possible states. This highlights an important way in which wishful thinking is tied to preferences: the marginal benefit of belief distortion is proportional to the variability of payoffs across states under the action the agent anticipates to take. Whenever an action leads to the same payoff regardless of the state, such benefit does not exist. Second, Receiver's belief distortion is biased towards the state that yields the highest possible payoff. In other words, Receiver is always overoptimistic about the most favorable outcomes.

Such distortions in his beliefs lead to a distortion in Receiver's behavior, when compared to a Bayesian agent. Formally, there might exist signals that induce a wishful and a Bayesian agent to take different actions, even if they depart from the same prior belief. In Section 4 we precisely characterize when this is the case and quantify how much wishful thinking impacts Receiver's behavior depending on Receiver's preferences and ability to self-deceive. We say that an action is *favored* by a wishful Receiver whenever it is taken at a strictly greater set of beliefs compared to the Bayesian case and show that two characteristics of the agent's preferences define which actions will be favored. The first is the *highest payoff* achievable under each action: wishful agents will be overoptimistic about the possibility of receiving such payoff, and as such might favor actions that can potentially yield the best outcome. The second aspect is the *payoff variability* of each action: the more the payoffs achievable under a given action vary across the possible states, the bigger the marginal benefit of distorting beliefs while playing that action and, as a result, the bigger the distortion. We show that whenever an action has both the highest possible payoff and the highest payoff variability among all actions, it will be favored. If an action induces the highest possible payoff and another action induces the highest payoff variability, then which of these two actions will be favored will depend on Receiver's ability to self-deceive.

A sender interested in inducing a particular action will therefore be more effective in persuading a wishful rather than a Bayesian receiver if such action is favored. As such, the effectiveness of information provision as a tool to incentivize agents might vary with individuals' material stakes: *persuasion is more effective when it is aimed at encouraging behavior that is risky but can potentially yield very high returns and less effective when it is aimed at encouraging more cautious behavior*. We illustrate this general insight in applications in which wishful beliefs can play an important role.

Application 1: Information Provision and Preventive Health Care. Individuals are consistently investing too little in preventive health care treatments, even if offered at low prices, especially in developing countries (Dupas, 2011; Chandra et al., 2019; Kremer et al., 2019, Section 3.1). Empirical studies provide mixed evidence that informational interventions might induce more individual investment in preventive health care devices (see, in particular, Dupas, 2011, Section 4, and Kremer et al., 2019, Section 3.3).

Recent literature conjectures that individuals might not be responsive to information campaigns because they prefer to hold optimistic prospects about their health risks (see Schwardmann, 2019 and Kremer et al., 2019, Section 3.3).[2] This argument can be formalized with our model: a health agency (Sender) designs an information campaign about the severity of an illness in order to promote a preventive treatment that can be adopted by individuals at some cost. Since not adopting the preventive treatment is the action that can potentially yield the highest payoff (in case the illness is not severe) and also the action with the highest payoff variability, it will be favored by wishful receivers. As such, information campaigns aimed at promoting preventive behavior are less effective.

Application 2: Public Persuasion and Political Polarization. Belief polarization along partisan lines is a pervasive and much debated feature of contemporary societies. Although such polarization can be partly caused by differential access to information, evidence suggests that it is exacerbated by the fact that individuals tend to make different motivated inferences about the *same* piece of information (Babad, 1995; Thaler, 2020).

In this application we explore the relationship between optimal information disclosure to motivated thinkers and polarization. We model a majority voting setting in which an electorate, differentiated in terms of partisan preferences, uses information disclosed by a politician to vote on a proposal. Motivated thinking leads voters with different preferences to adopt different beliefs after being exposed to a public signal, giving rise to disagreement. A cleavage in beliefs arises, separating the electorate in two groups: those voting against or for the proposal distort their beliefs in opposite directions, exacerbating polarization. Sender's optimal public experiment consists in persuading the median voter, maximizing the number of voters distorting beliefs in opposite directions. We show that if partisan preferences are symmetrically distributed around the median, then Sender's optimal information policy generates maximal belief polarization in the electorate.

[2]There exists compelling experimental evidence that such self-deception exists in the medical testing context (Lerman et al., 1998; Oster et al., 2013; Ganguly and Tasoff, 2017).

Application 3: Persuading a Wishful Investor. Individual investors usually need guidance and relevant information to take financial decisions. Expert advice is thus an essential feature of many retail financial services markets (Inderst and Ottaviani, 2012). Nevertheless, evidence shows that, because of commission-based compensations[3], professional advisors might sometimes not act in the best interest of their clients by making investment recommendations that take advantage of the clients' biases and mistaken beliefs (see, for instance, Mullainathan et al., 2012 or Beshears et al., 2018, Section 9). We show in this application that a financial broker interested in selling a risky product is always more effective when persuading a wishful investor, illustrating why some financial consulting firms seem to specialize in advice misconduct and cater to biased consumers (Egan et al., 2019), as well as why the online betting industry exerts so much persuasion effort. Indeed, Babad and Katz (1991) document that individuals generally display wishful thinking when they take part in lotteries: they prefer to think they will win and are therefore more receptive to information encouraging risky bets.

Related literature. The persuasion and information design literature[4] has initially focused on the problem of influencing rational Bayesian decision-makers as in the seminal contributions of Kamenica and Gentzkow (2011) and Bergemann and Morris (2016). By introducing non-Bayesian updating in the form of motivated beliefs formation, we contribute to the literature studying persuasion of receivers subject to mistakes in probabilistic inferences.[5,6] Levy et al. (2018) analyze a Bayesian persuasion problem where a sender can send multiple signals to a receiver subject to correlation neglect. Benjamin et al. (2019) provide an example of persuasion game where Receiver exhibits base-rate neglect when updating beliefs. In de Clippel and Zhang (2020) the receiver holds subjective beliefs which may be any arbitrary distortion of the Bayesian posterior.

More broadly, our paper belongs to the literature examining receiver-side "frictions" in

[3]Consumers often pay indirectly for financial advice through commissions that are channeled by product providers to brokers, investment advisers, and other intermediaries (Inderst and Ottaviani, 2012).

[4]See Bergemann and Morris (2019) and Kamenica (2019) for reviews of this literature.

[5]See Benjamin (2019) for a review of the literature. In particular, wishful thinking belongs to preference-biased inferences reviewed in Benjamin (2019), Section 9.

[6]It is interesting to note that an active literature also explores how errors in strategic reasoning (Eyster, 2019) affect equilibrium outcomes in strategic communication games. Although in our model Receiver understands all the strategic issues, we believe, nevertheless, that it is important to mention that players' misunderstanding of their strategic environment might also lead them to make errors in statistical inference even if they update beliefs via Bayes rule, as in Mullainathan et al. (2008), Ettinger and Jehiel (2010), Hagenbach and Koessler (2020) and Eliaz et al. (2021a,b) who consider communication games where players make inferential errors because of some "coarse" understanding of their environment.

Bayesian persuasion. In Ely et al. (2015), Lipnowski and Mathevet (2018) and Schweizer and Szech (2018) the receiver values information for non-instrumental reasons which impacts how much information should be disclosed by Sender. In Bloedel and Segal (2018), Lipnowski et al. (2020), Che et al. (2020) and Wei (2021) the receiver bears a cost from maintaining attention which might lead to information overload if Sender discloses too much. In static and dynamic Bayesian persuasion settings, some papers study the effect of exogenous (Bizzotto et al., 2018; Bizzotto and Vigier, 2020) as well as endogenous (Matysková, 2018; Montes, 2019) information acquisition by Receiver additionally to Sender's information. In Galperti (2019) Receiver holds a misspecified prior assigning probability zero on the true state of the world. Sender thus has to carefully design the information structure such as to trigger Receiver to change his initial "worldview."

Our model of motivated beliefs formation[7] follows the wishful thinking model of Caplin and Leahy (2019).[8] The decision-maker forms beliefs by trading-off benefits and costs of maintaining overly optimistic beliefs. Modelling the belief formation process as an actual optimization problem over the set of beliefs dates back from the seminal paper of Brunnermeier and Parker (2005) who, differently from Caplin and Leahy (2019), assume that the decision-maker chooses beliefs by trading-off the anticipated utility against the material cost those beliefs entail. Another closely related paper is Bracha and Brown (2012), who model the optimal choice of beliefs and actions of a single decision-maker as the result of a simultaneous game between two selves, with the self choosing beliefs holding a similar objective function as the agent in Caplin and Leahy (2019). The optimal belief approach to motivated belief updating has been used in different settings to model the effects of endogenous overoptimism on various economic outcomes (see, e.g., Brunnermeier et al., 2007; Gollier and Muermann, 2010; Oster et al., 2013; Brunnermeier et al., 2017; Jouini and Napp, 2018; Schwardmann, 2019; Bridet and Schwardmann, 2020). It is important to note that while anticipatory utility may be a strong motive for manipulating one's beliefs, it is not the only possible one. This differentiates wishful thinking from the more general concept of motivated reasoning, which is usually defined as the degree to which individuals' cognition is affected by their motivations.[9] Different motivations from anticipated payoffs have been explored in the literature such as cognitive

[7] See Bénabou and Tirole (2016) for a review of the literature.

[8] It is worth mentioning that recent papers by Mayraz (2019) and Kovach (2020) introduce behavioral foundations for wishful thinking. These works axiomatize the kind of belief distortion Receiver holds in our model as a result of his optimal choice of beliefs.

[9] See Krizan and Windschitl (2009) for a more detailed discussion on the differences between wishful thinking and motivated reasoning.

dissonance avoidance (Akerlof and Dickens, 1982; Golman et al., 2016), preference to believe in a "Just World" (Bénabou and Tirole, 2006), maintaining high motivation when individuals are aware of being subject to a form of time-inconsistency (Bénabou and Tirole, 2002, 2004) or satisfying the need to belong to a particular identity (Bénabou and Tirole, 2011).

2 Model

States and prior beliefs. A state of the world θ is drawn by Nature from a state space Θ according to a prior distribution $\mu_0 \in \text{int}(\Delta\Theta)$.[10] Receiver (he) and Sender (she) do not observe the state ex-ante but its prior distribution is common knowledge. In our model, Receiver's and Sender's subjective beliefs generally differ. We denote by η (resp. μ) Receiver's (resp. Sender's) beliefs.

Actions and payoffs. Receiver has to chose an action a from a compact action space A with at least two actions and has utility function $u\colon A \times \Theta \to \mathbb{R}$. Receiver's choice affects Sender's utility $v\colon A \to \mathbb{R}$.[11] Before Receiver takes his action Sender can commit, without restriction, to any signal structure $\sigma\colon \Theta \to \Delta S$ where S is an endogenously chosen set of signal realizations.

Receiver's optimal behavior. Given subjective belief η Receiver's optimal action correspondance is

$$A(\eta) = \arg\max_{a \in A} \int_\Theta u(a,\theta)\eta(\mathrm{d}\theta).$$

Without loss of generality, we assume that no action is trivially dominated, i.e. for any action $a \in A$ there always exists some belief η such that $a \in A(\eta)$. When the set $A(\eta)$ has more than one element we break the tie in favor of Sender. That is, when Receiver holds belief η, the action played in equilibrium is given by a selection $a(\eta) \in A(\eta)$ which maximizes Sender's expected utility.[12]

[10] In what follows, all spaces Ω are assumed to be nonempty Polish spaces endowed with the Borel σ-algebra $\mathcal{B}(\Omega)$. The set $\Delta\Omega$ denotes the set of Borel probability measures over the measure space $(\Omega, \mathcal{B}(\Omega))$. We always endow $\Delta\Omega$ with the weak* topology. If the support of a measure $\mu \in \Delta\Omega$ is finite we adopt the shorthand notation $\mu(\{\omega\}) = \mu(\omega)$ for any $\omega \in \text{supp}(\mu)$.

[11] The map $u(a,\cdot)\colon \Theta \to \mathbb{R}$ is assumed to be continuous, bounded and Borel-measurable for any $a \in A$.

[12] There might be more than one such selection if there exists some $\eta \in \Delta\Theta$ at which Sender is indifferent between some actions in $A(\eta)$. In that case, we pick arbitrarily one of those actions.

Posterior beliefs. After observing any signal realization $s \in S$, Sender updates her prior belief by performing Bayes rule. Her belief after observing signal realization s is then

$$\mu(\tilde{\Theta}|s) = \frac{\int_{\tilde{\Theta}} \sigma(s|\theta)\mu_0(\mathrm{d}\theta)}{\int_{\Theta} \sigma(s|\theta)\mu_0(\mathrm{d}\theta)},$$

for any Borel set $\tilde{\Theta} \subseteq \Theta$. In constrast to Sender, Receiver is not Bayesian. Instead, when forming beliefs, Receiver trades-off his anticipated payoff against the cost of holding possibly non-Bayesian beliefs. Let us define Receiver's psychological well-being function as

$$W(\eta, \mu) = \int_{\Theta} u(a(\eta), \theta)\eta(\mathrm{d}\theta) - \frac{1}{\rho} C(\eta, \mu)$$

for any $\eta, \mu \in \Delta\Theta$, where ρ is a strictly positive weight parameter describing Receiver's ability to self-deceive and $C\colon \Delta\Theta \times \Delta\Theta \to \mathbb{R}_+$ is a belief distortion cost function. In the main text, we make the assumption that the cost function corresponds to the Kullback-Leibler divergence between Receiver's subjective belief and Sender's Bayesian posterior, denoted $\mathrm{D}_{\mathrm{KL}}(\eta\|\mu)$ and defined by:

$$\mathrm{D}_{\mathrm{KL}}(\eta\|\mu) = \int_{\Theta} \frac{\mathrm{d}\eta}{\mathrm{d}\mu}(\theta) \ln\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}(\theta)\right)\mu(\mathrm{d}\theta),$$

for any $\eta, \mu \in \Delta\Theta$, where $\frac{\mathrm{d}\eta}{\mathrm{d}\mu}$ is the Radon-Nikodym derivative of η with respect to μ, defined whenever η is absolutely continuous with respect to μ. This assumption is made for tractability but does not qualitatively affect the main results. Precisely, we show that our results on Receiver's optimal beliefs continue to hold when the psychological cost functions belongs to a more general class of statistical divergences in Appendix A.1.

Receiver's optimal belief solves

$$W(\mu) = \max_{\eta \in \Delta\Theta} W(\eta, \mu),$$

for any Bayesian belief $\mu \in \Delta\Theta$.[13] Accordingly, Receiver's optimal belief correspondence

[13] As already noted by Bracha and Brown (2012) as well as Caplin and Leahy (2019), this optimization problem has a similar mathematical structure to the multiplier preferences developed in Hansen and Sargent (2008) and axiomatized in Strzalecki (2011). Precisely, the agent in Strzalecki (2011) solves

$$\max_{a \in A} \min_{\eta \in \Delta\Theta} \int_{\Theta} u(a, \theta)\eta(\mathrm{d}\theta) + \frac{1}{\rho}\mathrm{D}_{\mathrm{KL}}(\eta\|\mu), \tag{1}$$

is given by

$$\eta(\mu) = \arg\max_{\eta \in \Delta\Theta} W(\eta, \mu),$$

for any $\mu \in \Delta\Theta$. Given that the Bayesian belief is μ, the psychological well-being of Receiver in the interim phase is given by $W(\mu)$. We assume that when Receiver is indifferent, in terms of psychological well-being, between several subjective posteriors at some Bayesian belief μ he picks the one that maximizes Sender's expected utility. This tie breaking rule ensures that the optimal belief correspondence is singleton-valued and simplifies the characterization of the optimal information policy.

Information design problem. It is known since Kamenica and Gentzkow (2011) that it is equivalent to think in terms of signal structure or in terms of Bayes-plausible distributions over posteriors. Hence, Sender commits, at the ex-ante stage, to an information policy $\tau \in \mathcal{T}(\mu_0)$, where

$$\mathcal{T}(\mu_0) = \left\{ \tau \in \Delta\Delta\Theta : \int_{\Delta\Theta} \mu(\tilde{\Theta}) \tau(\mathrm{d}\mu) = \mu_0(\tilde{\Theta}) \text{ for any Borel set } \tilde{\Theta} \subseteq \Theta \right\},$$

is the set of Bayes-plausible distribution over posterior beliefs given the prior μ_0.[14]

We assume *Sender knows Receiver's tendency to be a wishful thinker*, i.e., anticipates correctly Receiver's subjective belief $\eta(\mu)$ for any $\mu \in \Delta\Theta$. Since Receiver's optimal belief characterizes how he would distort his belief away from any realized Bayesian posterior, Sender can choose the best information policy by backward induction, knowing: (i) how Receiver will distort his posterior; (ii) how he will behave given the realized distorted posterior. For any Bayesian posterior $\mu \in \Delta\Theta$, the set of Receiver's optimal actions is

$$A(\eta(\mu)) = \arg\max_{a \in A} \int_{\Theta} u(a, \theta) \eta(\mu)(\mathrm{d}\theta),$$

for any given $\mu \in \Delta\Theta$. In that model, the parameter ρ measures the degree of confidence of the decision-maker in the belief μ or, in other words, the importance he attaches to belief misspecification. Conclusions on the belief distortion in that setting are naturally reversed with respect to our model: a receiver forming beliefs according to Equation (1) would form overcautious beliefs. Studying how a rational Sender would persuade a Receiver concerned by robustness seems an interesting path for future research.

[14] By Theorem 15.17 in Aliprantis and Border (2006), $\Delta\Theta$ is itself a Polish space that we equip with the weak* topology and the corresponding Borel σ-algebra.

and Sender's indirect utility function is therefore

$$v(\mu) = \max_{a \in A(\eta(\mu))} \int_{\Theta} v(a, \theta) \mu(\mathrm{d}\theta),$$

for any $\mu \in \Delta\Theta$. Hence, Sender's value from persuading a Receiver with motivated beliefs under the prior μ_0 is

$$V(\mu_0) = \max_{\tau \in \mathcal{T}(\mu_0)} \int_{\Delta\Theta} v(\mu) \tau(\mathrm{d}\mu). \tag{2}$$

Discussion of the modelling assumptions. Receiver's belief formation process optimally trades-off the benefits and costs associated with maintaining non-Bayesian beliefs. We make assumptions about both benefits and costs.

On the one hand, we assume that the benefit for Receiver from maintaining inaccurate beliefs comes from the expected utility he anticipates from his decision problem. Intuitively, it represents the idea that individuals might derive utility from the anticipation of future outcomes, be them good or bad. This hypothesis has been widely used in the literature to study how anticipatory emotions affect physical choices (see, e.g., Loewenstein, 1987; Caplin and Leahy, 2001) as well as choices of beliefs (Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005; Caplin and Leahy, 2019). Receiver's choice of beliefs is thus a way of satisfying his psychological need to be optimistic about the best-case outcomes or, on the contrary, to avoid the dread and anxiety associated with the worst-case outcomes. This hypothesis is supported experimentally by Engelmann et al. (2019), who find significant evidence that wishful thinking is caused by the desire to reduce anxiety associated with anticipating bad events.

On the other hand, we assume distorting beliefs away from the Bayesian benchmark is subject to some psychological cost. This assumption reflects the idea that, under a motivated cognition process (Kunda, 1987, 1990), individuals may use sophisticated mental strategies such as manipulating their own memory (Bénabou, 2015; Bénabou and Tirole, 2016)[15], avoiding freely available information (Golman et al., 2017) or creating elaborate narratives supporting their bad choices or inaccurate claims to justify their preferred beliefs.[16] Our assumptions on the cost function captures, in "reduced form", the fact that

[15] For experimental evidence on memory manipulation see, e.g., Saucet and Villeval (2019), Carlson et al. (2020) and Chew et al. (2020).

[16] One can relate this possible microfoundation of the belief distortion cost to the literature on lying costs (Abeler et al., 2014, 2019) since, when Receiver is distorting away his subjective belief from the rational Bayesian beliefs, he is essentially lying to himself. We thank Emeric Henry for suggesting us this interpretation of the cost function.

implementing such mental strategies comes at a cost when desired beliefs deviate from from the Bayesian rational ones. In contrast, Brunnermeier and Parker (2005) model the cost of erroneous beliefs as the instrumental loss associated with the inaccurate choices induced by such beliefs. It is worth noting that Coutts (2019) provides experimental evidence in favor of the psychological rather than instrumental costs associated with belief distortion.

3 Receiver's motivated beliefs and behavior

This section completes and extends the results in Caplin and Leahy (2019). Precisely, we characterize optimal beliefs and behavior of a wishful decision-maker without any restriction on the action and state spaces. We start by fixing action a and define $\eta_a(\mu)$ to be the optimal belief of Receiver with no other choice but a. Formally,

$$\eta_a(\mu) = \arg\max_{\eta \in \Delta\Theta} \int_\Theta u(a,\theta)\eta(\mathrm{d}\theta) - \frac{1}{\rho} \mathrm{D}_{\mathrm{KL}}(\eta \| \mu).$$

Accordingly, Receiver's optimal well-being at belief $\eta_a(\mu)$ is

$$W_a(\mu) = \max_{\eta \in \Delta\Theta} \int_\Theta u(a,\theta)\eta(\mathrm{d}\theta) - \frac{1}{\rho} \mathrm{D}_{\mathrm{KL}}(\eta \| \mu),$$

We call $\eta_a(\mu)$ Receiver's optimal belief motivated by action a under posterior μ. We can therefore identify Receiver's optimal beliefs by: (i) finding the optimal belief motivated by action a under μ, resulting in psychological well-being $W_a(\mu)$; (ii) finding which action it is optimal to motivate by maximizing $W_a(\mu)$ with respect to a. The next proposition characterizes Receiver's optimal beliefs.

Proposition 1. *Receiver's optimal belief motivated by action a under posterior μ is given by*

$$\eta_a(\mu)(\tilde{\Theta}) = \frac{\int_{\tilde{\Theta}} \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta)}{\int_\Theta \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta)},$$

for any Borel set $\tilde{\Theta} \subseteq \Theta$, any $a \in A$ and any $\mu \in \Delta\Theta$, while Receiver's psychological well-being when under belief $\eta_a(\mu)$ equals

$$W_a(\mu) = \frac{1}{\rho} \ln\left(\int_\Theta \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta) \right),$$

for any $a \in A$ and any $\mu \in \Delta\Theta$. Hence, Receiver's optimal belief is

$$\eta(\mu)(\tilde{\Theta}) = \frac{\int_{\tilde{\Theta}} \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta)}{\int_{\Theta} \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta)}, \tag{3}$$

for any Borel set $\tilde{\Theta} \subseteq \Theta$, whenever $\mu \in \Delta\Theta$ satisfies

$$\int_{\Theta} \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta) \geq \int_{\Theta} \exp(\rho u(a',\theta))\mu(\mathrm{d}\theta),$$

for all $a' \neq a$.

Proof. See Appendix A.1. □

The former proposition states that Receiver's optimal belief is discontinuous in μ, with discontinuities located at Bayesian posterior beliefs such that Receiver is indifferent, in terms of psychological well-being, between holding beliefs η_a or $\eta_{a'}$ for some $a \neq a'$.

We can see from expression (3) that Receiver holds a distorted subjective belief if and only if the resulting action $a \in A$ does not lead to a sure payoff. Formally, for any $a \in A$, we have $\eta_a(\mu) \neq \mu$ if, and only if, there exists $\theta \neq \theta'$ such that $u(a,\theta) \neq u(a,\theta')$. Hence, *only preferences constitute a motive for Receiver to distort his beliefs*. Furthermore, Receiver distorts upwards the probabilities of states associated with high payoffs and vice-versa. A wishful Receiver is therefore always *overoptimistic about his preferred outcomes*.

As Proposition 1 shows, wishful thinking naturally has an effect on the beliefs of Receiver. The next result shows that, accordingly, wishful thinking has behavioral implications for Receiver.

Corollary 1. *Under optimal wishful beliefs, Receiver's optimal action correspondence satisfies*

$$A(\eta(\mu)) = \arg\max_{a \in A} \int_{\Theta} \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta),$$

for any $\mu \in \Delta\Theta$.

Remark that this result comes as a direct consequence of Proposition 1 as, by definition, any action a is optimal under the belief motivated by action a. As already observed by Caplin and Leahy (2019), the previous result states, in essence, that a Receiver forming wishful beliefs behaves as a Bayesian agent whose preferences are distorted by the function $z \mapsto \exp(\rho z)$ for any $z \in \mathbb{R}$. Importantly, from Sender's point of view, a wishful

Receiver's behavior is therefore indistinguishable from that of a Bayesian rational agent with utility function $\exp(\rho u(a,\theta))$.

Corollary 1 shows that wishful thinking materializes in the form of "motivated errors" in the sense of Exley and Kessler (2019): by choosing psychologically desirable beliefs, Receiver commits systematic errors in his decision-making, i.e., acts as if he had cognitive limitations or behavioral biases relatively to a Bayesian decision-maker.

4 Persuading a wishful Receiver

In this section, we compare Sender's value of persuading a motivated compared to a Bayesian Receiver. In what follows, we constrain the action space to be binary, $A = \{0,1\}$, and consider successively the cases of finite and continuous state spaces. Moreover, we assume that Sender has utility function $v(a) = a$ for any $a \in A$. These restrictive assumptions allow us to gain in analytical tractability.

4.1 Finite state space

We start by assuming that $\Theta = \{\underline{\theta}, \overline{\theta}\}$. We will first compare the value of persuasion when the state space is binary and then show that the result extends to any finite state space with more than two elements.

Denote $u(a,\underline{\theta}) = \underline{u}_a$ and $u(a,\overline{\theta}) = \overline{u}_a$ for any $(a,\theta) \in A \times \Theta$. Assume that Receiver wants to "match the state," such that $\overline{u}_1, \underline{u}_0 > \overline{u}_0, \underline{u}_1$. Define the *payoff variability under action* 0 by $u_0 = \underline{u}_0 - \overline{u}_1$, the *payoff variability under action* 0 by $u_1 = \overline{u}_1 - \underline{u}_1$ and *the highest achievable payoff* by $u_{\max} = \underline{u}_0 - \overline{u}_1$. With a small abuse of notation, denote $\eta = \eta(\overline{\theta})$ and $\mu = \mu(\overline{\theta})$.

By Corollary 1, comparing the value of persuasion for Sender when facing a wishful rather than Bayesian receiver is equivalent to compare the value of persuasion when facing two Bayesian receivers with respective utility functions $\exp(\rho u(a,\theta))$ and $u(a,\theta)$. Denote μ^B (resp. $\mu^W(\rho)$) the belief at which a Receiver with preferences $u(a,\theta)$ (resp. $\exp(\rho u(a,\theta))$) is indifferent between the two actions. Those beliefs are respectively equal to

$$\mu^B = \frac{\underline{u}_0 - \underline{u}_1}{\underline{u}_0 - \underline{u}_1 + \overline{u}_1 - \overline{u}_0}$$

and

$$\mu^W(\rho) = \frac{\exp(\rho\underline{u}_0) - \exp(\rho\underline{u}_1)}{\exp(\rho\underline{u}_0) - \exp(\rho\underline{u}_1) + \exp(\rho\overline{u}_1) - \exp(\rho\overline{u}_0)}.$$

We say that a wishful Receiver favors action $a = 1$ if $\mu^W < \mu^B$. Whenever that condition is satisfied, a wishful Receiver takes action $a = 1$ under a larger set of beliefs than a Bayesian. Next proposition characterizes whenever it is the case.

Lemma 1. *Action $a = 1$ is favored by a wishful Receiver if, and only if:*

(i) $u_{\max} \leq 0$ *and* $u_0 < u_1$, *or;*

(ii) $u_{\max} < 0$, $u_0 > u_1$ *and* $\rho > \overline{\rho}$, *or;*

(iii) $u_{\max} > 0$, $u_0 < u_1$ *and* $\rho < \overline{\rho}$.

where $\overline{\rho}$ is a strictly positive threshold such that

$$\mu^W(\overline{\rho}) = \mu^B.$$

Proof. See Appendix A.2 □

Two key aspects of the payoff matrix thus determine which action is favored by a wishful Receiver: *the highest achievable payoff* as well as *the payoff variability* for both actions. It is easy to grasp the importance of the highest payoff. Since the wishful thinker will always distort his beliefs in the direction of more favorable outcomes, in the limit, when there is no cost of distorting the Bayesian belief, Receiver would fully delude herself and always play the action that potentially yields such a payoff. The payoff variability u_a, on the other hand, is precisely Receiver's marginal benefit of distorting his belief under action a. Hence, the higher the payoff variability associated with action a, the more the uncertainty about θ is relevant when such action is played and the bigger the marginal gain in anticipatory utility the wishful thinker would get from distorting beliefs.

Lemma 1 states that if an action a has both the highest payoff $\underline{u}_0$ or $\overline{u}_1$ and the greatest payoff variability u_a among all actions $a \in A$, it will always be favored. If an action has either the highest payoff or the greatest payoff variability, then the self-deception ability ρ will define whether or not it will be favored: for low self-deception ability the action with the highest payoff will be favored, whereas for high self-deception ability it is the action with the greatest payoff variability that will be favored. The reason is the following. For sufficiently high values of Receiver's self-deception ability, Receiver can afford stronger overoptimism about the most desired outcome despite that this action is not associated with the highest marginal psychological benefit, and distorts behavior accordingly. In contrast, for sufficiently low values of ρ, Receiver cannot afford too much overoptimism

about the most desired outcome. Hence, he prefers to distort beliefs at the margin that yields the highest marginal psychological benefit, such that the action associated with the highest payoff variability is favored.

Let us now turn our attention to the following questions: when is Sender better-off facing a wishful Receiver compared to a Bayesian and how does the (Blackwell) informativeness of Sender's optimal policy compares when persuading a wishful or a Bayesian Receiver? Remember that Sender chooses an information policy $\tau \in \Delta[0,1]$ maximizing

$$\int_{[0,1]} v(\mu)\tau(\mathrm{d}\mu),$$

where

$$v(\mu) = \begin{cases} 1 & \text{if} \quad \mu \geq \mu^W \\ 0 & \text{if} \quad \mu < \mu^W \end{cases},$$

for every $\mu \in [0,1]$, subject to the Bayes plausibility constraint

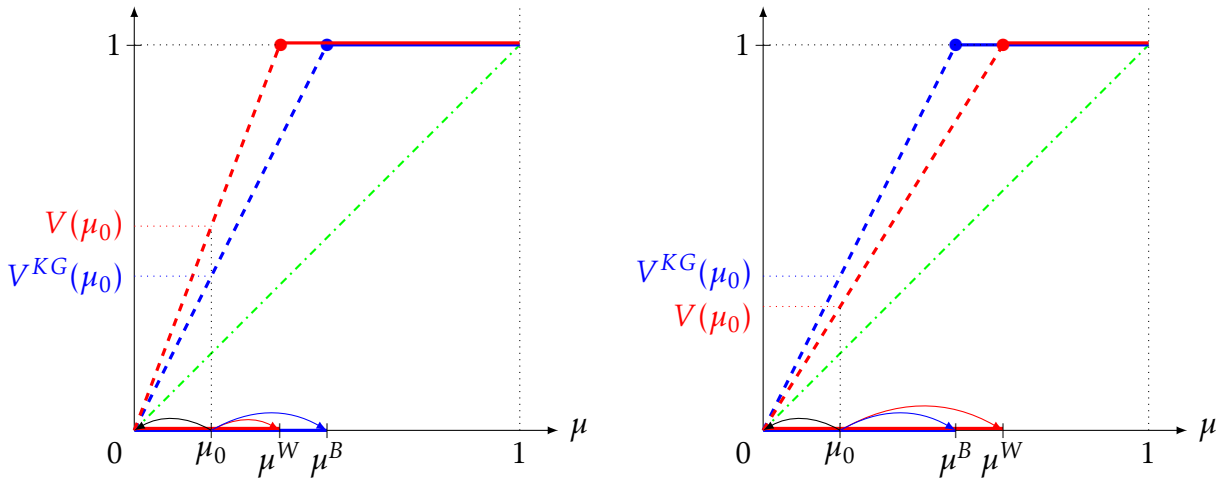
$$\int_{[0,1]} \mu\tau(\mathrm{d}\mu) = \mu_0.$$

That is, the threshold belief μ^W corresponds to the *smallest Bayesian posterior Sender needs to induce to persuade a wishful Receiver to take action* $a = 1$. Therefore, Lemma 1 has immediate consequences for Sender.

Corollary 2. *Sender's optimal information policy is weakly less (Blackwell) informative and achieves a weakly higher payoff when interacting with a wishful Receiver compared to a Bayesian for any prior* $\mu_0 \in]0,1[$ *if and only if one of the conditions (i), (ii) or (iii) in Lemma 1 is satisfied.*

To illustrate Corollary 2 we represent in Figure 1 the concavifications of Sender's indirect utility when Receiver is wishful or Bayesian in two different cases. The case corresponding to Lemma 1 is represented in Figure 1a. Sender is always better-off persuading a wishful compared to a Bayesian receiver as $V(\mu_0) \geq V^{KG}(\mu_0)$ for any $\mu_0 \in]0,1[$. On the other hand, if Receiver's preferences or ability to self-deceive do not satisfy any of the properties in Lemma 1, then Sender is weakly worse-off under any prior. This case is represented on Figure 1b.

When Sender wants to induce an action that is (resp. is not) favored by a wishful Receiver, persuasion is always "easier" (resp. "harder") for Sender in the following sense: Sender needs a strictly less (resp. strictly more) Blackwell informative policy than KG to



(a) At least one property in Lemma 1 is satisfied. (b) No property in Lemma 1 is satisfied.

Figure 1: Expected payoffs under optimal information policies. Red curves: expected payoffs under wishful thinking. Blue curves: expected payoffs when Receiver is Bayesian. Dashed-dotted green lines: expected payoffs under a fully revealing experiment.

persuade Receiver to take his preferred action. Equivalently, if experiments were costly to produce, as in Gentzkow and Kamenica (2014), then Sender would always need to consume less (resp. more) resources to persuade a wishful Receiver to take his preferred action than a Bayesian. The hypothesis of a binary state space facilitates the comparisons between the Bayesian-optimal and the wishful-optimal information policies as it ensures that the Bayesian-optimal and the wishful-optimal information policies are Blackwell comparable. Although the informativeness comparisons in Corollary 2 do not necessarily extend when the state space contains more than two elements it is not difficult to show that Sender's welfare comparisons still hold under any arbitrary finite state space.

Proposition 2. *Assume Θ is a finite set with at least two elements. Sender always achieves a weakly higher payoff for any prior $\mu_0 \in \Delta\Theta$ if, and only if, for any pair of states $\theta, \theta' \in \Theta$, one of the conditions (i), (ii) or (iii) in Lemma 1 is satisfied.*

Proof. See Appendix A.3. □

Indeed, if the states satisfy the properties of Lemma 1 pairwise, then the set of posteriors under which a motivated receiver would take action $a = 1$ is always a superset of the set of posteriors under which a Bayesian receiver would take action $a = 1$. We illustrate Proposition 2 in the following example with three states.

Example with ternary state space. In this example $\Theta = \{0, 1, 2\}$. We start by defining the sets

$$\Delta_a^B = \{\mu \in \Delta\Theta : a \in A(\mu)\},$$

and

$$\Delta_a^W = \{\mu \in \Delta\Theta : a \in A(\eta(\mu))\},$$

for any $a \in A$. The set Δ_a^B (resp. Δ_a^W) is precisely the subset of beliefs supporting an action a as optimal for a Bayesian (resp. wishful) Receiver. We say that action a is favored by a wishful Receiver if $\Delta_a^B \subset \Delta_a^W$. Moreover, denote $\mu^B_{\theta,\theta'}$ (resp. $\mu^W_{\theta,\theta'}$) the belief making a Bayesian (resp. wishful) Receiver indifferent between actions $a = 0$ and $a = 1$ when $\mu(\theta), \mu(\theta') > 0$ but $\mu(\theta'') = 0$ for any $\theta, \theta', \theta'' \in \Theta$. Preferences of Receiver are represented by the following utility function:

$u(a,\theta)$	$\theta = 0$	$\theta = 1$	$\theta = 2$
$a = 0$	2	3	-1
$a = 1$	1	0	4

Notice that for the two pairs of states $(0, 2)$ and $(1, 2)$, the associated payoffs satisfy property (i) in Lemma 1. That is, action $a = 1$ is associated to the highest payoff $u(a_1, 2) = 4$ as well as the highest payoff variability $u(a_1, 2) - u(a_0, 2) = 5$. As a consequence, Lemma 1 applies whenever focusing on those two pairs of states. Then, we have $\mu^W_{0,2} > \mu^B_{0,2}$ and $\mu^W_{1,2} > \mu^B_{1,2}$. Remark now, that $\Delta_1^B = \text{co}(\{\mu^B_{0,2}, \mu^B_{1,2}, \delta_2\})$ and $\Delta_1^W = \text{co}(\{\mu^W_{0,2}, \mu^W_{1,2}, \delta_2\})$, where δ_θ denotes the Dirac distribution on state $\theta \in \Theta$ (see Figure 2). Consequently, $\Delta_1^B \subset \Delta_1^W$ so action $a = 1$ is favored by Receiver and Sender is better-off. When the state space is finite, a policy $\tau \in \mathcal{T}(\mu_0)$ such that all elements in $\text{supp}(\tau)$ are affinely independents is (weakly) more Blackwell-informative than a policy $\tau' \in \mathcal{T}(\mu_0)$ if, and only if, and $\text{supp}(\tau') \subset \text{co}(\text{supp}(\tau))$ (see Lipnowski et al., 2020, Lemma 2). The support of the Bayesian-optimal policy τ^B (resp. wishful-optimal policy τ^W) is $\{\mu^B_-, \mu^B_{0,2}\}$ (resp. $\{\mu^W_-, \mu^W_{0,2}\}$). Hence, $\text{co}(\text{supp}(\tau^W)) = \{\mu \in \Delta\Theta : \exists t \in [0,1], \mu = t\mu^W_- + (1-t)\mu^W_{0,2}\}$. It is visible on Figure 2 that $\{\mu^B_-, \mu^B_{0,2}\} \not\subset \text{co}(\text{supp}(\tau^W))$. Hence, τ^B and τ^W are not Blackwell comparable.

We now show in two applications that Corollary 2 has important economic consequences for preventive health care information provision as well as for political beliefs polarization.

Information provision and preventive health care. A public health agency (Sender) informs an individual (Receiver) about the prevalence of a certain disease. Receiver forms

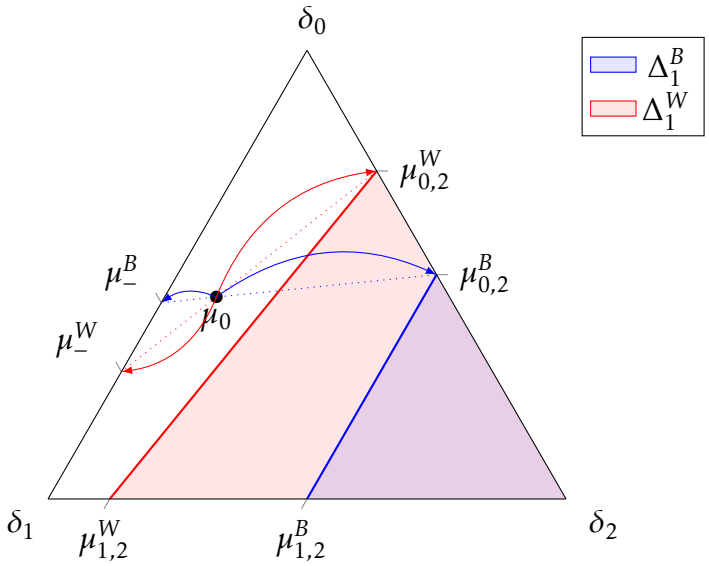


Figure 2: The Bayesian-optimal policy τ^B (in blue) vs. the wishful-optimal policy τ^W (in red) with respective supports $\{\mu_-^B, \mu_{0,2}^B\}$ and $\{\mu_-^W, \mu_{0,2}^W\}$.

beliefs about the infection risk, which can be either high or low: $0 < \underline{\theta} < \overline{\theta} < 1$. The probability of contracting that illness also depends on whether the individual adopts a preventive treatment or not, where $a = 1$ designates adoption. Investment in the treatment entails a cost $c > 0$ to Receiver.[17] Moreover, let us assume that the effectiveness of the treatment, i.e., the probability that the treatment works is $\alpha \in [0,1]$ so that the probability of falling ill, conditional on adoption, is $(1-\alpha)\theta$. The payoff from staying healthy is normalized to 0 whereas the payoff from being infected equals $-\varsigma < 0$ where ς is the severity of the disease. Receiver's utility function is

$$u(a,\theta) = (1-a)(-\varsigma\theta) + a(-(1-\alpha)\theta\varsigma - c)$$

for any $(a,\theta) \in A \times \Theta$. We assume that $\varsigma\alpha\underline{\theta} < c < \varsigma\alpha\overline{\theta}$ so Receiver faces a trade-off: he would prefer not to invest if he was sure the probability of infection was low and, conversely, would prefer to invest in the treatment if he was sure the risk of infection is high. Also remark that Receiver always expect to experience a negative payoff, as $u(a,\theta) < 0$ for any $(a,\theta) \in A \times \Theta$. So $\int_\Theta u(a,\theta)\eta(\mathrm{d}\theta)$ can be interpreted as the anticipated anxiety of Receiver under belief $\eta \in [0,1]$ and action a.

A public health agency wants to maximize the probability of individuals adopting the

[17] One might interpret that cost to be the price of the treatment or the either material or psychological cost from undertaking medical procedures.

preventive treatment.[18] The agency informs individuals about the prevalence of the disease by designing and committing to a Bayes-plausible information policy τ. A Bayesian Receiver would be indifferent between adopting or not the treatment at belief

$$\mu^B = \frac{c - \alpha\underline{\theta}\varsigma}{\alpha(\overline{\theta} - \underline{\theta})\varsigma}.$$

In contrast, by Proposition 1 and Corollary 1, the optimal beliefs and behavior of a wishful Receiver are given by

$$\eta(\mu) = \begin{cases} \dfrac{\mu}{\mu + (1-\mu)\exp(\rho\varsigma(\overline{\theta} - \underline{\theta}))} & \text{if} \quad \mu < \mu^W \\ \\ \dfrac{\mu\exp(-\rho(1-\alpha)\varsigma(\overline{\theta} - \underline{\theta}))}{\mu\exp(-\rho(1-\alpha)\varsigma(\overline{\theta} - \underline{\theta})) + (1-\mu)} & \text{if} \quad \mu \geq \mu^W \end{cases},$$

and

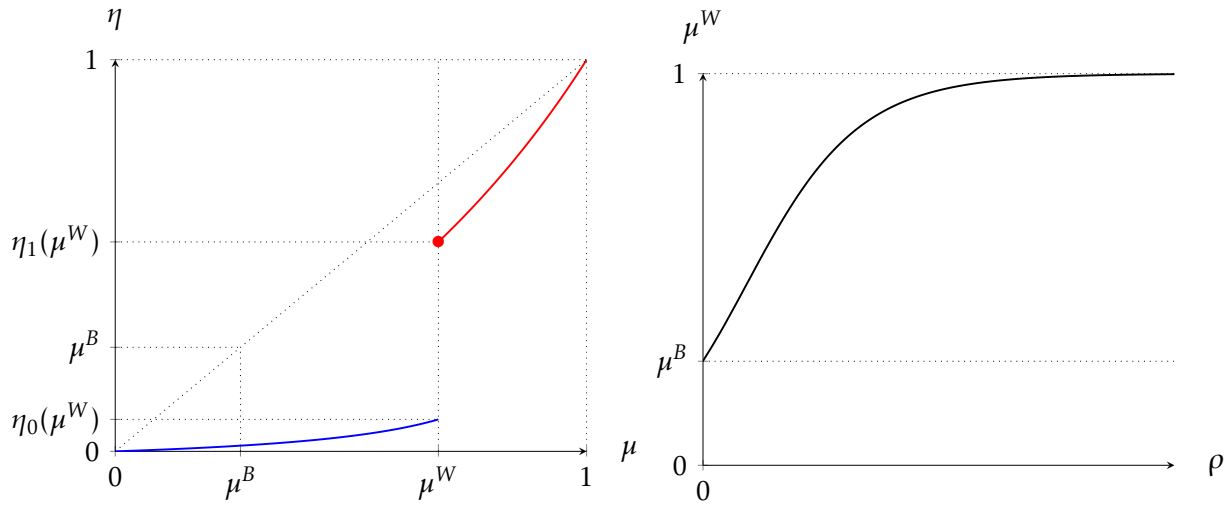
$$a(\eta(\mu)) = \begin{cases} 1 & \text{if} \quad \mu \geq \mu^W \\ 0 & \text{if} \quad \mu < \mu^W \end{cases},$$

for any posterior belief $\mu \in [0,1]$, where

$$\mu^W = \frac{\exp(-\rho\underline{\theta}\varsigma) - \exp(\rho(-(1-\alpha)\underline{\theta}\varsigma - c))}{\exp(-\rho\varsigma\underline{\theta}) - \exp(\rho(-(1-\alpha)\underline{\theta}\varsigma - c)) + \exp(\rho(-(1-\alpha)\overline{\theta}\varsigma - c)) - \exp(-\rho\overline{\theta}\varsigma)}.$$

We illustrate the belief distortion of Receiver in Figure 3a. Receiver is always overoptimistic about his probability of staying healthy, as $\eta(\mu) \leq \mu$ for any $\mu \in [0,1]$. Remark that non-adoption is associated with the highest possible payoff $-\varsigma\underline{\theta}$ as well as the highest payoff variability $\varsigma(\overline{\theta} - \underline{\theta})$. Accordingly, by Lemma 1, Receiver always privileges non adoption as illustrates Figure 3b. As a result of Corollary 2, Sender always needs to induce higher beliefs for Receiver to adopt the treatment than she would need if she faced a Bayesian agent, all the more so when Receiver's ability to self-deceive ρ becomes larger. Therefore in this example, overoptimism of Receiver always goes against Sender. The reason for

[18] Maximizing the probability of adoption is a sensible objective. Most infections cause negative externalities due to their transmission through social interactions. Therefore, a benevolent planner who wants to reduce the likelihood of transmission of an infection would do well to maximize the rate of adoption of the preventive treatment (for example, maximize condom distribution to control AIDS transmission, maximize injection of vaccines to control viral infections, or maximize mask use to control the spread of airborne diseases).



(a) Optimal belief $\eta(\mu)$ as a function of μ.

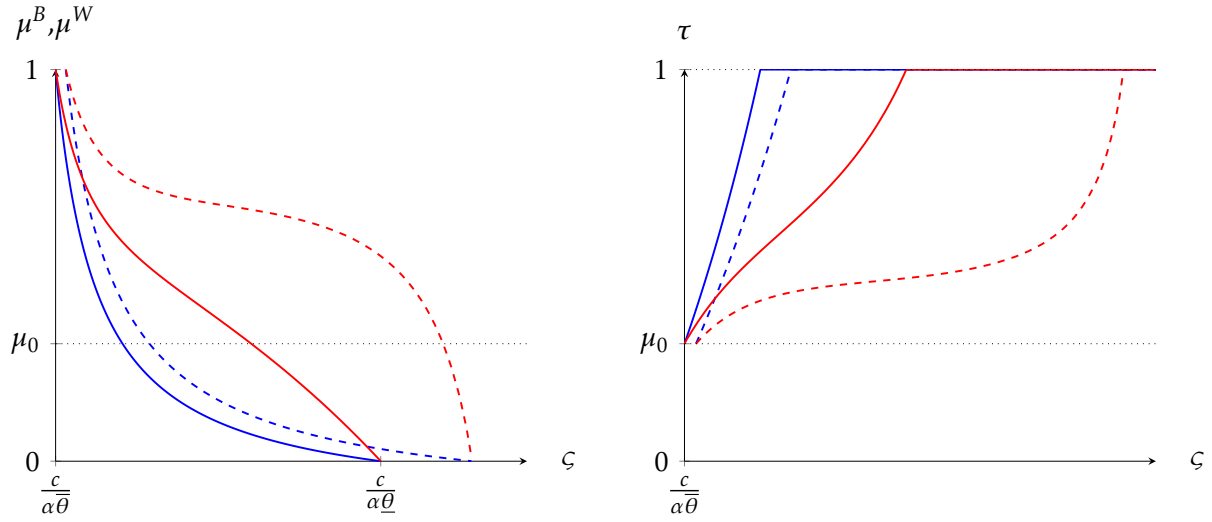
(b) Behavioral threshold μ^W as a function of ρ.

Figure 3: The optimal belief correspondence for $\varsigma = 2$, $c = 0.5$, $\alpha = 0.8$, $\underline{\theta} = 0.1$, $\overline{\theta} = 0.9$ and $\rho = 2$. Receiver is always overoptimistic concerning his health risk for any induced posterior, except at $\mu = 0$ or $\mu = 1$. Moreover, the belief threshold μ^W as a function of ρ is strictly increasing and admits μ^B as a lower bound.

this can be interpreted easily: Receiver's optimism bias and Sender's interests are always misaligned.

It is interesting to see how Sender's probability to induce adoption evolves with respect to the severity of the disease ς as well as the effectiveness of the treatment α.[19] We represent on Figure 4b the probability that Sender induces adoption of the treatment under the optimal information policy as a function of ς. Notice that the probability of inducing adoption is less sensitive to the severity of the disease, i.e., becomes "flatter," when facing a wishful Receiver compared to the Bayesian when the treatment becomes less effective. The intuition is the following: when the treatment is fully effective, i.e., $\alpha = 1$, Receiver's payoff in case he invests in the treatment becomes state independent. Therefore, he does not have any incentive to self-deceive when taking action $a = 1$. As a result, μ^W decreases and Receiver holds perfectly Bayesian beliefs when $\mu \geq \mu^W$. However, whenever there is uncertainty about the treatment efficacy, i.e., $\alpha < 1$, uncertainty about infection risk matters and gives room to self-deception even when taking the treatment. Decreasing α increases the anticipated anxiety of Receiver leading to more optimistically biased beliefs, a higher μ^W and, in turn, complicates persuasion for Sender for any severity s. Remark on

[19] This probability is pinned down by the Bayes-plausibility constraint and equal to $\tau^{KG} = \mu_0/\mu^B$ in the Bayesian case and $\tau = \mu_0/\mu^W$ in the wishful case.



(a) Behavioral thresholds μ^B (in blue) and μ^W (in red) as functions of severity ς.

(b) Probability τ of inducing treatment adoption as a function of severity ς.

Figure 4: Red (resp. blue) curves correspond to wishful (resp. Bayesian) Receiver. We set parameters to $c = 0.5$, $\alpha = 0.8$, $\underline{\theta} = 0.1$, $\overline{\theta} = 0.9$ and $\rho = 2$. Full lines correspond to the case where $\alpha = 1$ whereas dashed curves correspond to $\alpha = 0.8$.

Figure 4b that τ decreases sharply with α for a fixed s. In fact, one could show that as α decreases, τ is becomes closer and closer to μ_0 for any s, meaning that the agency cannot achieve a substantially higher payoff than under non disclosure.[20]

Public persuasion and political polarization. A Sender (e.g., a politician, a lobbyist) persuades an odd-numbered finite group of voters $N = \{1, \ldots, n\}$ (e.g., a committee or parliamentary members) to adopt a proposal $x \in X = \{0, 1\}$, where $x = 0$ corresponds to the status-quo. Uncertainty is binary, $\Theta = \{0, 1\}$, and the audience uses only the information disclosed by Sender to vote on the proposal. Let $a^i \in A = \{0, 1\}$ be the ballot cast by voter i, where $a^i = 0$ designates voting for the status-quo. The proposal is accepted if it is supported by a simple majority of voters. We assume Sender is only interested in the proposal being accepted, so her utility is $v(x) = x$ for every $x \in X$. In contrast, any voter

[20] One additional implication of this result is the following. Assume the true treatment efficacy is α but Receiver perceives the efficacy to be $\hat{\alpha} < \alpha$ (e.g. because Receiver adheres to anti-vaccines movements or generally mistrusts the pharmaceutical industry). In that case, the doubts expressed by Receiver about the treatment efficacy makes him even more anxious which, in turn, makes self-deception stronger and, thus, downplays the effectiveness of the agency's information policy whatever is the severity of the disease.

$i \in N$ has utility function

$$u^i(x,\theta) = x\theta\beta^i + (1-x)(1-\theta)(1-\beta^i)$$

for any $(x,\theta) \in X \times \Theta$ where $\beta^i \in [0,1]$ parametrizes the partisan preference of voter i. That is, all voters agree that the proposal should be implemented only when $\theta = 1$, but they vary in how much they value the (correct) implementation of the proposal versus the (correct) maintenance of the status-quo. The proposal is accepted if it is supported by a simple majority of voters. We assume β^i is symmetrically distributed around $\frac{1}{2}$ in the population of voters. Denote $\beta^m = \frac{1}{2}$ the median voter's preference.

Voters all form motivated beliefs and ρ is assumed homogeneous among all voters. We assume N to be large so, consistently with theories of expressive voting, electors vote sincerely as no voter has an instrumental interest in voting since the probability of being pivotal is null. We further assume that each voter expects the median voter to vote as himself, thus expecting their vote to reflect the outcome of the election. This assumption reflects voters tendency for overoptimism, and is particularly sensible since, as will become apparent shortly, in equilibrium the median voter will be indifferent between voting for or against the proposal. As a result, the direction as well as the magnitude of voters' belief distortion depends only on their partisan preferences β and self-deception ability ρ.

As a result, the direction as well as the magnitude of voters' belief distortion depends only on their partisan preferences.[21] By Proposition 1, voter i's belief under posterior $\mu \in [0,1]$ is given by

$$\eta(\mu,\beta^i) = \begin{cases} \dfrac{\mu}{\mu + (1-\mu)\exp(\rho(1-\beta^i))} & \text{if} \quad \mu < \mu^W(\beta^i) \\ \\ \dfrac{\mu\exp(\rho\beta^i)}{\mu\exp(\rho\beta^i) + (1-\mu)} & \text{if} \quad \mu \geq \mu^W(\beta^i) \end{cases}.$$

where

$$\mu^W(\beta^i) = \frac{\exp(\rho(1-\beta^i)) - 1}{\exp(\rho(1-\beta^i)) + \exp(\rho\beta^i) - 2}.$$

[21] It has been shown in psychology (Babad et al., 1992; Babad, 1995, 1997) as well as in behavioral economics (Thaler, 2020) that voters political beliefs are often motivated by their partisan orientation.

Given sincere voting, voter i's behavior under optimal belief $\eta(\mu, \beta^i)$ is given by

$$a(\eta(\mu,\beta^i)) = \begin{cases} 1 & \text{if} \quad \mu \geq \mu^W(\beta^i) \\ 0 & \text{if} \quad \mu < \mu^W(\beta^i) \end{cases}.$$

Due to the heterogeneity in β, there is always some level of belief polarization among wishful voters for any $\mu \in]0,1[$. Let us measure such polarization by the sum of the absolute difference between each pair of beliefs in the audience

$$\pi(\mu) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\eta(\mu,\beta^i) - \eta(\mu,\beta^j)| \tag{4}$$

for any $\mu \in [0,1]$.[22]

Proposition 3. *Under Sender's optimal information policy, the signal that leads to the implementation of the proposal also generates the maximum polarization among voters.*

Proof. See Appendix A.4. □

To build an intuition of why this is the case, let's first first note that, in our model, belief polarization and action polarization are closely related. Agents voting for the implementation of the proposal will distort their beliefs upwards, whereas agents voting for the status quo will distort their beliefs downwards. We can thus see that maximum belief polarization should be attained for some belief for which action polarization is maximized, that is, for some belief at which $\frac{n+1}{2}$ agents are voting one way and the remaining $\frac{n-1}{2}$ are voting another way. This will be the case for any $\mu \in [\mu^W(\beta^{m-1}), \mu^W(\beta^{m+1})[$.

Due to sincere voting, the result of the election always coincides with the vote of the median voter under posterior belief μ. Accordingly, Sender's indirect utility is

$$v(\mu) = \mathbb{1}\{\mu \geq \mu^W(\beta^m)\},$$

for any $\mu \in [0,1]$. The optimal information policy for Sender is thus supported on $\{0, \mu^W(\beta^m)\}$ whenever $\mu_0 \in]0, \frac{1}{2}[$, and on $\{\mu_0\}$ whenever $\mu_0 \in]\mu^W(\beta^m), 1[$. The posterior $\mu^W(\beta^m)$, which leads to the implementation of the proposal, belongs to the interval $[\mu^W)(\beta^{m-1}), \mu^W(\beta^{m+1})[$

[22] Note that we can always index the voters in an ascending order of β, such that $\eta(\mu,\beta^i) \geq \eta_j(\mu)$ for all $\mu \in \Delta\Theta$ whenever $i < j$, such that Equation (4) does indeed represent the absolute difference between each pair of beliefs.

and, as such, is in the neighbourhood of the belief that maximizes polarization for any distribution of preferences. When such distribution is symmetric around the median voter, polarization is maximized exactly at the middle point in that interval, which is $\mu^W(\beta^m)$.

We illustrate Proposition 3 below in Section 4.1 in a setup with 3 voters. Following

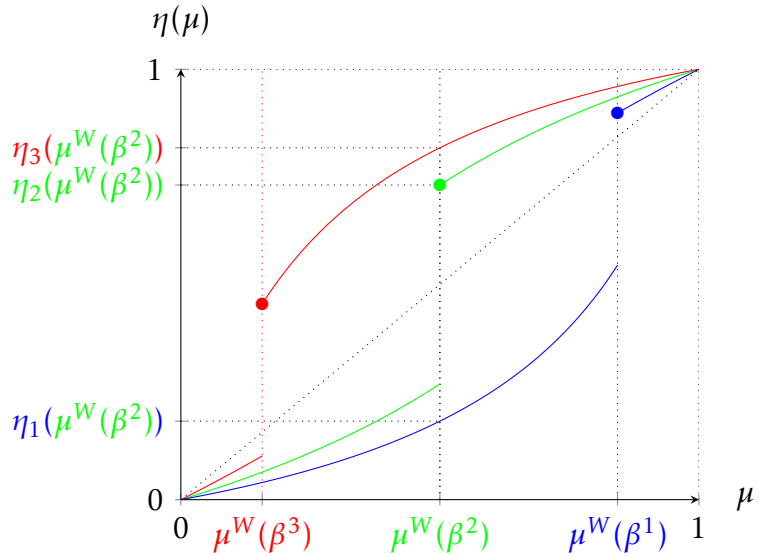


Figure 5: Beliefs distortions in the electorate for $\rho = 2$, $\beta_1 = 1/4$, $\beta_2 = 1/2$ and $\beta_3 = 3/4$. Polarization equals $\pi(\mu) = 2(\eta(\mu, \beta^1) - \eta(\mu, \beta^3))$ which is maximized at $\mu^W(\beta^2) = \frac{1}{2}$.

Corollary 1, wishful thinking induces voters to switch from disapproval to approval at different Bayesian posteriors $\mu^W(\beta^i)$. The optimal information policy τ for Sender is the one that maximizes the probability of the median voter voting for the approval. That is, $\text{supp}(\tau) = \{0, \mu^W(\beta^m)\}$ and $\mu^W(\beta^m) = 1/2$ is induced with probability $p = \mu^W(\beta^m)/\mu_0$ whenever $\mu_0 \in]0, \mu^W(\beta^2)[$ and $\text{supp}(\tau) = \{\mu_0\}$ whenever $\mu_0 \in]\mu^W(\beta^2), 1[$.

Let us now turn to polarization. First, it is quite easy to see in Section 4.1 that

$$\pi(\mu) = 2(\eta(\mu, \beta^1) - \eta(\mu, \beta^3))$$

for any $\mu \in [0, 1]$, as the distances to the median belief add up to $\eta(\mu, \beta^1) - \eta(\mu, \beta^3)$. Thus, it suffices to check where $\eta(\mu, \beta^1) - \eta(\mu, \beta^3)$ is maximized. Quite naturally, polarization is maximized when the posterior belief induced by Sender is in between $\mu^W(\beta^3)$ and $\mu^W(\beta^1)$. In particular, it is exactly maximized at the posterior belief $\mu^W(\beta^2) = \frac{1}{2}$ which is exactly the posterior belief Sender induces to obtain the approval of the proposal under her optimal policy.

Proposition 3 establishes that the intuition developed in this example is generally valid

when the partisan preferences of voters are symmetrically distributed around the median. In other words, attempts by a rational sender to maximize the probability of approval induces, as an externality, maximal belief polarization among wishful voters. This result differs from the literature studying the possible heterogeneity of beliefs due to deliberate attempts at persuasion[23], which tends to focus on polarization arising from differential access to information. Our model gives an alternative mechanism to the rise of polarization, based on motivated beliefs: a sender can induce polarization involuntarily when her message is subject to motivated interpretations, and such polarization might be especially large whenever sender's strategy involves targeting an agent with a median preference.

In the next subsection we extend out framework to the case of a continuous state space and linear preferences. We show that results in the finite state space case extend to this setting. We also highlight why we might expect persuasion to be more effective in the context of risky investment decisions.

4.2 Continuous state space: persuading a wishful investor

A financial broker (Sender) designs reports about the return of some risky financial product to inform a potential client (Receiver). The return of the product is $\theta \in \Theta = [\underline{\theta}, \overline{\theta}]$, where $\underline{\theta} < 0 < \overline{\theta}$. Returns are distributed according to the prior distribution μ_0. Let F be the cumulative distribution function associated with μ_0 and let us assume that μ_0 admits a continuous and strictly positive density function f over $[\underline{\theta}, \overline{\theta}]$. Receiver has some saved up money he is willing to invest and chooses action $a \in A = \{0, 1\}$, where $a = 0$ represents the choice of non-investing in which case Receiver's payoff is 0 and $a = 1$ represents investing, in which case Receiver's payoff is the product realized return θ. The broker is remunerated on the basis of a flat fee $v > 0$ that is independent of the true product's profitability. Hence, Receiver's payoff is $u(a, \theta) = a\theta$ while Sender's payoff is $v(a, \theta) = va$ for any $(a, \theta) \in A \times \Theta$.

Receiver forms motivated beliefs about the return of the financial product. By Proposition 1 his optimal beliefs are given by

$$\eta(\mu)(\tilde{\Theta}) = \begin{cases} \mu(\tilde{\Theta}) & \text{if } \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta) < 1 \\ \frac{\int_{\tilde{\Theta}} \exp(\rho\theta)\mu(\mathrm{d}\theta)}{\int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta)} & \text{if } \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta) \geq 1 \end{cases},$$

[23] See Arieli and Babichenko (2019) for general considerations on the private persuasion of multiple receivers and see Chan et al. (2019) for an application to voting.

for any $\mu \in \Delta\Theta$ and any Borel set $\tilde{\Theta} \subseteq \Theta$, and, by Corollary 1, his equilibrium behavior is given by

$$a(\eta(\mu)) = \begin{cases} 0 & \text{if } \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta) < 1 \\ 1 & \text{if } \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta) \geq 1 \end{cases}.$$

Therefore, Sender's indirect utility is equal to

$$v(\mu) = \begin{cases} 0 & \text{if } \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta) < 1 \\ v & \text{if } \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta) \geq 1 \end{cases}.$$

for any $\mu \in \Delta\Theta$. To make the problem interesting, we assume that neither a Bayesian nor a wishful Receiver would take action $a = 0$ under the prior. That is, $\hat{m} = \int_{\underline{\theta}}^{\overline{\theta}} \theta\mu_0(\mathrm{d}\theta) < 0$ and $\hat{x} = \int_{\underline{\theta}}^{\overline{\theta}} \exp(\rho\theta)\mu_0(\mathrm{d}\theta) < 1$.[24]

Under these assumptions, remark that a signal structure σ that induces a distribution τ over posterior beliefs μ matters for Receiver and Sender only through the *distribution of exponential moments* $x = \int_\Theta \exp(\rho\theta)\mu(\mathrm{d}\theta)$ *it induces*. Let X be the space of such moments, that is, $X = \mathrm{co}(\exp(\rho\Theta))$, where $\exp(\rho\Theta)$ is the graph of the function $\theta \mapsto \exp(\rho\theta)$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$. That is, $X = [\underline{x}, \overline{x}]$ where $\underline{x} = \exp(\rho\underline{\theta})$ and $\overline{x} = \exp(\rho\overline{\theta})$. Let G be the prior cumulative distribution function over the random variable $\exp(\rho\theta)$ induced by F, that is

$$G(x) = F\left(\frac{\ln(x)}{\rho}\right),$$

for any $x \in [\underline{x}, \overline{x}]$. By standard arguments (Gentzkow and Kamenica, 2016), the problem of finding an optimal signal structure σ reduces to finding a cumulative distribution function H that maximizes

$$\int_{\underline{x}}^{\overline{x}} v(x)\mathrm{d}H(x)$$

subject to

$$\int_{\underline{x}}^{z} H(x)\mathrm{d}x \leq \int_{\underline{x}}^{z} G(x)\mathrm{d}x$$

for every $z \in [\underline{x}, \overline{x}]$. The solution to such a problem is well-known and can be found either using optimization under stochastic dominance constraints techniques (Gentzkow and Kamenica, 2016; Ivanov, 2020; Kleiner et al., 2021) or linear programming duality

[24] It is in fact always true that $\hat{m} < 0$ when $\hat{x} < 1$. Hence, assuming $\hat{m} < 0$ additionally to $\hat{x} < 1$ is without loss.

arguments (Kolotilin, 2018; Dworczak and Martini, 2019; Dizdar and Kováč, 2020). In our context, the optimal signal is a binary partition of the state space. That is, Sender reveals whether the realized state is above or below some threshold state.

Proposition 4. *There exists a unique* $\theta^W \in [\underline{\theta}, \overline{\theta}]$ *verifying*

$$\frac{1}{1-F(\theta^W)} \int_{\theta^W}^{\overline{\theta}} \exp(\rho\theta) f(\theta) \mathrm{d}\theta = 1$$

and such that Sender pools all states $\theta \in [\theta^W, \overline{\theta}]$ *under the same signal* $s = 1$*, i.e.,* $\sigma(1|\theta) = 1$ *for all* $\theta \in [\theta^W, \overline{\theta}]$*, and similarly pools all states* $\theta \in [\underline{\theta}, \theta^W]$ *under the same signal* $s = 0$*. Hence, the probability of inducing action* $a = 1$ *for Sender is equal to*

$$\int_{\theta^W}^{\overline{\theta}} \sigma(1|\theta) f(\theta) \mathrm{d}\theta = 1 - F(\theta^W).$$

Proof. See Ivanov (2020), Section 3. □

It is therefore optimal for Sender to partition the state space at the threshold state making Receiver indifferent between investing or not at the prior. In our context, such an information policy can intuitively be seen as the investment recommendation rule which maximizes the probability that Receiver invests given the prior distribution of returns F.

Using the exact same arguments as above, one can deduce that the probability of inducing action $a = 1$ when Receiver is Bayesian is given by $1 - F(\theta^B)$ where θ^B is the unique threshold verifying the equation

$$\frac{1}{1-F(\theta^B)} \int_{\theta^B}^{\overline{\theta}} \theta f(\theta) \mathrm{d}\theta = 0.$$

Therefore, Sender is more effective at persuading a wishful Receiver if and only if $\theta^W < \theta^B$.

Proposition 5. *It is always true that* $\theta^W < \theta^B$*. Hence, Sender is always more effective at persuading a wishful than a Bayesian investor.*

Proof. See Appendix A.5. □

The above result relates to Proposition 2: buying the risky product is favored by the wishful investor since it is the action that yields both the highest possible payoff and the

highest payoff variability. This example thus illustrates how the results in the finite state space case naturally extend to an infinite state space setting with linear preferences. It further helps explaining the pervasiveness of persuasion efforts in financial and betting markets, illustrating why some financial consulting firms seem to specialize in advice misconduct and cater to biased consumers (Egan et al., 2019).

5 Conclusion

In this paper we study optimal persuasion in the presence of a wishful Receiver. By modeling wishful thinking as a process that optimally trades-off gains in anticipatory utility with the cost of distorting beliefs, we characterize the correspondence between motivated and objective beliefs, highlighting the particularities that such belief formation process entails.

In particular, we show that wishful thinking impacts behavior, causing some actions to be favored in the sense that they are taken at a greater set of beliefs. This has important implications for the strategic design of information, as it adds some nuance on the way preferences and information determine behavior. Concretely, we show that, in the presence of wishful thinking, persuasion is more effective when it is aimed at inducing an action that is risky but can potentially yield a very large return and less effective when it is aimed at inducing a more cautious action. We use this model to illustrate why information disclosure seems less effective than expected at inducing preventive health behavior and more effective than expected at inducing dubious financial investments. Wishful thinking opens a channel for preferences to interfere in belief formation, raising the question of what kind of belief polarization could we observe in a population in which agents have access to the same information but vary in their preferences. An information designer interested in the approval of a proposal would, by optimally targeting the median voter in her choice of signal structure, induce, as an externality, maximum polarization among the electorate whenever the proposal is approved.

Interestingly, some studies already investigate the effects of wishful thinking on the outcomes of strategic interactions (see, Yildiz, 2007; Banerjee et al., 2020; Heller and Winter, 2020). Further investigation on ways in which individual preferences might impact information processing and how these may impact social phenomena such as belief polarization in non-strategic and strategic settings seem to be promising paths for future research.

References

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics*, 113:96–104. 10

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4):1115–1153. 10

Akerlof, G. A. and Dickens, W. T. (1982). The Economic Consequences of Cognitive Dissonance. *American Economic Review*, 72(3):307–319. 7

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis*. Springer-Verlag, Berlin/Heidelberg. 9

Arieli, I. and Babichenko, Y. (2019). Private Bayesian persuasion. *Journal of Economic Theory*, 182:185–217. 25

Babad, E. (1995). Can Accurate Knowledge Reduce Wishful Thinking in Voters' Predictions of Election Outcomes? *The Journal of Psychology*, 129(3):285–300. 4, 22

Babad, E. (1997). Wishful thinking among voters: motivational and cognitive Iinfluences. *International Journal of Public Opinion Research*, 9(2):105–125. 22

Babad, E., Hills, M., and O'Driscoll, M. (1992). Factors Influencing Wishful Thinking and Predictions of Election Outcomes. *Basic and Applied Social Psychology*, 13(4):461–476. 22

Babad, E. and Katz, Y. (1991). Wishful Thinking—Against All Odds. *Journal of Applied Social Psychology*, 21(23):1921–1938. 5

Banerjee, S., Davis, J., and Gondhi, N. (2020). Motivated Beliefs in Coordination Games. *SSRN Electronic Journal*. 28

Bénabou, R. (2015). The Economics of Motivated Beliefs. *Revue d'économie politique*, 125(5):665–685. 10

Bénabou, R. and Tirole, J. (2002). Self-Confidence and Personal Motivation. *Quarterly Journal of Economics*, 117(3):871–915. 7, 10

Bénabou, R. and Tirole, J. (2004). Willpower and Personal Rules. *Journal of Political Economy*, 112(4):848–886. 7

Bénabou, R. and Tirole, J. (2006). Belief in a Just World and Redistributive Politics. *Quarterly Journal of Economics*, 121(2):699–746. 7

Bénabou, R. and Tirole, J. (2011). Identity, Morals, and Taboos: Beliefs as Assets *. *Quarterly Journal of Economics*, 126(2):805–855. 7

Bénabou, R. and Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, 30(3):141–164. 2, 6, 10

Benjamin, D., Bodoh-Creed, A., and Rabin, M. (2019). Base-Rate Neglect: Foundations and Implications. 5

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 2*, volume 2, chapter 2, pages 69–186. Elsevier B.V. 2, 5

Bergemann, D. and Morris, S. (2016). Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium. *American Economic Review*, 106(5):586–591. 5

Bergemann, D. and Morris, S. (2019). Information Design: A Unified Perspective. *Journal of Economic Literature*, 57(1):44–95. 5

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2018). Behavioral Household Finance. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics: Applications and Foundations 1*, chapter 3, pages 177–276. Elsevier B.V. 5

Bizzotto, J., Rüdiger, J., and Vigier, A. (2018). Dynamic Persuasion With Outside Information. *SSRN Electronic Journal*. 6

Bizzotto, J. and Vigier, A. (2020). Can a better informed listener be easier to persuade? *Economic Theory*, (August). 6

Bloedel, A. W. and Segal, I. (2018). Persuasion with Rational Inattention. *SSRN Electronic Journal*. 6

Bracha, A. and Brown, D. J. (2012). Affective decision making: A theory of optimism bias. *Games and Economic Behavior*, 75(1):67–80. 6, 8

Bridet, L. and Schwardmann, P. (2020). Selling Dreams: Endogenous Optimism in Lending Markets. *CESifo Working Paper*, (8271). 6

Broniatowski, M. and Keziou, A. (2006). Minimization of ϕ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442. 38

Brunnermeier, M. K., Gollier, C., and Parker, J. A. (2007). Optimal Beliefs, Asset Prices, and the Preference for Skewed Returns. *American Economic Review*, 97(2):159–165. 6

Brunnermeier, M. K., Papakonstantinou, F., and Parker, J. A. (2017). Optimal Time-Inconsistent Beliefs: Misplanning, Procrastination, and Commitment. *Management Science*, 63(5):1318–1340. 6

Brunnermeier, M. K. and Parker, J. A. (2005). Optimal Expectations. *American Economic Review*, 95(4):1092–1118. 6, 10, 11

Caplin, A. and Leahy, J. (2001). Psychological Expected Utility Theory and Anticipatory Feelings. *Quarterly Journal of Economics*, 116(1):55–79. 10

Caplin, A. and Leahy, J. (2019). Wishful Thinking. *NBER Working Paper Series*. 6, 8, 10, 11, 12

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., and Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, 11(1):2100. 10

Chan, J., Gupta, S., Li, F., and Wang, Y. (2019). Pivotal persuasion. *Journal of Economic Theory*, 180:178–202. 25

Chandra, A., Handel, B., and Schwartzstein, J. (2019). Behavioral economics and health-care markets. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics: Applications and Foundations 2*, chapter 6, pages 459–502. Elsevier B.V. 4

Che, Y.-K., Kim, K., and Mierendorff, K. (2020). Keeping the Listener Engaged: a Dynamic Model of Bayesian Persuasion. *arXiv*. 6

Chew, S. H., Huang, W., and Zhao, X. (2020). Motivated False Memory. *Journal of Political Economy*, 128(10):3913–3939. 10

Coutts, A. (2019). Testing models of belief bias: An experiment. *Games and Economic Behavior*, 113:549–565. 2, 11

de Clippel, G. and Zhang, X. (2020). Non-Bayesian Persuasion. *Working Paper*. 5

Dizdar, D. and Kováč, E. (2020). A simple proof of strong duality in the linear persuasion problem. *Games and Economic Behavior*, 122:407–412. 27

Dupas, P. (2011). Health Behavior in Developing Countries. *Annual Review of Economics*, 3(1):425–449. 4

Dupuis, P. and Ellis, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley. 37

Dworczak, P. and Martini, G. (2019). The Simple Economics of Optimal Persuasion. *Journal of Political Economy*, 127(5):1993–2048. 27

Egan, M., Matvos, G., and Seru, A. (2019). The Market for Financial Adviser Misconduct. *Journal of Political Economy*, 127(1):233–295. 5, 28

Eliaz, K., Spiegler, R., and Thysen, H. C. (2021a). Persuasion with endogenous misspecified beliefs. *European Economic Review*, 134:103712. 5

Eliaz, K., Spiegler, R., and Thysen, H. C. (2021b). Strategic interpretations. *Journal of Economic Theory*, 192:105192. 5

Ely, J., Frankel, A., and Kamenica, E. (2015). Suspense and Surprise. *Journal of Political Economy*, 123(1):215–260. 6

Engelmann, J., Lebreton, M., Schwardmann, P., van der Weele, J. J., and Chang, L.-A. (2019). Anticipatory Anxiety and Wishful Thinking. *SSRN Electronic Journal*. 2, 10

Ettinger, D. and Jehiel, P. (2010). A Theory of Deception. *American Economic Journal: Microeconomics*, 2(1):1–20. 5

Exley, C. and Kessler, J. (2019). Motivated Errors. *NBER Working Paper Series*. 13

Eyster, E. (2019). Errors in strategic reasoning. In *Handbook of Behavioral Economics: Applications and Foundations 2*, volume 2, chapter 3, pages 187–259. Elsevier B.V. 5

Galperti, S. (2019). Persuasion: The Art of Changing Worldviews. *American Economic Review*, 109(3):996–1031. 6

Ganguly, A. and Tasoff, J. (2017). Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future. *Management Science*, 63(12):4037–4060. 4

Gentzkow, M. and Kamenica, E. (2014). Costly Persuasion. *American Economic Review: Papers & Proceedings*, 104(5):457–462. 16

Gentzkow, M. and Kamenica, E. (2016). A Rothschild-Stiglitz Approach to Bayesian Persuasion. *American Economic Review: Papers & Proceedings*, 106(5):597–601. 26

Gollier, C. and Muermann, A. (2010). Optimal Choice and Beliefs with Ex Ante Savoring and Ex Post Disappointment. *Management Science*, 56(8):1272–1284. 6

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information Avoidance. *Journal of Economic Literature*, 55(1):96–135. 10

Golman, R., Loewenstein, G., Moene, K. O., and Zarri, L. (2016). The Preference for Belief Consonance. *Journal of Economic Perspectives*, 30(3):165–188. 7

Hagenbach, J. and Koessler, F. (2020). Cheap talk with coarse understanding. *Games and Economic Behavior*, 124:105–121. 5

Hansen, L. P. and Sargent, T. J. (2008). *Robustness*. Princeton University Press. 8

Heger, S. A. and Papageorge, N. W. (2018). We should totally open a restaurant: How optimism and overconfidence affect beliefs. *Journal of Economic Psychology*, 67(July):177–190. 2

Heller, Y. and Winter, E. (2020). Biased-Belief Equilibrium. *American Economic Journal: Microeconomics*, 12(2):1–40. 28

Inderst, R. and Ottaviani, M. (2012). Financial Advice. *Journal of Economic Literature*, 50(2):494–512. 5

Ivanov, M. (2020). Optimal monotone signals in Bayesian persuasion mechanisms. *Economic Theory*. 26, 27

Jiao, P. (2020). Payoff-Based Belief Distortion. *The Economic Journal*, 130(629):1416–1444. 2

Jouini, E. and Napp, C. (2018). The Impact of Health-Related Emotions on Belief Formation and Behavior. *Theory and Decision*, 84(3):405–427. 6

Kamenica, E. (2019). Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11:249–272. 5

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615. 2, 5, 9

Kleiner, A., Moldovanu, B., and Strack, P. (2021). Extreme Points and Majorization: Economic Applications. *Econometrica*, 89(4):1557–1593. 26

Kolotilin, A. (2018). Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635. 27

Kovach, M. (2020). Twisting the truth: Foundations of wishful thinking. *Theoretical Economics*, 15(3):989–1022. 6

Kremer, M., Rao, G., and Schilbach, F. (2019). Behavioral development economics. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics: Applications and Foundations 2*, chapter 5, pages 345–458. Elsevier B.V. 4

Krizan, Z. and Windschitl, P. D. (2009). Wishful Thinking about the Future: Does Desire Impact Optimism? *Social and Personality Psychology Compass*, 3(3):227–243. 6

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, 53(4):636–647. 2, 10

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498. 2, 10

Lerman, C., Hughes, C., Lemon, S. J., Main, D., Snyder, C., Durham, C., Narod, S., and Lynch, H. T. (1998). What you don't know can hurt you: adverse psychologic effects in members of BRCA1-linked and BRCA2-linked families who decline genetic testing. *Journal of Clinical Oncology*, 16(5):1650–1654. 4

Levy, G., Moreno de Barreda, I., and Razin, R. (2018). Persuasion with Correlation Neglect. *Working Paper*. 5

Lipnowski, E. and Mathevet, L. (2018). Disclosure to a Psychological Audience. *American Economic Journal: Microeconomics*, 10(4):67–93. 6

Lipnowski, E., Mathevet, L., and Wei, D. (2020). Attention Management. *American Economic Review: Insights*, 2(1):17–32. 6, 17

Loewenstein, G. (1987). Anticipation and the Valuation of Delayed Consumption. *The Economic Journal*, 97(387):666–684. 10

Matysková, L. (2018). Bayesian Persuasion with Costly Information Acquisition. *SSRN Electronic Journal*, (March). 6

Mayraz, G. (2011). Wishful Thinking. *SSRN Electronic Journal*. 2

Mayraz, G. (2019). Priors and Desires—A Bayesian Model of Wishful Thinking and Cognitive Dissonance. *Working Paper*. 6

Mijović-Prelec, D. and Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1538):227–240. 2

Montes, A. (2019). Bayesian Persuasion with Rational Inattentive Receiver. *Working Paper*. 6

Mullainathan, S., Noeth, M., and Schoar, A. (2012). The Market for Financial Advice: An Audit Study. Technical report, National Bureau of Economic Research, Cambridge, MA. 5

Mullainathan, S., Schwartzstein, J., and Shleifer, A. (2008). Coarse Thinking and Persuasion *. *Quarterly Journal of Economics*, 123(2):577–619. 5

Oster, E., Shoulson, I., and Dorsey, E. R. (2013). Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease. *American Economic Review*, 103(2):804–830. 4, 6

Panik, M. J. (1993). *Fundamentals of Convex Analysis*. Springer Netherlands, Dordrecht. 44

Saucet, C. and Villeval, M. C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*, 117:250–275. 10

Schwardmann, P. (2019). Motivated health risk denial and preventative health care investments. *Journal of Health Economics*, 65:78–92. 4, 6

Schweizer, N. and Szech, N. (2018). Optimal Revelation of Life-Changing Information. *Management Science*, 64(11):5250–5262. 6

Strzalecki, T. (2011). Axiomatic Foundations of Multiplier Preferences. *Econometrica*, 79(1):47–73. 8

Thaler, M. (2020). The 'Fake News' Effect: Experimentally Identifying Motivated Reasoning Using Trust in News. *SSRN Electronic Journal*. 4, 22

Wei, D. (2021). Persuasion under costly learning. *Journal of Mathematical Economics*, 94:102451. 6

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5):806–820. 2

Yildiz, M. (2007). Wishful Thinking in Strategic Environments. *Review of Economic Studies*, 74(1):319–344. 28

Appendix

A Proofs

A.1 Proof for Proposition 1 and Corollary 1

Let Θ be any Polish space and let $\Delta\Theta$ be the set of probability measures on Θ endowed with its Borel σ-algebra, let also $\mathcal{C}_b(\Theta)$ be the set of bounded continuous and Borel-measurable real-valued functions on Θ. The psychological well-being function is given by

$$W(\eta,\mu) = \int_\Theta u(a,\theta)\eta(\mathrm{d}\theta) - \frac{1}{\rho}\mathrm{D}_{\mathrm{KL}}(\eta\|\mu),$$

whenever $a(\eta) = a$ for some $a \in A$. For any $\eta, \mu \in \Delta\Theta$, by application of the Donsker-Varadhan variational formula (see Dupuis and Ellis, 1997, Lemma 1.4.3) we have

$$\mathrm{D}_{\mathrm{KL}}(\eta\|\mu) = \sup_{u(a,\cdot)\in\mathcal{C}_b(\Theta)} \int_\Theta \rho u(a,\theta)\eta(\mathrm{d}\theta) - \ln\left(\int_\Theta \exp(u(a,\theta))\mu(\mathrm{d}\theta)\right). \tag{5}$$

Taking the Legendre-Fenchel's dual to the variational equality (5) (see Dupuis and Ellis, 1997, Proposition 1.4.2) we get

$$\ln\left(\int_\Theta \exp(u(a,\theta))\mu(\mathrm{d}\theta)\right) = \sup_{\eta\in\Delta\Theta} \int_\Theta \rho u(a,\theta)\eta(\mathrm{d}\theta) - \mathrm{D}_{\mathrm{KL}}(\eta\|\mu). \tag{6}$$

Hence, we have

$$W_a(\mu) = \frac{1}{\rho}\ln\left(\int_\Theta \exp(u(a,\theta))\mu(\mathrm{d}\theta)\right),$$

for any $a \in A$, any $\mu \in \Delta\Theta$ and any $\rho \in]0,+\infty[$. Moreover, the supremum in Equation (6) is attained uniquely by the probability measure $\eta_a \in \Delta\Theta$ defined by

$$\eta_a(\mu)(\tilde{\Theta}) = \frac{\int_{\tilde{\Theta}} \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta)}{\int_\Theta \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta)},$$

for any Borel set $\tilde{\Theta}$ (see, again, Dupuis and Ellis, 1997, Proposition 1.4.2).

Finally, under some Bayesian posterior $\mu \in \Delta\Theta$, Receiver holds belief $\eta_a(\mu)$ if and only

if $W_a(\mu) \geq W_{a'}(\mu)$ for all $a' \neq a$, that is

$$\int_\Theta \exp(\rho u(a,\theta))\mu(\mathrm{d}\theta) \geq \int_\Theta \exp(\rho u(a',\theta))\mu(\mathrm{d}\theta),$$

for all $a' \neq a$.

In fact, we can extend the result beyond the case of the Kullback-Leibler divergence. Define the φ-divergence between η and μ as

$$\mathrm{D}_\varphi(\eta\|\mu) = \int_\Theta \varphi\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}(\theta)\right)\mu(\mathrm{d}\theta),$$

where $\varphi\colon \mathbb{R} \to [0,+\infty[$ is a proper, closed, convex and essentially smooth function such that $\varphi(1) = 0$ and such that its domain is an interval with endpoints $a < 1 < b$ (which may be finite or infinite). Let us also define the Legendre-Fenchel conjugate of φ, denoted φ^*, by

$$\varphi^*(x') = \max_{x\in\mathbb{R}} xx' - \varphi(x)$$

for any $x' \in \mathbb{R}$. Then, the following proposition holds.

Proposition 6. *Receiver's optimal belief motivated by action a under posterior μ uniquely satisfies*

$$\varphi'\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}(\theta)\right) = \rho u(a,\theta),$$

for any $\theta \in \Theta$, any $a \in A$ and any $\mu \in \Delta\Theta$, while Receiver's optimal psychological well-being equals

$$W_a(\mu) = \frac{1}{\rho}\int_\Theta \varphi^*(\rho u(a,\theta))\mu(\mathrm{d}\theta),$$

for any $a \in A$ and any $\mu \in \Delta\Theta$. Hence, Receiver's optimal belief verifies

$$\varphi'\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}(\theta)\right) = \rho u(a,\theta),$$

whenever $\mu \in \Delta\Theta$ satisfies

$$\int_\Theta \varphi^*(\rho u(a,\theta))\mu(\mathrm{d}\theta) \geq \int_\Theta \varphi^*(\rho u(a',\theta))\mu(\mathrm{d}\theta),$$

for all $a' \neq a$.

Proof. See Broniatowski and Keziou (2006), Theorem 4.3. □

A.2 Proof for Lemma 1

Let us study the properties of the belief threshold μ^W as a function of ρ and payoffs. First of all, let us define the function

$$\mu^W(\rho) = \frac{\exp(\rho\underline{u}_0) - \exp(\rho\underline{u}_1)}{\exp(\rho\underline{u}_0) - \exp(\rho\underline{u}_1) + \exp(\rho\overline{u}_1) - \exp(\rho\overline{u}_0)}.$$

for any $\rho \in]0, +\infty[$. To avoid notational burden, we omit the superscript W in the proof. We can find the limit of $\mu(\rho)$ at 0 by applying l'Hôpital's rule

$$\begin{aligned}\lim_{\rho\to 0}\mu(\rho) &= \lim_{\rho\to 0}\frac{\underline{u}_0\exp(\rho\underline{u}_0) - \underline{u}_1\exp(\rho\underline{u}_1)}{\underline{u}_0\exp(\rho\underline{u}_0) - \underline{u}_1\exp(\rho\underline{u}_1) + \overline{u}_1\exp(\rho\overline{u}_1) - \overline{u}_0\exp(\rho\overline{u}_0)}\\ &= \frac{\underline{u}_0 - \underline{u}_1}{\underline{u}_0 - \underline{u}_1 + \overline{u}_1 - \overline{u}_0}\\ &= \mu^B.\end{aligned}$$

So, we are back to the case of a Bayesian Receiver whenever the cost of distortion becomes infinitely high. After multiplying by $\exp(-\rho\underline{u}_0)$ at the numerator and the denominator of $\mu(\rho)$ we get

$$\mu(\rho) = \frac{1 - \exp(\rho(\underline{u}_1 - \underline{u}_0))}{1 - \exp(\rho(\underline{u}_1 - \underline{u}_0)) + \exp(\rho(\overline{u}_1 - \underline{u}_0)) - \exp(\rho(\overline{u}_0 - \underline{u}_0))}.$$

So the limit of μ^W at infinity only depends on the sign of $\overline{u}_1 - \underline{u}_0$ as, by assumption, $\underline{u}_1 - \underline{u}_0 < 0$ and $\overline{u}_0 - \underline{u}_0 < 0$. Hence, $\lim_{\rho\to+\infty}\mu(\rho) = 1$ when $\overline{u}_1 - \underline{u}_0 < 0$ and $\lim_{\rho\to+\infty}\mu(\rho) = 0$ when $\overline{u}_1 - \underline{u}_0 > 0$. Finally, in the case where $\underline{u}_0 = \overline{u}_1$ we have

$$\begin{aligned}\lim_{\rho\to+\infty}\mu(\rho) &= \lim_{\rho\to+\infty}\frac{1 - \exp(\rho(\underline{u}_1 - \underline{u}_0))}{2 - \exp(\rho(\underline{u}_1 - \underline{u}_0)) - \exp(\rho(\overline{u}_0 - \underline{u}_0))}\\ &= \frac{1}{2}.\end{aligned}$$

Let us now check the variations of the function. After differentiating with respect to ρ and rearranging terms, one can remark that the derivative of $\mu(\rho)$ must verify the following logistic differential equation with varying coefficient

$$\mu'(\rho) = \alpha(\rho)\mu(\rho)(1 - \mu(\rho)),$$

where

$$\alpha(\rho) = \frac{\underline{u}_0 \exp(\rho \underline{u}_0) - \underline{u}_1 \exp(\rho \underline{u}_1)}{\exp(\rho \underline{u}_0) - \exp(\rho \underline{u}_1)} - \frac{\overline{u}_1 \exp(\rho \overline{u}_1) - \overline{u}_0 \exp(\rho \overline{u}_0)}{\exp(\rho \overline{u}_1) - \exp(\rho \overline{u}_0)},$$

for all $\rho \in]0, +\infty[$, together with the initial condition $\mu(0) = \mu^B$. Hence, α completely dictates the variations of $\mu(\rho)$. Let us study the properties of the function α defined on $]0, +\infty[$. First, still applying again l'Hôpital's rule, its limits are given by

$$\begin{aligned} \lim_{\rho \to 0} \alpha(\rho) &= \frac{\underline{u}_0 - \overline{u}_0 - (\overline{u}_1 - \underline{u}_1)}{2} \\ &= \frac{1}{2}(u_0 - u_1) \end{aligned}$$

and

$$\begin{aligned} \lim_{\rho \to +\infty} \alpha(\rho) &= \underline{u}_0 - \overline{u}_1 \\ &= u_{\max}. \end{aligned}$$

Second, after rearranging terms, its derivative is given by

$$\alpha'(\rho) = \frac{(\underline{u}_0 - \underline{u}_1)^2}{\cosh(\rho(\underline{u}_0 - \underline{u}_1)) - 1} - \frac{(\overline{u}_1 - \overline{u}_0)^2}{\cosh(\rho(\overline{u}_1 - \overline{u}_0)) - 1},$$

for any $\rho \in]0, +\infty[$, where cosh is the hyperbolic cosine function defined by

$$\cosh(x) = \frac{e^x + e^{-x}}{2},$$

for any $x \in \mathbb{R}$. Remark that the function defined by

$$f(x) = \frac{x^2}{\cosh(\rho x) - 1} \tag{7}$$

is strictly decreasing on $]0, +\infty[$. So, we have $\alpha'(\rho) < 0$ and therefore μ^W strictly decreasing for all $\rho \in]0, +\infty[$ if and only if $\underline{u}_0 - \underline{u}_1 > \overline{u}_1 - \overline{u}_0$. Accordingly, α is always a strictly monotonic function if and only if $\underline{u}_0 \neq \overline{u}_1$ and $\overline{u}_0 \neq \underline{u}_1$. Hence, excluding the extreme case where $\underline{u}_0 = \overline{u}_1$ and $\overline{u}_0 = \underline{u}_1$ so $\alpha'(\rho) = 0$ and $\mu(\rho) = \mu^B$ for all $\rho \in \mathbb{R}_+^*$, three interesting cases arise, all depicted on Figure 6 for different payoff matrices:

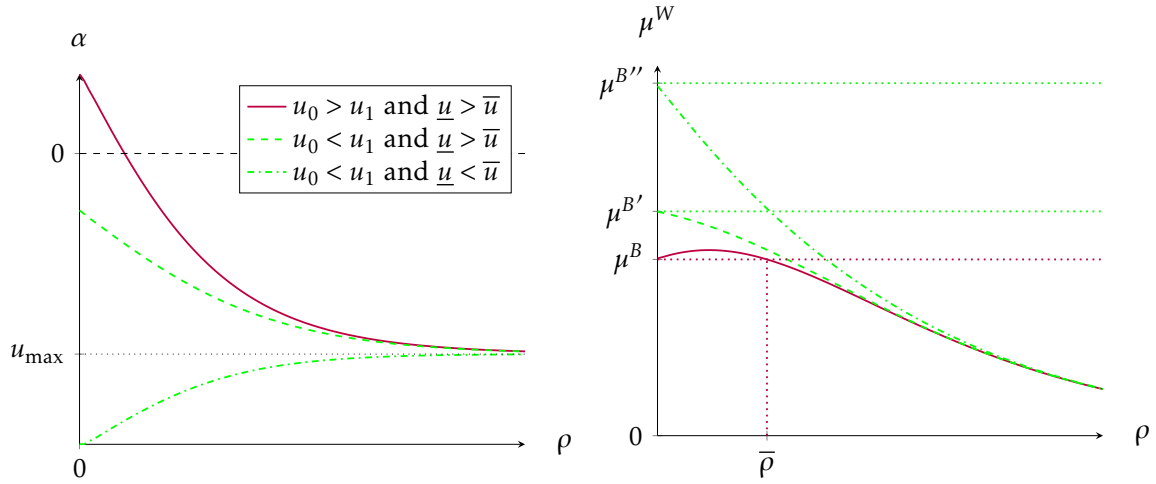
(i) If $u_{\max} < 0$, function α has a constant sign for any $\rho \in]0, +\infty[$ if and only if $u_0 < u_1$, in which case μ^W is strictly decreasing from μ^B to 0. In case $u_0 > u_1$, α has a varying sign

so μ^W starts from μ^B and is sequentially strictly increasing and strictly decreasing toward 0.

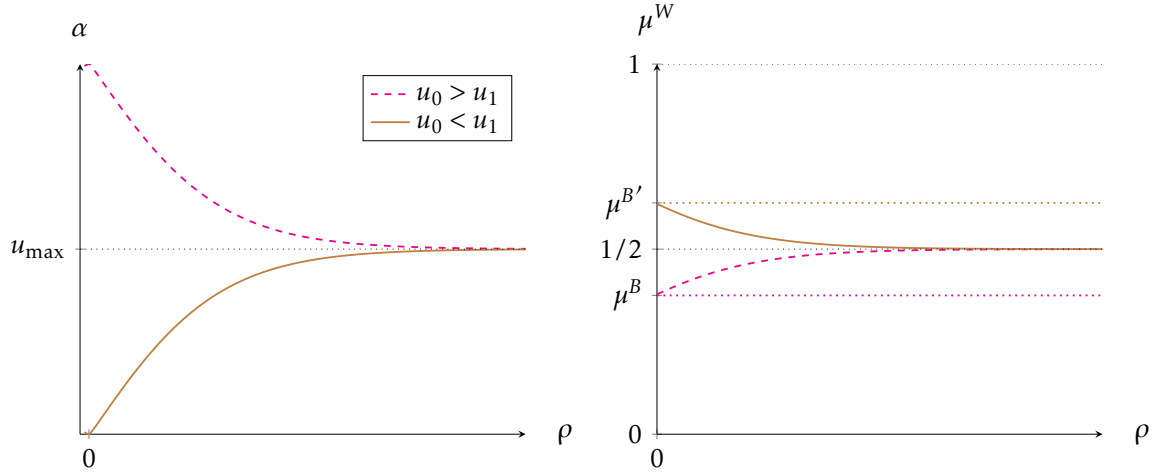
(ii) If $u_{\max} = 0$, function α has a constant sign for any $\rho \in]0, +\infty[$. In this case μ^W is strictly increasing from μ^B to 1/2 if and only if $u_0 > u_1$.

(iii) If $u_{\max} > 0$, function α has a constant sign for any $\rho \in]0, +\infty[$ if and only if $u_0 > u_1$, in which case μ^W is strictly increasing from μ^B to 1. In case $u_0 < u_1$, α has a varying sign so μ^W starts from μ^B and is sequentially strictly decreasing and strictly increasing toward 1.

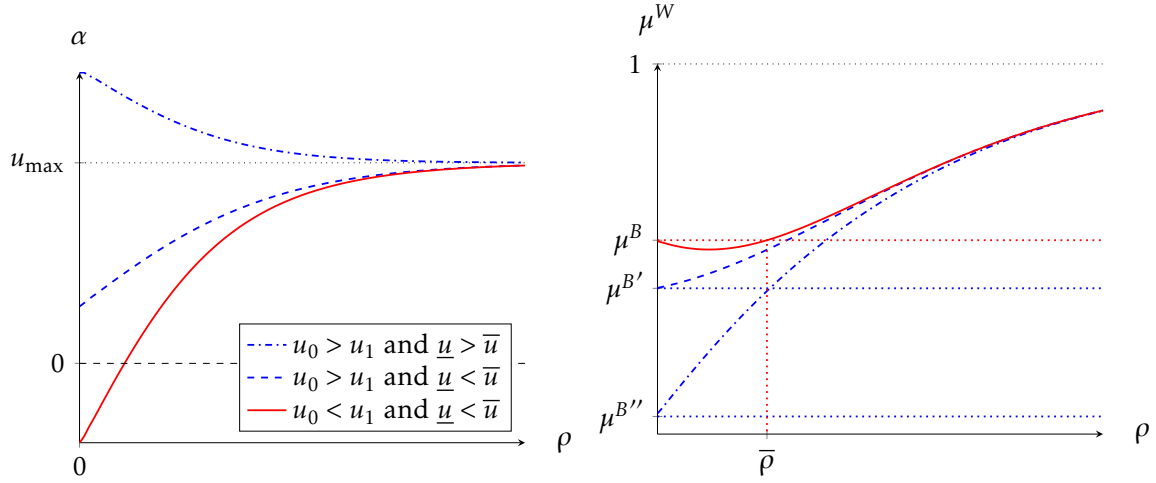
Accordingly, in case μ^W is non-monotonic in ρ, there always exists some $\overline{\rho} > 0$ such that $\mu^W(\overline{\rho}) = \mu^B$. This concludes the proof.



(a) Functions α and μ^W when $u_{\max} < 0$.



(b) Functions α and μ^W when $u_{\max} = 0$.



(c) Functions α and μ^W when $u_{\max} > 0$.

Figure 6: Functions α and μ^W for different payoff matrices $(u_a^\theta)_{a,\theta \in A \times \Theta}$. Action $a = 1$ is favored by a wishful Receiver whenever $\mu^W \leq \mu^B$.

A.3 Proof for Proposition 2

Assume $|\Theta| = n$ where $2 \leq n \leq \infty$. We start by defining the sets

$$\Delta_a^B = \{\mu \in \Delta\Theta : a \in A(\mu)\},$$

and

$$\Delta_a^W = \{\mu \in \Delta\Theta : a \in A(\eta(\mu))\},$$

for any $a \in A$. The set Δ_a^B (resp. Δ_a^W) is the subset of beliefs supporting an action a as optimal for a Bayesian (resp. wishful) Receiver. We say that action a is favored by a wishful Receiver if $\Delta_a^B \subset \Delta_a^W$. We want to show that $\Delta_1^B \subset \Delta_1^W$ if, and only if, the payoff matrix $(u(a,\theta))_{(a,\theta)\in A\times\Theta}$ and the self-deception ability ρ verify at least one of property (i), (ii) or (iii) in Lemma 1 for every pair of states $\theta, \theta' \in \Theta$.

Extreme point representation for Δ_1^B and Δ_1^W. First, remark that Δ_a^B and Δ_a^W are both convex polytopes in $\mathbb{R}^{|\Theta|}$ defined by

$$\Delta_a^B = \Delta\Theta \cap \left\{\mu \in \mathbb{R}^{|\Theta|} : \forall a' \in A, \sum_{\theta\in\Theta} u(a,\theta)\mu(\theta) \geq \sum_{\theta\in\Theta} u(a',\theta)\mu(\theta)\right\},$$

and

$$\Delta_a^W = \Delta\Theta \cap \left\{\mu \in \mathbb{R}^{|\Theta|} : \forall a' \in A, \sum_{\theta\in\Theta} \exp(\rho u(a,\theta))\mu(\theta) \geq \sum_{\theta\in\Theta} \exp(\rho u(a',\theta))\mu(\theta)\right\}.$$

The sets Δ_a^B and Δ_a^B are thus compact and convex sets in $\mathbb{R}^{|\Theta|}$ with finitely many extreme points. Let us now characterize the sets of extreme points of Δ_1^B and Δ_1^W. For any $\mu \in \mathbb{R}^{|\Theta|}$, define the systems of equations

$$\mathbf{A}^B \cdot \mu = \mathbf{b}, \quad \mu \geq 0$$

and

$$\mathbf{A}^W \cdot \mu = \mathbf{b}, \quad \mu \geq 0$$

where

$$\mathbf{A}^B = \begin{pmatrix} u^B(\theta_1) & \dots & u^B(\theta_n) \\ 1 & \dots & 1 \end{pmatrix},$$

and

$$\mathbf{A}^B = \begin{pmatrix} u^W(\theta_1) & \dots & u^W(\theta_n) \\ 1 & \dots & 1 \end{pmatrix},$$

are $2 \times n$ matrices, where $u^B(\theta) = u(1,\theta) - u(0,\theta)$ and $u^W(\theta) = \exp(\rho u(1,\theta)) - \exp(\rho u(0,\theta))$ for any $\theta \in \Theta$, and

$$\mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

In what follows, we always assume that $(u^B(\theta))_{\theta\in\Theta}$ and $(u^W(\theta))_{\theta\in\Theta}$ are such that $\text{rank}(\mathbf{A}^B) = \text{rank}(\mathbf{A}^W) = 2$.[25] Let us recall some mathematical preliminaries.

Definition 1 (Basic feasible solution)**.** *Let $\theta, \theta' \in \Theta$ be any pair of states. A vector μ^* is a basic feasible solution to $\mathbf{A}^B \cdot \mu = \mathbf{b}$ (resp. $\mathbf{A}^W \cdot \mu = \mathbf{b}$), $\mu \geq 0$, for θ, θ' if $\mathbf{A}^B \cdot \mu^* = \mathbf{b}$ (resp. $\mathbf{A}^W \cdot \mu = \mathbf{b}$), $\mu^*(\theta), \mu^*(\theta') > 0$ and $\mu^*(\theta'') = 0$ for any $\theta'' \neq \theta, \theta'$.*

Lemma 2 (Extreme point representation for convex polyhedra)**.** *A vector $\mu \in \mathbb{R}^{|\Theta|}$ is an extreme point of the convex polyhedron Δ_1^B (resp. Δ_1^B) if, and only if μ is a basic feasible solution to $\mathbf{A}^B \cdot \mu = \mathbf{b}$, $\mu \geq 0$ (resp. $\mathbf{A}^W \cdot \mu = \mathbf{b}$, $\mu \geq 0$).*

Proof. See Panik (1993) Theorem 8.4.1. □

Therefore, to find extreme points of Δ_1^B, we just have to solve the system of equations

$$\begin{cases} \mu(\theta)u^B(\theta) + \mu(\theta')b(\theta') = 0 \\ \mu(\theta) + \mu(\theta') = 1 \\ \mu(\theta), \mu(\theta') \geq 0 \end{cases} \tag{8}$$

for any pair of states θ, θ'. When either $\mu(\theta) = 0$ or $\mu(\theta') = 0$, the solution to (8) is given by the Dirac measure δ_θ only if $u^B(\theta) \geq 0$. Denote $\mathcal{E}_1^B$ the set of such beliefs. The set $\mathcal{E}_1^B$ then corresponds to the set of degenerate beliefs under which a Bayesian Receiver would take action $a = 1$. Now, if $\mu(\theta), \mu(\theta') > 0$ then the solution to (8) is given by

$$\mu^B_{\theta,\theta'} = \frac{u(0,\theta') - u(1,\theta')}{u(0,\theta') - u(1,\theta') + u(0,\theta) - u(1,\theta)}.$$

Such a belief is exactly the belief on the edge of the simplex between δ_θ and $\delta_{\theta'}$ at which a Bayesian decision-maker is indifferent between action $a = 0$ and $a = 1$. Denote $\mathcal{I}^B$ the set

[25] This amounts to assuming that payoff are not constant across states.

of such beliefs. Hence, we have

$$\text{ext}(\Delta_1^B) = \mathcal{E}_1^B \cup \mathcal{I}^B.$$

Following the same procedure, the set of extreme points of Δ_1^W is given by $\mathcal{E}_1^W \cup \mathcal{I}^W$, where $\mathcal{E}_1^W$ is the set of degenerate beliefs at which $u^W(\theta) \geq 0$ and $\mathcal{I}^W$ is the set of beliefs

$$\mu^W_{\theta,\theta'}(\rho) = \frac{\exp(\rho u(0,\theta')) - \exp(\rho u(1,\theta'))}{\exp(\rho u(0,\theta')) - \exp(\rho u(1,\theta')) + \exp(\rho u(0,\theta)) - \exp(\rho u(1,\theta))},$$

for any $\theta, \theta' \in \Theta$. Now, applying Krein-Milman theorem, we can state that

$$\Delta_1^B = \text{co}\left(\mathcal{E}_1^B \cup \mathcal{I}^B\right)$$

and

$$\Delta_1^W = \text{co}\left(\mathcal{E}_1^W \cup \mathcal{I}^W\right)$$

Sufficiency. Assume the payoff matrix $(u(a,\theta))_{(a,\theta)\in A\times\Theta}$ and the self-deception ability ρ verify at least one of property (i), (ii) or (iii) in Lemma 1 for every pair of states $\theta, \theta' \in \Theta$. Therefore, we have $\mu^W_{\theta,\theta'}(\rho) > \mu^B_{\theta,\theta'}$ for any $\theta, \theta' \in \Theta$. This implies $\mathcal{I}_1^B \subset \Delta_1^W$, since action $a = 1$ is favored by a wishful Receiver on each edge of the simplex. Moreover, it is trivially satisfied that $\mathcal{E}_1^B = \mathcal{E}_1^W$. Hence, since any point in Δ_1^B can be written as a convex combination of points in $\mathcal{E}_1^B \cup \mathcal{I}^B \subset \Delta^W$, it follows that $\Delta_1^B \subset \Delta_1^W$.

Necessity. Assume now that $\Delta_1^B \subset \Delta_1^W$. Therefore, we have $\mu^W_{\theta,\theta'}(\rho) > \mu^B_{\theta,\theta'}$ for any $\theta, \theta' \in \Theta$ which implies that $(u(a,\theta))_{(a,\theta)\in A\times\Theta}$ and the self-deception ability ρ verify at least one of property (i), (ii) or (iii) in Lemma 1 for every pair of states $\theta, \theta' \in \Theta$.

A.4 Proof for Proposition 3

First, let's remark that the terms of the sum in Equation (4) can be rearranged in the following way:

$$\begin{aligned}\pi(\mu) = &\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \eta(\mu,\beta^i) - \eta(\mu,\beta^j) \\ =&(n-1)\eta^1(\mu) + (n-2)\eta^2(\mu) - \eta^2(\mu) + \\ &\cdots + \frac{n-1}{2}\eta(\mu,\beta^m) - \frac{n-1}{2}\eta(\mu,\beta^m) + \cdots + \\ &\eta(\mu,\beta^{n-1}) - (n-2)\eta(\mu,\beta^{n-1}) - (n-1)\eta^n(\mu) \\ =&\sum_{i=1}^{m}(n+1-2i)(\eta(\mu,\beta^i) - \eta(\mu,\beta^{n+1-i})),\end{aligned}$$

for any $\mu \in [0,1]$, where $m = (n+1)/2$. That is, we can express it in terms of the differences in beliefs among voters who are equidistant from the median. To see that this is true, we need to first realize that each belief will appear $n-1$ times in Equation (4) (since each belief will be paired once with each of the other $n-1$ beliefs). The beliefs of voters below the median will more often than not appear as positive (the belief of the first voter will be positive in all of its pairings, the belief of the second voter will be positive in all of its pairing except for the pairing with the first voter, etc), whereas the beliefs of voters above the median will be negative more often than not. If we rearrange the terms of the sum in order to pair symmetric voters, the term $(\eta(\mu,\beta^1) - \eta_n(\mu))$ will appear $n-1$ times, whereas the term $(\eta_2(\mu) - \eta(\mu,\beta^{n-1}))$ will appear $n-3$ times, since out of the $n-1$ times $\eta_2(\mu)$ appears on Equation (4), $n-2$ of them will be positive and 1 will be negative (the converse is true for $\eta(\mu,\beta^{n-1})$). You can continue the same reasoning for all the pairs of symmetric voters, and get to the formulation of $\pi(\mu)$ presented above. Note, also, that the belief of the median voter will be summed and subtracted at the same rate, such that he does not matter in our measure of polarization.

Consider the distance between beliefs of any pair of symmetric voters $\eta(\mu,\beta^i) - \eta(\mu,\beta^{n+1-i})$ for $i \in \{1,\ldots,m\}$. Given our symmetry assumption these two agents are symmetric, such that $\beta^i = 1-\beta^{n+1-i}$. It is not difficult to show that any of those pairwise distances is maximized when agent i is distorting its belief upwards and agent $n+1-i$ is distorting its belief downwards. That is, when $\mu \in [\mu^W(\beta^i), \mu^W(\beta^{n+1-i})]$.

First, the distance between symmetric beliefs in such an interval can be rewritten as

$$\eta(\mu,\beta^i)-\eta(\mu,\beta^{n+1-i})=\frac{\mu\exp(\rho\beta^i)}{\mu\exp(\rho\beta^i)+(1-\mu)}-\frac{\mu}{\mu+(1-\mu)\exp(\rho\beta^i)}.$$

for any $i\in\{1,\dots,m\}$ and $\mu\in[\mu^W(\beta^i),\mu^W(\beta^{n+1-i})]$.

Second, by taking the first order condition in this interval and rearranging it we get

$$\frac{\mu+(1-\mu)\exp(\rho\beta^i)}{\mu\exp(\rho\beta^i)+(1-\mu)}=1,$$

such that the difference between symmetric beliefs is maximized uniquely at

$$\mu=\mu^W(\beta^m)=\frac{1}{2},$$

for any $i\in\{1,\dots,m\}$, $\beta^i\in\,]0,1[$ and any $\rho\in\,]0,+\infty[$. Since

$$\mu^W(\beta^m)=\arg\max_{\mu\in[0,1]}\eta(\mu,\beta^i)-\eta(\mu,\beta^{n+1-i})$$

for any $i\in\{1,\dots,m\}$, we get

$$\mu^W(\beta^m)=\arg\max_{\mu\in[0,1]}\pi(\mu),$$

which concludes the proof.

A.5 Proof for Proposition 5

First, we define the function

$$\psi(z)=\frac{1}{1-F(z)}\int_z^{\overline{\theta}}\exp(\rho\theta)f(\theta)\mathrm{d}\theta,$$

for any $z\in[\underline{\theta},\overline{\theta}[$ and adopt the convention that $\psi(\overline{\theta})=\exp(\rho\overline{\theta})$. It is not difficult to show that ψ is a continuous and strictly increasing function from $\psi(\underline{\theta})=\hat{x}<1$ to $\psi(\overline{\theta})=\exp(\rho\overline{\theta})$. Define similarly the function

$$\varphi(z)=\frac{1}{1-F(z)}\int_z^{\overline{\theta}}\theta f(\theta)\mathrm{d}\theta,$$

for any $z \in [\underline{\theta}, \overline{\theta}[$ and $\varphi(\overline{\theta}) = \overline{\theta}$. Again, it is not difficult to show that φ is a continuous and strictly increasing function from $\varphi(\underline{\theta}) = \hat{m} < 0$ to $\varphi(\overline{\theta}) = \overline{\theta}$.

Since ψ is strictly increasing, it thus suffices to show that $\psi(\theta^B) > 1 (= \psi(\theta^W))$ to prove that $\theta^W < \theta^B$. Applying Jensen's inequality, it follows that

$$\psi(z) > \exp(\rho\varphi(z)),$$

for any $z \in]\underline{\theta}, \overline{\theta}[$, where the strict inequality comes from the strict convexity of $z \mapsto \exp(\rho z)$ and the non degeneracy of F. In particular, Jensen's inequality holds with equality at $\underline{\theta}$ and $\overline{\theta}$, but applying the intermediate value theorem, it is not difficult to show that θ^B (as well as θ^W) must lie in the open interval $]\underline{\theta}, 0[$. Thus, we have in particular

$$\psi(\theta^B) > 1,$$

since $\varphi(\theta^B) = 0$ and $\theta^B \neq \underline{\theta}, \overline{\theta}$.