Universidad de San Andrés
Departamento de Economía
Maestría en Economía

Interpretive Evaluation: Effects of Confirmation Bias on the Retribution to Talent

Autor: **Marcelo Denis WOO**
32.784.045

Mentor: **Christian RUZZIER**

Victoria, Buenos Aires
25 de agosto, 2019

Tesis de Maestría en Economía de

Marcelo Denis WOO

"Evaluación interpretativa: efectos del sesgo de confirmación en la retribución al talento"

Resumen

Este trabajo estudia los efectos del sesgo de confirmación en la retribución a talento inobservable en mercados de trabajo competitivos. Bajo evaluación con sesgo de confirmación, el candidato siempre ejerce esfuerzo positivo para influir las señales, pero el esfuerzo decrece en el tiempo, convergiendo a cero. Mientras que creencias bayesianas convergen a tasa lineal t al talento, creencias interpretativas convergen a tasa exponencial 2^t a una media ponderada entre el talento y el prejuicio inicial. Si las creencias iniciales están sesgadas, entonces para cualquier precisión de prior $h_0 > 0$, salarios interpretativos no convergen a la productividad verdadera, y el sesgo en retribución es persistente, incluso con signalling óptimo ad infinitum del candidato. De este modo, el sesgo de confirmación se convierte en una nueva fuente de falla de mercado, de naturaleza persistente. El sesgo en retribución crece en el sesgo del prejuicio inicial y en la precisión relativa h_0/h_ε. Para el caso en que el mercado condiciona creencias iniciales en características observables (género, etnicidad), analizamos diversas métricas de sesgos retributivos relativos al talento. El gap inter-individual de salarios, usado comúnmente en discusiones públicas, refleja tanto un gap de prejuicios inter-grupal como también un gap de talentos inter-individual, siendo así una medida con confounding e inconclusiva sobre sesgos grupales. Una medida más apropiada de diferencias en prejuicios inter-grupales es el gap de sesgos en retribución, al ser ortogonal a -y controlar por- el gap de talentos inter-individual.

Palabras clave: [economía de la información, economía conductual, sesgo de confirmación, aprendizaje, discriminación, economía laboral]

"Interpretive Evaluation: Effects of Confirmation Bias on the Retribution to Talent"

Abstract

This thesis studies the effects of confirmation bias on the retribution to unobservable talent in a competitive labor market. Under evaluation with confirmatory-bias, the candidate always exerts a positive level of effort to influence information, but effort decreases over time, converging to zero. While Bayesian beliefs converge at linear rate t to the true talent, confirmatory-biased beliefs converge at an exponential rate 2^t to a weighted average between talent and the initial prejudice. If initial beliefs are biased, then for any prior precision $h_0 > 0$, confirmation-biased wages never converge to the talent, so the Retribution Bias is *persistent, even with ad infinitum optimal signalling by the candidate. Thus, confirmation bias becomes a new source of market inefficiency, of persistent nature. The Retribution Bias increases in the initial prejudice gap and in the relative prior-to-signal precision h_0/h_ε. For the case when the market conditions initial beliefs on observable characteristics (e.g., gender, ethnicity), we analyze different measures of bias in retribution relative to individual talent. The inter-individual wage gap, commonly used in public discussion, reflects both an inter-group prejudice gap and an inter-individual talent gap, and therefore is a confounding and inconclusive measure of group-based biases. A more appropriate measure of the inter-group prejudice gap is the Retribution Bias Gap, since it is orthogonal to -and therefore controls for- the inter-individual talent gap.*

Keywords: [Information Economics, Behavioral Economics, Confirmation Bias, Learning, Discrimination, Labor Economics]

Códigos JEL: [D83, D86, D87, D9, D91, J15, J16, J7]

Acknowledgements

I am grateful to those who have been important in writing this thesis.

To my **family and friends**, who have been most supportive and patient. I hope that the relevance of this thesis' topic, and the potential contribution to our societies serve as a partial retribution for their kind understanding and patience during the last months.

To my colleagues at **Samsung Electronics**, in Argentina and in South Korea. It has been a priviledge to work over the last years in this leading international tech company, where I could appreciate how global-scale organizations function from the inside and observe much of empirical human behavior in organizations. Particularly, given its bi-cultural setting, and my own dual cultural background, I could appreciate from both Argentinean and Korean perspectives how culturally conditioned priors can lead to different interpretations of information and hence different conclusions, and cultural blindspots as well, which triggered my fascination for Social Psychology. It was the subsequent exploration of this discipline that, in turn, inspired the present work.

I express my sincere gratitude to professor **Christian Ruzzier**, who has generously accepted to mentor my thesis, even after my 9 years in industry, out of academia. He has been most kind in reviewing my model's drafts and supporting my interest on researching in this still niche area of the Economics discipline. Even when the literature on this topic is not especially abundant, his guidance has been crucial in providing connections for the right framework for my ideas. This work would have had a greatly diminished impact without his valuable comments. Needless to say, all remaining errors, if any, are entirely mine.

To my **uncle Jeong**, for providing a focused environment where my initial ideas could develop further than I could imagine. This thesis would have not matured this far had he not offered his study room generously for my usage.

To **Diana Shim**, my early mentor in life. Beliefs are path dependent, not only in the sequence of information we get exposed to, but also in the sequence of people and the worldviews, experiences, and values they embody. Just as early information can influence long term beliefs, her early presence has left an indelible mark shaping my values. Even as I moved the center of gravity of my life towards further academic and professional fields and novel personal interest areas, these later experiences in my life have not erased her early, continuing influence.

1 Introduction

"The human understanding, when it has once adopted an opinion, draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects."

—Francis Bacon

1.1 Motivation

Confirmation bias, as is related to human learning processes, seems to be ubiquitous in every human domain involving learning, updating beliefs, making inference, evaluation, and judgement, and many social sciences have studied its effect on their own fields, ranging from Social Psychology to International Relations, Political Science, and Economics. In this section, we provide a necessarily brief summary, hoping to give the reader an idea of the breadth and relevance of this phenomenon, as it has been acknowledged by several disciplines.

The empirical observation that human beings tend to draw conclusions in a way that seeks to confirm their pre-existing beliefs has been documented as early as the fifth century B.C, when the Greek historian Thucydides, in his account of the war between Sparta and Athens in *The History of the Peloponnesian War*, famously stated: "For it is a habit of mankind to entrust to careless hope what they long for, and to use sovereign reason to thrust aside what they do not fancy".

Among modern academic disciplines, the field of **Social Psychology** has been the first to systematically point out to human cognitive biases. In particular, for confirmation bias, the earliest work is Wason (1960), which showed evidence that subjects who have an hypothesis in mind tend to seek only confirming evidence when drawing conclusions about simple tasks.

Thereafter, an abundant body of work has been developed; because of space constraints we provide the most canonical here. One of the foundational works is Lord, Ross, and Lepper (1979), which showed that people who hold strong views about complex social issues are prone to consider empirical evidence in a biased way. They are likely "to accept 'confirming' evidence at face value while subjecting 'disconfirming' evidence to critical evaluation, and as a result to draw undue support for their initial positions from mixed or random empirical findings. Thus, the result of exposing contending factions in a social dispute to an *identical* body of relevant empirical evidence may not be a narrowing of disagreement but rather an *increase in polarization*". In their experiment, people favoring and opposing capital punishment (as a particular and contemporary instance of a complex social issue) were given two studies, one confirming and one disconfirming their existing beliefs about the deterrent efficacy of capital punishment. Both proponents and opponents rated the results and methodologies that confirmed their beliefs as more convincing and valid, and shifted their attitudes correspondingly. Thus, the study showed that the exposition to even the *same* body of evidence can lead to a *polarization* of beliefs.

Anderson, Ross, and Lepper (1980) showed that social theories (which are related to beliefs about social issues) are *perseverant*, as they can survive the total discrediting of the initial evidence that gave origin to them. In their study, individuals were given two case studies suggestive of either a positive or a negative relationship between risk taking and success as a firefighter, after which they were thoroughly debriefed about the ficticious nature of the case studies. Subsequent evaluation of the subjects' personal beliefs revealed that *even when initially based on weak data, social theories can survive the total discrediting of the initial evidential base.*

Especially relevant to our work is Darley and Gross (1983), as it showed how *schemas*[1] *or stereotypes about social class can influence the perception of performance.* In particular, they studied "the process leading to the confirmation of a perceiver's expectancies[expectations] about another when the social label that created the expectancy provides poor or tentative evidence about another's true dispositions or capabilities". In their experiment, two groups of subjects were first given a picture and some information about a fourth-grade girl student named Hannah. To induce a stereotype about her social class, Hannah was pictured in front of a nice suburban house for one-half of the participants and pictured in front of an impoverished house in an urban area for the other half. All subjects then watched a video that showed Hannah taking an intelligence test, where she got some questions right and some wrong. The number of right and wrong questions were the same for both conditions. The participants who thought that Hannah had come from an upper-class background remembered that she had gotten more correct answers than those who thought she was from a lower-class background. This study then shows that *stereotypes can influence memory and the perception of performance.*

[1] A pattern of thought or behavior that organizes categories of information and the relationships among them.

In the field of **International Relations**, the school of thought of Realism[2] asserts that under the anarchic nature of relations among sovereign states that have offensive, possibly lethal power, the uncertainty about other actor's intentions become a crucial driving force that leads states to a self-enforcing, perennial state of war. The possibility that other states may conceal potentially lethal intentions leads states to a preparation for war, which combined with their very incapacity to effectively signal unobservable intentions sets the stage for a constant threat and fear of war[3]. In this continual high-stakes environment, cognitive limitations have been studied as a potential determinant of foreign policy decision-making. Since foreign policy leaders are responsible for vital, pressing decisions on the basis of imprecise information, it is posited that they reduce the complexity and make sense of the ambiguity through their *images* ("organized representations of the world") and their *belief systems* (worldviews that contain "beliefs, explanations, hypotheses, feelings, predispositions, attitudes, and so on"). The effect is that "political leaders act on their individual images and *perceptions* of the world rather than on objective reality". However, such images and belief systems have been acknowledged to be *resistant to change* for a number of reasons, including an inner striving for cognitive consistency among the components of the belief system (so that any major change in the images and elements of that system can be avoided), and a tendency to minimize cognitive dissonances by the assimilation of information and by selective filtering. (Cashman, 1980).

More recently, the phenomenon of confirmation bias has been brought to prominence in the field of **Political Science** over the past decade, as one potential source of the observed polarization of electorates. In addition to the fact pointed out by Psychological studies that individuals tend to exacerbate differences in their priors and polarize their beliefs even when they are exposed to the *same* body of information, the recent availability of individually optimized search engines in the internet, and the free choice of association in social media has brought a new source of confirmation bias. In the current era of information overload, as individuals (or equivalently, their associated algorithms) choose which information they get to encounter and their relative frequency, now the problem of confirmation bias is even aggravated, given that individuals with different priors get different pieces of information, which are more likely to be consistent with their preexisting views rather than inconsistent. Thus, modern individuals, by filtering the excess of information in non-representative ways, end up creating and entering into the so called "echo chambers", where they mostly hear and share voices and views that are similar to their own. This promotion and overrepresentation of "consistent information" with respect to one's beliefs exacerbates the polarizing phenomenon already noted by Lord, Ross, and Lepper (1979).

Another trend in the political world has been pointed out by political scientist Francis Fukuyama. Since the emergence of the World Wide Web in the

[2] Realism, incidentally, embraces the aforementioned *The History of the Peloponessian War* by Thucydides as one of its classic, foundational texts

[3] See Mearsheimer (2001)

1990s, the production and transmission of information has become extremely cheap, so that it is now more widespread and accessible than ever. The development of social media in the 2000s furthered this momentum, and featured the rapid spread of user-generated information, and allowed the mass coordination of the "color revolutions" around the world. However, some governments have in response exploited the potential for disseminating (mis)information in highly accessed digital platforms by creating and promoting what have been identified as "fake news" to sway the constituencies' mental images, beliefs, and preferences in the policy space[4]. Given the proliferation of information whose veracity or falsehood is hard and costly to prove, every piece of information can now be politicized, leading to a decay of trust and an increase of ambiguity[5]. In short, both recent trends (the biased reduction of excessive information and the propagation of mis-information) have contributed to the polarization of worldviews in the political constituencies' beliefs about social, political, and economic issues.

In **Economics**, however, given its classical emphasis on the assumption of rationality of economic agents, attention on the role of confirmation biases has been mostly deferred until the appearance of Behavioral Economics, which highlights human cognitive limitations instead. Nonetheless, a notable exception is Schrag and Rabin (1999), which by 1999 modeled confirmation bias and provided a mathematical foundation for the polarization and divergence of beliefs, even in the presence of infinite information. In their words, "a person suffers from confirmatory bias if he tends to *misinterpret ambiguous evidence* as confirming his current hypothesis about the world. (...) Many people misread their observations of individual behavior as supporting their prior stereotypes about groups to which these individuals belong". The model also confirms an intuition from Psychology literature: "confirmatory bias leads to *overconfidence*, in the sense that people in average believe more strongly than they should in their favored hypothesis". They also show that "even an infinite amount of information does not necessarily overcome the effects of confirmatory bias: over time an agent may with positive probability come to believe with near certainty in the wrong hypothesis."

A recent paper from Fryers, Harms, and Jackson (2013) expands the work of Schrag and Rabin (1999), in at least two directions. First, the model introduces a human memory storage limitation. Information that is *ambiguous* about the state of the nature, *ab*, can only be stored as a single bit of information, either *a* or *b*. Thus, this model sets a foundation for the interpretation and storage of ambiguous information "that can be thought as providing a 'why' behind long term bias, and how this can also lead to belief polarization". Second, they develop a model of confirmation bias for the case of continuous state of nature

[4] More recently, the development of "deep fakes" (AI-based generation of new digital images and videos based on a minimal amount of training data, that can effectively mimic original content) will likely add a new source of noise to this phenomenon.

[5] In Economics language, signals are becoming noisier, with higher variance σ_{ε}^2, and as it will become clear later in this work, if learning is confirmation-biased, this kind of environment sets a condition in which initial beliefs (or "prejudices") become more prominent in the learning process and, importantly, in the location of the final belief.

and signal distributions, which “provide some new results showing that bias always occurs (with probability one), and shows how it depends on early signals and not just the prior”.

In spite of these developments in Economics, and in spite of the ubiquity of studies on the effects of confirmation bias in other social domains, as far as the author knows it seems that up to date no study in Economics has linked the phenomenon of confirmation bias to how labor markets retribute individuals of varying levels of productivity and belonging to different social categories, over which society (and hence the market) might have differentiated priors (or prejudices). A study of this topic seems to be the more urgent and relevant given the contemporaneous polarization of worldviews in the political constituencies throughout the world, propelled by the biased selection of information (in the form of individually customized bubble filtering algorithms and our self-selection into our own “echo chambers” in social media and news feeds) and by the proliferation of mis-information (in the form of fake news and deep fakes). Looking ahead, there are concerns over the degree in which historical human-induced biases could be promoted by the adoption and deployment of AI-based learning systems across industries, governments, and healthcare. If we human beings are as vulnerable to confirmation bias as Thucydides noted back in his age, as Social Psychologists have been documenting ever since the 1960s, and as Political Scientists have been intensely debating over the last decade to the present day, we ought at least ponder what might be the effects of this potentially pervasive, often unconscious cognitive bias, and how it might lead to both inefficiencies and unfairness in the marketplace. This work is offered as a small contribution to that enterprise.

1.2 Our model

In this paper, we aim to study the effects of confirmation bias on the retribution to talent. To do so, we start from the interpretive learning model of Fryer, Harms and Jackson (2013), in which a confirmation-biased individual interprets a stream of exogenous, independent, and normally distributed information and then updates her beliefs based on her interpretations. We extend this model by adding an evaluated individual (a “candidate”) whose payment depends on the beliefs of the confirmation-biased evaluator, and who can take actions to influence the information she gets to see, introducing thereby a strategical dimension to the problem of confirmatory biased learning. Given that we are interested in the effects of confirmation bias on the retribution to unobservable talent, and the inherently dynamic nature of the process of learning and influencing, we analyze this problem within the framework of the reputation formation model of Holmstrom (1999). In our extended model, the “candidate” (i.e., a “manager” or an “applicant” for a job or promotion) has an unobservable characteristic η that is related to his productivity (e.g., his talent or his ability) which is not known to both the evaluator and the candidate. When the evaluator gets a signal about η and updates her beliefs about the distribution of η to decide the candidate’s payment, she suffers from a confirmation bias, which mediates -possibly unconsciously- her learning process, in such a way that tends to interpret signals about talent as if they were closer to her pre-existing beliefs.

Given that the candidate faces an interpretive (i.e., confirmatory-biased) evaluator and that he has a chance to influence signals, what is his best strategy and how does it differ from the case of an unbiased (Bayesian) evaluator? How does the equilibrium belief reached by an interpretive evaluator differ from that of a Bayesian one? How much bias in the retribution to talent, if any, can we expect when confirmation biases affect the evaluation process based on information that is influenced by the candidate? Finally, can the candidate overcome the (potentially unfair) effects of confirmation bias, if he were given enough opportunities to act and optimally influence information?

We first highlight that confirmation-biased (interpretive) learning features overweight of initial beliefs, overconfidence, and preference for early information -so the order in which signals are received is important, even when they are statistically independent. We find that just like under Bayesian evaluation, the optimal sequence of effort of a candidate under confirmatory biased evaluation is strictly decreasing and converges to zero. However, in the latter case the convergence is attained at an exponential rate 2^t, which is faster than the Bayesian linear rate t. While with enough information Bayesian wages converge to the true talent, the interpretive wage usually does not, and converges instead to a weighted average between the talent and the initial prior mean m_0. Therefore, confirmation bias becomes a new source of inefficiency in the marketplace, of permanent nature, even with the candidate's optimal strategy played infinite times. The long-run bias in the retribution with respect to talent is determined by the distance between the talent and the initial prejudice, and the relative prior-to-signal precision.

When the evaluator conditions her initial beliefs on social groups, and can observe the social group a candidate belongs to, differences in long-run wages between two given individuals can arise from 2 sources: inter-individual differences in talent, and inter-group differences in prejudices. Therefore, the inter-individual wage gap is an inconclusive and confounding measure of prejudices, since it does not control for inter-individual talent gaps. As an alternative measure, the Retribution Bias Gap, defined as the difference between the retribution biases of two individuals, which is therefore orthogonal to the talent dimension, is proposed theoretically as a measure of the long-run effects of group-based prejudices.

This work is outlined as follows. In Section 2, we make a review of Holmstrom (1999) reputation building and career concerns model. By linking dynamic learning about unobserved talent to the market equilibrium wage, it constitutes a natural place to start. In addition, as it features Bayesian learning, it sets a benchmark to which the case of confirmatory biased learning can be then compared. In Section 3, we unpack the continuous version of the confirmation biased learning model of Fryer, Harms, and Jackson (2013). Since it is likely to be unknown by many readers, a thorough presentation is provided. Unlike FJH, we express the process in terms of precisions rather than variances, which notably simplifies the math and sharpens economic intuition. We also contribute by providing foundations from previous Social Psychology literature to features of this model. In Section 4, we incorporate this confirmation-biased learner to Holmstrom(1999) model and analyze its implications for the equilibrium wage. Since we are interested on whether sufficiently high amounts of signalling may lead the evaluators to the true individual talent, we provide a convergence analysis and see how the asymptotic retribution in confirmatory biased markets differ from a Bayesian-learning market. In Section 5, we turn to the question of how differing group-based prejudices lead to varying retribution outcomes. We let the market have initial beliefs conditional on an observable attribute (which may be just a social category the candidate or worker belongs to). We evaluate the appropriateness of different measures of Bias in Retribution by analyzing how they are indicative of inter-individual talent gaps and inter-group prejudice gaps. Section 6 we provide a conclusion with main findings, point out to limitations of this work, and suggest future research agenda.

2 Evaluation and Retribution under Bayesian Learning

To study the problem of the evaluation and retribution to talent, we start with the reputation building model of Holmstrom (1999). As it considers a job market that learns and sets expectations through rational, Bayesian inference, the model sets a baseline to which the results of the confirmatory biased learning can be compared. In this section we provide a quick review of the model.

We consider a candidate (e.g., a manager) who works in a competitive labor market and who is endowed with labor, which he can exchange for a wage. In this market, no output-contingent contracts are available. However, since the wage in each period is based on the expected output, and the expected output depends on the assessed ability, an "implicit contract" takes place, linking today's performance to future wages through the updating of beliefs about ability. The wage is first set by the market, and then the candidate chooses his level of effort. Given the unavailability of output-contingent contracts, however, we can think that at each period the candidate is paid in advance for his output. To create incentives for labor, we consider a multi-period interaction, where the labor market has uncertainty about the candidate's unobservable characteristics. In this context, present performance serves as information about future performance.

Assume there is a valid measure η of a (fixed) characteristic of the candidate that is incompletely known to the candidate and the market and that is related to his productivity (e.g., his talent or ability). Both the candidate and the evaluator have initial beliefs about η to be normally distributed with mean m_0 and precision $h_0 \equiv 1/\sigma_0^2$. Over time, the evaluator will learn about η through the observation of the candidate's output, which at any time t is given by:

$$y_t = \eta + a_t + \varepsilon_t, \qquad t = 1, 2, 3, ... \tag{1}$$

where $a_t \in [0, \infty)$ is the candidate's effort and ε_t is a stochastic noise term which randomly affects the output. To learn about η through y_t, it will be necessary to know the distribution of ε_t. Assume that ε_t's are independent and identically distributed with $\varepsilon_t \sim \mathcal{N}(0, 1/h_\varepsilon)$.

The candidate has risk-neutral preferences given by:

$$U(c, a) = \sum_{t=1}^{\infty} \beta^{t-1}[c_t - g(a_t)] \tag{2}$$

where $g(.)$ is an increasing and convex function that measures the disutility of effort. This utility function is assumed to be publicly known.

Let $y^t \equiv (y_1, y_2, ..., y_t)$ denote the history of outputs up to time t, also assumed to be publicly known. Wages in period t will be set as a function $w_t(y^{t-1})$ of history, as well as the candidate's effort $a_t(y^{t-1})$.

A risk-neutral evaluator in a competitive market will set a wage equal to her expectation of output, given the history of outputs up to $t-1$. Because in markets with no output-contingent contracts wages are set before the choice of effort, no conditioning on y_t is possible, since history is available only up to $t-1$.

$$\begin{aligned} w_t(y^{t-1}) &= E[y_t \,|\, y^{t-1}] \\ &= E[\eta \,|\, y^{t-1}] + a_t(y^{t-1}) \end{aligned} \tag{3}$$

Note that the competitive market wage is determined by the present belief about the candidate's talent η and the candidate's decision rule.

On the other hand, the candidate solves his utility maximization problem to derive his decision rule $a_t^*(y^{t-1})$:

$$\underset{a_t}{Max} \sum_{t=1}^{\infty} \beta^{t-1}\{E[w_t(y^{t-1})] - E[g(a_t(y^{t-1}))]\} \tag{4}$$

The evaluator's learning about the unobservable talent η occurs as follows. Although the candidate's action a_t is not directly observable by the evaluator, she can however derive it by solving the candidate's decision problem (4). Therefore, by observing y_t the evaluator can equivalently observe z_t, given by:

$$z_t \equiv y_t - a_t^*(y^{t-1}) = \eta + \varepsilon_t \tag{5}$$

where $a_t^*(y^{t-1})$ denotes the candidate's decision rule.

In other words, by deriving the candidate's decision rule, the evaluator is able to observe a noisy measure of the talent, which (given the distribution of ε) will be distributed as $z_t \sim \mathcal{N}(\eta, 1/h_\varepsilon)$. Based on the observation of z_t, the evaluator learns about η through Bayesian inference. Since both the initial prior belief about η and the signal z_t are normally distributed, the posterior distribution is also normal, with means and precisions given respectively by:

$$
\begin{aligned}
m_t &= m_{t-1}\left(\frac{h_{t-1}}{h_{t-1}+h_\varepsilon}\right) + z_t\left(\frac{h_\varepsilon}{h_{t-1}+h_\varepsilon}\right) \\
&= m_0\left(\frac{h_0}{h_0+t\,h_\varepsilon}\right) + \left(\frac{h_\varepsilon}{h_0+t\,h_\varepsilon}\right)\sum_{j=1}^{t} z_j \qquad (6)
\end{aligned}
$$

$$
\begin{aligned}
h_t &= h_{t-1} + h_\varepsilon \\
&= h_0 + t\,h_\varepsilon \qquad (7)
\end{aligned}
$$

Thus, for any t, the Bayesian posterior mean is essentially a *weighted average* of the initial prior mean m_0 and all the available information set z^{t-1}. The weight $\left(\frac{h_0}{h_0+t\,h_\varepsilon}\right)$ placed on the initial prior mean m_0, increases in the prior precision h_0, so when prior beliefs are deeply held the initial prior mean has a higher importance in the determination of the posterior belief for any given t. On the other hand, each individual signal $z_j, j = 1, 2, ..., t$ receives a weight of $\left(\frac{h_\varepsilon}{h_0+t\,h_\varepsilon}\right)$, so when the information variance is lower, information receives higher importance in the belief updating process. It is also a feature of Bayesian inference that every signal is given a uniform weight, and that therefore the order by which signals arrive is irrelevant, a property known as *exchangeability*. As time t increases (and more information is received) the weights of both the initial belief and each individual signal all decrease. The precision h_t also increases linearly in t, meaning that as information grows the Bayesian evaluator is more and more certain about her assessment of the location of η, which she believes to be around m_t.

Given that $m_t(z^{t-1}) = E[\eta|y^{t-1}]$, equation (3) becomes:

$$
w_t(y^{t-1}) = m_t(z^{t-1}) + a_t^*(y^{t-1}) \qquad (8)
$$

In equilibrium, the wage at time t is composed by the evaluator's contemporary belief about η given the history of signals z^{t-1}, plus the anticipated equilibrium decision rule $a_t^*(y^{t-1})$.

Taking expectations to (8), and using (5) and (1) yields:

$$
E[w_t(y^{t-1})] = \left(\frac{h_0}{h_0+t\,h_\varepsilon}\right)m_0 + \left(\frac{h_\varepsilon}{h_0+t\,h_\varepsilon}\right)\sum_{j=1}^{t}\{m_j + a_j - E[a_j^*(y^{j-1})]\} + E[a_t^*(y^{t-1})] \qquad (9)
$$

The marginal return of labor at time t, α_t, will then be:

$$
\alpha_t = \frac{h_\varepsilon}{h_t} = \frac{h_\varepsilon}{h_0+t\,h_\varepsilon} \qquad (10)
$$

The marginal return to labor is hence independent of the past history, and only depends on the period t, and the relative precisions of the prior belief and of the signal error.

From (4), the candidate's decision rule is given by the condition:

$$\gamma_t \equiv \sum_{s=t}^{\infty} \beta^{s-t} \alpha_s = g'(a_t^*) \tag{11}$$

That is, the optimal effort is such that the discounted *ad infinitum* stream of present and future marginal returns to current effort equals its marginal disutility. Intuitively, the candidate gains from influencing today's posterior, insofar it will become the prior in tomorrow's learning, which will determine subsequent payments, hence the *ad infinitum* nature of the factor γ_t in the optimal effort choice.

Proposition 1. *If the variance of the signal is bounded ($h_\varepsilon > 0$) and the candidate is minimally patient ($\beta > 0$), the effort of a candidate who faces a Bayesian evaluator is strictly positive for any t. That is, $a_t^* > 0$ for any t: the candidate always finds optimal to improve the output.*

Given that $\alpha_t = \frac{h_\varepsilon}{h_0 + t\, h_\varepsilon}$ is a decreasing sequence, γ_t also is. Given that the function of disutility of effort $g(.)$ is convex and increasing, it should be the case that the equilibrium sequence of the candidate's effort a_t^* be decreasing as well.

2.1 Convergence under Bayesian Learning

It is worth analyzing the convergence of the Bayesian beliefs before advancing to the interpretive case. Given the strategical choice of effort of an agent who influences output and therefore changes the information observed by the evaluator, how do the Bayesian evaluator's beliefs result?

Effort We begin our analysis of convergence with the candidate's effort.

Proposition 2. *The optimal effort sequence of a candidate who faces a Bayesian evaluator strictly decreases over time: $\frac{\partial a_t^*}{\partial t} < 0$.*
Furthermore, as $t \to \infty$, the optimal effort converges to zero: $a_t^* \xrightarrow[t\to\infty]{} 0$. The velocity of convergence is 1/t.

Given that $\alpha_t = \frac{h_\varepsilon}{h_0 + t\, h_\varepsilon} \xrightarrow{t\to\infty} 0$, $\gamma_t \xrightarrow{t\to\infty} 0$. Given that the function of disutility of effort $g(.)$ is convex and increasing, it should therefore be the case that $a_t^* \xrightarrow{t\to\infty} 0$.

Intuitively, because a Bayesian evaluator places a uniform weight on every signal, the marginal contribution of a signal at early stages (low ts) is relatively large given the reduced number of total signals; hence, the candidate has stronger incentives to exert effort to produce a high output. Conversely, at later stages (high ts), because of the large amounts of evidence already available to the evaluator, the marginal contribution of a new signal on the posterior becomes minimal, and hence the candidate has fewer incentives to exert effort. This at least partially accounts for the commonly observed behavior among repeated social interactions in which evaluated individuals exhibit higher amounts of effort when they are ”young”, at the initial stages of their reputation building, decreasing it as they become ”older” in the relationship. This is a feature noticed by Holmstrom (1999), which occurs even under Bayesian learning. This dynamic occurs because of the differencial availabilities of information, which sets differential incentives for effort over time. It is worth noticing that the velocity of convergence is “linear”, at rate $1/t$.

Beliefs about talent From (6), we find that, provided enough information, the Bayesian evaluator’s beliefs converge with probability 1 to the true talent of the candidate.

Proposition 3. *As information grows ($t \to \infty$), a Bayesian evaluator’ beliefs converge to the true talent η with probability 1, even in the presence of strategic signalling by the candidate. That is,*

$$\text{plim}_{t\to\infty}\, m_t(a_t^*) = \eta$$

The velocity of convergence of Bayesian beliefs is $1/t$.

Wages Given that we have shown the convergence of the effort and the belief, the convergence of the expected wage is straightforward:

Proposition 4. *As $t \to \infty$, the wage paid by a Bayesian evaluator converges to:*

$$\text{plim}_{t\to\infty}\, w_t(a_t^*) = \eta$$

As we have seen, as t grows to infinity, a Bayesian evaluator’s belief will converge to the talent η, whereas the candidate’s effort will shrink to zero. Therefore, a competitive, Bayesian-learning market will converge to a retribution to labor that is consistent with the true talent η. It is again worth emphasizing that convergence of wages occurs at rate $1/t$ as well.

3 Confirmation-Biased Learning

To model confirmation bias, we consider an interpretive evaluator as proposed by Fryer, Harms and Jackson (2013) (Henceforth, FHJ). We build upon FHJ model in at least four ways. First, we express the equations in terms of *precisions* rather than *variances*. This notably simplifies the math, making the equations elegantly tractable. Moreover, for the purposes of analyzing confirmation bias, it is more intuitive to think in terms of precisions (indicating the degree of *conviction* or *entrenchment* about one’s beliefs) rather than variances (acknowledging the variability of the belief). Second, in their continuous version, FHJ study the behavior of the mean of the distribution, but make no explicit analysis about the evolution of precisions. We provide it here, and, as we will see, this reveals interesting insights about confirmation biased learning. Third, we explicitly posit that, following a two-step Bayesian process, with each piece of information not only the belief increases its precision, but also the precision of the *interpretation* of the information increases as well. Fourth, as we analyze this learning process, we highlight connections with the traditional Psychology literature on confirmation bias that provides empirical foundations for this model.

An interpretive evaluator follows a two-step belief updating process. Before updating her beliefs based on the incoming information, the evaluator reads and *interprets* the information, as if it were closer to her preexisting beliefs. Then, she makes a Bayesian updating based on her *interpretation* instead of the raw information. More specifically, let $\hat{z}$ denote the interpretation of the raw information z and $\hat{m}$ the posterior mean based on the interpretation $\hat{z}$. The two-steps involved are as follows:

1. Interpretation of the information

$$\hat{z}_t = \hat{m}_{t-1}\left(\frac{h^m_{t-1}}{h^z_t}\right) + z_t\left(\frac{h_\varepsilon}{h^z_t}\right) \tag{12}$$

   where h^z_t and h^m_t denote, respectively, the precisions of the interpretation $\hat{z}_t$ and of the interpretive belief $\hat{m}_t$ at time t.
   The interpretive evaluator reads and *interprets* (possibly ambiguous) information z_t based on her pre-existing belief $\hat{m}_{t-1}$, following a Bayes rule. This has the effect of "pulling" the information towards her pre-existing belief.

   This interpretation step that alters information can be thought of as a model of the "information assimilation bias" pointed out by Lord, Ross, and Lepper (1979).

2. Belief updating, based on interpretation $\hat{z}_t$

$$\hat{m}_t = \hat{m}_{t-1}\left(\frac{h^m_{t-1}}{h^m_t}\right) + \hat{z}_t\left(\frac{h^z_t}{h^m_t}\right) \tag{13}$$

   The evaluator updates her prior belief based on her *interpretation* $\hat{z}_t$.

3.1 Dynamics of Confirmatory Biased Learning

To get a better feeling of the implications of this learning process, it is useful to examine the evolution of beliefs over time. Since most readers are likely not familiar with this model, we unpack it and provide a step-by-step presentation here.

We begin at $t = 1$, when the very first piece of raw information z_1 arrives. The information is first interpreted, as follows:

$$\begin{aligned}\hat{z}_1 &= m_0\left(\frac{h_0}{h^z_1}\right) + z_1\left(\frac{h_\varepsilon}{h^z_1}\right)\\ h^z_1 &= h_0 + h_\varepsilon\end{aligned}$$

In the confirmatory biased interpretation $\hat{z}_1$, the location of the information z_1 is pulled towards the initial prior mean (or "prejudice") m_0. In addition, by conforming the information to the initial belief, the variance of the information is reduced, and its perceived precision h_1^z is increased by the addition of one extra prior precision h_0. In other words, a confirmatory biased evaluator *overestimates* the precision of the interpreted information.

The evaluator then updates her beliefs based on her interpretation of the information $\hat{z}_1$:

$$\hat{m}_1 = m_0 \left(\frac{h_0}{h_1^m} \right) + \hat{z}_1 \left(\frac{h_1^z}{h_1^m} \right)$$
$$h_1^m = h_0 + h_1^z = 2h_0 + h_\varepsilon$$

The interpretive learner ends up with a posterior precision of $2h_0 + h_\varepsilon$, whereas a Bayesian evaluator would have a precision of only $h_0 + h_\varepsilon$. Hence, an interpretive evaluator features **overconfidence** about her beliefs. This is one of the features of confirmation bias, as Rabin and Schrag (1999) have noted and demonstrated as well in a discrete model. Their definition still holds for the continuous case: "overconfidence, in the sense that people in average believe more strongly than they should in their favored hypothesis". Mathematically, in the continuous model, this is due to the addition of one extra initial precision h_0.

We can express the interpretive posterior mean as a function of the raw information:

$$\hat{m}_1 = m_0 \left(\frac{2h_0}{2h_0 + h_\varepsilon} \right) + z_1 \left(\frac{h_1^z}{2h_0 + h_\varepsilon} \right)$$

The prior belief m_0 is weighted twice, more than a Bayesian would, which has the effect of pulling the posterior even closer to the initial belief m_0. In effect, confirmation bias produces an **overweight of initial beliefs**, which ultimately leads to a **prior inertia** in the belief updating process.

At $t = 2$, the evaluator gets a new, independent piece of information. A Bayesian evaluator would treat this information as independent. But the interpretation step is as follows:

$$\hat{z}_2 = \hat{m}_1 \left(\frac{h_1^m}{h_2^z} \right) + z_2 \left(\frac{h_\varepsilon}{h_2^z} \right)$$
$$h_2^z = h_1^m + h_\varepsilon = 2h_0 + 2h_\varepsilon$$
$$\hat{z}_2 = m_0 \left(\frac{2h_0}{2h_0 + 2h_\varepsilon} \right) + z_1 \left(\frac{h_\varepsilon}{2h_0 + 2h_\varepsilon} \right) + z_2 \left(\frac{h_\varepsilon}{2h_0 + 2h_\varepsilon} \right)$$

At $t = 2$, the precision h_2^z of the interpretation $\hat{z}_2$ is $2h_0 + 2h_\varepsilon$, higher than the precision of the raw information z_2 (which is only h_ε). Notice how the initial prior precision h_0 is added twice in the interpretation of the second period information (more than what happened with the interpretation at $t = 1$). Moreover, the precision of previous information z_1 is added once in the interpretation of this *independent* second piece of information, via the prior $\hat{m}_1$.

Now, looking at the mean of the interpreted information, $\hat{z}_2$, it is not only the case that the initial belief m_0 is overweighted again (this time, appearing twice). It is also the case that the first piece of information z_1 has an influence on the interpretation of the second, *independent* piece of information. **Early information influences the interpretation of subsequent information**. This is consistent with the classical Asch (1946) experiments[6], which showed empirically that early information conditions the interpretation of later information. Mathematically, the influence of early signals on the interpretation of subsequent signals is shown by the presence of z_1 in the interpretation $\hat{z}_2$.

Based on this interpretation, the posterior belief is updated as:

$$\begin{aligned}
\hat{m}_2 &= \hat{m}_1 \left(\frac{h_1^m}{h_2^m} \right) + \hat{z}_2 \left(\frac{h_2^z}{h_2^m} \right) \\
h_2^m &= h_1^m + h_2^z = 4h_0 + 3h_\varepsilon \\
\hat{m}_2 &= m_0 \left(\frac{4h_0}{4h_0 + 3h_\varepsilon} \right) + z_1 \left(\frac{2h_\varepsilon}{4h_0 + 3h_\varepsilon} \right) + z_2 \left(\frac{h_\varepsilon}{4h_0 + 3h_\varepsilon} \right)
\end{aligned}$$

By $t = 2$ the initial prior gets 4 times its precision. Moreover, notice that the early signal z_1 is now given *twice* the precision that a Bayesian evaluator would assign. This is corresponds to another effect of confirmation bias, a **preference for early information**, which is a corollary of the fact that early information influences interpretation of future information. Notice how the information is weighted more strongly when it appears earlier in the series. Because of this, confirmatory biased evaluators are more influenced by *first impressions.* This implies that the *exchangeability* property of Bayesian learning does not hold anymore and instead the confirmatory biased learning features *informational path-dependencies.*

To get a grasp of what happens as information grows, consider $t = 4$. By $t = 4$, the interpretation rule becomes:

$$\begin{aligned}
\hat{z}_4 &= \hat{m}_3 \left(\frac{h_3^m}{h_4^z} \right) + z_4 \left(\frac{h_\varepsilon}{h_4^z} \right) \\
h_4^z &= h_3^m + h_\varepsilon = 8h_0 + 8h_\varepsilon \\
\hat{z}_4 &= m_0 \left(\frac{8h_0}{8h_0 + 8h_\varepsilon} \right) + z_1 \left(\frac{4h_\varepsilon}{8h_0 + 8h_\varepsilon} \right) + z_2 \left(\frac{2h_\varepsilon}{8h_0 + 8h_\varepsilon} \right) + z_3 \left(\frac{h_\varepsilon}{8h_0 + 8h_\varepsilon} \right)
\end{aligned}$$

[6] In Asch(1946) experiments, different individuals were given a sequence of words describing a person. The study found that the order in which positive and negative adjectives describing a person where presented influenced their overall perception of the described person. Sequences with positive adjectives at the beginning and negative ones at the end induced a more positive perception, while the inverse order elicited a more negative perception, suggesting that early information conditions the interpretation of the following information.

based on which the posterior belief is updated as:

$$
\begin{aligned}
\hat{m}_4 &= \hat{m}_3 \left(\frac{h_3^m}{h_4^m}\right) + \hat{z}_4 \left(\frac{h_4^z}{h_4^m}\right) \\
h_4^m &= h_3^m + h_4^z = 16h_0 + 15h_\varepsilon \\
\hat{m}_4 &= m_0 \left(\frac{16h_0}{16h_0 + 15h_\varepsilon}\right) + z_1 \left(\frac{8h_\varepsilon}{16h_0 + 15h_\varepsilon}\right) + z_2 \left(\frac{4h_\varepsilon}{16h_0 + 15h_\varepsilon}\right) \\
&\qquad + z_3 \left(\frac{2h_\varepsilon}{16h_0 + 15h_\varepsilon}\right) + z_4 \left(\frac{h_\varepsilon}{16h_0 + 15h_\varepsilon}\right)
\end{aligned}
$$

By recursion, the rules can be expressed as a function of t as:

$$
\hat{z}_t = m_0 \left(\frac{h_0}{h_0 + h_\varepsilon}\right) + \left[\sum_{j=1}^{t-1} z_j \frac{1}{2^j} + z_t \frac{1}{2^{t-1}}\right] \left(\frac{h_\varepsilon}{h_0 + h_\varepsilon}\right) \tag{14}
$$

$$
h_t^z = 2^{t-1}(h_0 + h_\varepsilon) \tag{15}
$$

$$
\hat{m}_t = m_0 \left(\frac{2^t h_0}{2^t(h_0 + h_\varepsilon) - h_\varepsilon}\right) + \sum_{j=1}^{t} z_j \left(\frac{2^{t-j} h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon}\right) \tag{16}
$$

$$
h_t^m = 2^t(h_0 + h_\varepsilon) - h_\varepsilon \tag{17}
$$

4 Evaluation and Retribution under Confirmation Biased Learning

We now introduce a confirmatory biased evaluator to the dynamic wage equlibrium model. Again, the evaluator can learn about η by observing z_t, but now a confirmatory biased, interpretive process mediates the learning.

Her posteriors will now be given by:

$$
\hat{m}_t = m_0 \left(\frac{2^t h_0}{2^t(h_0 + h_\varepsilon) - h_\varepsilon}\right) + \left(\frac{2^t h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon}\right) \sum_{j=1}^{t} z_j \left(\frac{1}{2^j}\right) \tag{18}
$$

$$
h_t^m = 2^t(h_0 + h_\varepsilon) - h_\varepsilon \tag{19}
$$

Using (18), by $\hat{m}_t(z^{t-1}) = E[\eta | y^{t-1}]$, and taking expectations, the competitive market equilibrium wage (3) becomes:

$$
\begin{aligned}
E[w_t(y^{t-1})] =& m_0 \left(\frac{2^t h_0}{2^t(h_0 + h_\varepsilon) - h_\varepsilon}\right) \\
&+ \left(\frac{2^t h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon}\right) \sum_{j=1}^{t} \left\{ [\hat{m}_j + a_j - E[a_j^{**}(y^{j-1})]](\frac{1}{2^j}) \right\} \\
&+ E[a_t^{**}(y^{t-1})]
\end{aligned} \tag{20}
$$

where a_t^{**} is the decision rule in the interpretive case.

The marginal return for labor a_t in the interpretive case is:

$$\hat{\alpha}_t = \frac{h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon} \tag{21}$$

Given the candidate's maximization problem and the beliefs of the interpretive evaluator, the candidate's decision rule is given by:

$$\hat{\gamma}_t \equiv \sum_{s=t}^{\infty} \beta^{s-t} \hat{\alpha}_s = g'(a_t^{**}) \tag{22}$$

The following proposition is analogous to the Bayesian case:

Proposition 5. *If the variance of the signal is bounded ($h_\varepsilon > 0$) and the candidate is minimally patient ($\beta > 0$), the effort of a candidate who faces an confirmation-biased evaluator is strictly positive for all t. That is, $a_t^{**} > 0$ for all t: the candidate always finds optimal to improve the output.*

Proofs are analogous to the Bayesian case.

The expected equilibrium wage in a confirmatory biased market is determined as:

$$E[w_t(y^{t-1})] = m_0 \left(\frac{2^t h_0}{2^t(h_0 + h_\varepsilon) - h_\varepsilon} \right) + \sum_{j=1}^{t} \hat{m}_j \left(\frac{2^{t-j} h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon} \right) + a_t^{**}(y^{t-1}) \tag{23}$$

which at any t is the sum of the current (confirmation-biased) belief and the candidate's optimal decision rule. As we can see, the confirmation-biased equilibrium wage is determined by the initial belief m_0 in a larger part than in the Bayesian equilibrium wage.

4.1 Convergence of Confirmation-Biased Market and Differences in Long Run Equilibria

We now analyze the main features of the long-run convergence, after the confirmatory-biased evaluators in the market have learned through arbitrarily high amounts of information. We then provide a comparison between the confirmatory biased and the Bayesian long-run equilibria.

4.1.1 Effort

We begin by analyzing the incentives faced by the candidate.

Proposition 6. *The optimal effort sequence of a candidate who faces a confirmatory biased evaluator, though always strictly positive, is decreasing in time. Furthermore, as $t \to \infty$, the optimal effort converges to zero:*

$$a_t^{**} \xrightarrow[t\to\infty]{} 0.$$

Proposition 7. *The effort of a candidate who faces an interpretive evaluator decreases towards 0 faster than that of a candidate who faces a Bayesian evaluator, since the speed of convergence of the marginal benefit of effort in the interpretive case is $1/2^t$, while it is $1/t$ in the Bayesian case.*

In both the interpretive and the Bayesian cases, the marginal returns to labor are strictly positive, and decreasing in t, so that the candidate exerts a positive yet decreasing amount of effort in either case. However, while the Bayesian convergence rate is linear (at speed $1/t$), the interpretive case is *exponential* (at speed $1/2^t$). Thus, the interpretive candidate reduces his effort faster, given that his evaluator converges more quickly in her beliefs due to the mentioned *overweight of initial beliefs* and her *over-influence of early signals.*

Given this result, we would expect that the interpretive candidate exert more effort than the Bayesian one at early stages. However, this is true only if the signal precision h_ε is sufficiently high relative to h_0, as stated in the following proposition:

Proposition 8. *If $\frac{h_\varepsilon}{h_0+h_\varepsilon} > \underline{h} \equiv ln(2)$, there exists $\bar{t} > 0$ such that for all $t < \bar{t}$, $a_t^{**} > a_t^*$ and $a_t^{**} < a_t^*$ for $t > \bar{t}$.*

If $\frac{h_\varepsilon}{h_0+h_\varepsilon} < \underline{h} \equiv ln(2)$, $a_t^{**} < a_t^*$ for all $t > 0$.

Proof. See appendix. □

Intuitively, if the signal precision is high relative to the prior precision (i.e., the interpretive evaluator has a relatively low degree of entrenchment of her beliefs), then the belief updating will rely more (though not sufficiently) on the information, and given the preference for early information, the candidate will at early stages exert higher amounts of effort than when he faces a Bayesian evaluator (who places uniform weights on information over time). However, if the prior precision is sufficiently high relative to the signal precision (so that the evaluator is deeply convinced about her beliefs), there are simply no strong incentives to exert great effort even at initial stages given that initial beliefs are excessively overweighted.

4.1.2 Beliefs

We now analyze the long run convergence of posterior beliefs.

Proposition 9. *As information grows to infinity ($t \to \infty$), the interpretive evaluator's posterior mean converges in expectation to:*

$$E[\hat{m}_t] \xrightarrow{t\to\infty} \left(\frac{h_0}{h_0+h_\varepsilon}\right) m_0 + \left(\frac{h_\varepsilon}{h_0+h_\varepsilon}\right) \eta$$

at an exponential convergence rate of 2^t.

Even with arbitrarily high information, the interpretive evaluator does no better than a Bayesian learner with access to only one piece of information that purely reveals the true value η.

Unlike a Bayesian learner, an interpretive evaluator never converges to the real value of η after an infinite number of signals. If $h_0 > 0$ -a fairly general condition-, the influence of the initial belief always remains, and the candidate has no way to counter it and to place the evaluator's belief at the real value of η, even with his optimal signal influencing played *ad infinitum*. For a sufficiently patient candidate, this means that even though under some conditions he might exert higher levels of effort than those he would expend had he faced a Bayesian evaluator, he will nonetheless never get the interpretive evaluator have his belief at the true value of η.

The weight of the initial belief m_0 decreases in the signal-to-prior precision ratio h_ε/h_0. This means that, ceteris paribus, a relatively smaller initial prior precision h_0 (larger prior variance σ_0^2) would move the asymptotic posterior $\hat{m}_t$ closer to η. However, the very idea of a significantly large prior variance σ_0^2 seems to be at odds with the figure of a confirmatory biased evaluator. Alternatively, increasing the other component of this ratio, namely the signal precision h_ε, has the same effect. Therefore, an alternative to close the gap between the interpretive belief and the true value η is to improve or find an alternative signalling mechanism with lower variance, so that the evaluator's inferences rely more on the new information. In any case, however, we can see that unless $h_0 = 0$ (which amounts to an infinite variance σ_ε^2), the effect of the initial belief will inevitably remain.

Most interestingly, in the long run, the belief of a confirmatory biased evaluator behave as that of a Bayesian evaluator who observes the true value of η just one time. That is, even with the best actions the candidate can do, an interpretive evaluator does no better than a Bayesian evaluator who observes a pure signal with the true value of η *only once*. This characterization will help understand up to what degree an interpretive evaluator's initial beliefs can be expected to be changed as the result of an infinite stream of strategically influenced information.

The convergence takes place at a rate 2^t, faster than the Bayesian belief, meaning that the biased equilibria will be reached faster, with the candidate's efforts converging to zero faster as well.

4.1.3 Wages

We now turn to analyze the long run equilibrium wages. In the Bayesian case:

$$w_t = m_t + a_t^*$$
$$w_t \xrightarrow[t\to\infty]{} \eta$$

since the Bayesian belief converges to η and the effort decreases to zero. So, Bayesian wages converge to the true talent.

On the other hand, the equilibrium wage in the interpretive market is:

$$\hat{w}_t = \hat{m}_t + a_t^{**} \tag{24}$$

$$\hat{w}_t \xrightarrow[t\to\infty]{} m_0 \left(\frac{h_0}{h_0 + h_\varepsilon}\right) + \eta \left(\frac{h_\varepsilon}{h_0 + h_\varepsilon}\right) \tag{25}$$

Since interpretive effort decays to zero, the equilibrium wage converges to a value equal to the asymptotic belief. Therefore, there will be a *persistent* effect of the initial belief in the retribution to talent.

Retribution Bias It is of interest to measure the bias in retributions to labor as generated by confirmation biased learning. We define the **Retribution Bias** as the difference between the interpretive wage and the Bayesian wage:

$$\text{Retribution bias} \equiv \hat{w}_t - w_t \tag{26}$$

$$= (\hat{m}_t - m_t) - (a_t^{**} - a_t^*) \tag{27}$$

The following proposition holds:

Proposition 10. *In the long run (as $t \to \infty$), the retribution bias $\equiv \hat{w}_t - w_t$ converges to:*

$$\textit{Retribution bias} \xrightarrow{t\to\infty} \tfrac{h_0}{h_0+h_\varepsilon}\,(m_0 - \eta)$$

In the long run, the Retribution Bias is determined by two components. First, the distance between the initial belief and the true ability $(m_0 - \eta)$. That is, how far is the initial assessment of talent from the talent itself. If the interpretive evaluator holds negative initial beliefs about the ability of the candidate, there will be *persistent* under-retributions to his talent, no matter the amount of information. Conversely, if the initial belief overrates the candidate's talent, a payment that is overly excessive will persist over time. The second factor determining the retribution bias is the ratio $\frac{h_0}{h_0+h_\varepsilon}$. The higher the *initial* prior precision (the degree of "conviction" about one's initial beliefs), the greater retribution bias will persist given an initial distance. Therefore, in determining the long-run retribution bias, not only does the initial under-assessment of talent matter, but also the degree of conviction or entrenchment of beliefs about the initial assessment.

Holmstrom (1999) highlighted that the effort of candidates to improve their own reputations lead them to oversupply effort at early stages and undersupply it at later stages, creating a dynamic incentive problem leading to a market inefficiency. We have found an additional potential source of surplus loss. In the presence of confirmation biased learning, the inefficiencies found in Holmstrom (1999) are even aggravated by a mis-appreciation of talent. Moreover, unlike Holmstrom (1999), misplaced beliefs about talent can be persistent over time, locked in a non-corrective stationary state.

5 Prejudices Based on Social Categories: Initial Beliefs Conditional on Observable Attributes

In this section, we consider the case when a confirmatory biased evaluator differentiates his initial beliefs conditioning them on an observable characteristic g. In general terms, g can be any observable attribute. More specifically, g can be thought of as a group or social category to which the candidate belongs (such as gender, ethnicity, or any other). g can be either a univariate class, or a combination of multi-dimensional classes.

To simplify the analysis, consider a set of candidates $i = 1, 2, ..., I$. Each candidate i belongs to a group $g(i) \in \{A, B\}$. Assume the evaluator conditions her initial beliefs based on the characteristic g, such that:

$$\eta|g(i) = \begin{cases} \eta|A \sim \mathcal{N}(m_0^A, 1/h_0^A) & \text{if } g(i) = A \\ \eta|B \sim \mathcal{N}(m_0^B, 1/h_0^B) & \text{if } g(i) = B \end{cases} \tag{28}$$

where m_0^A can be either equal to m_0^B or not.

For the sake of the exposition, let's suppose that individual i belongs to group A and individual j belongs to group B. Their talents are respectively η_i and η_j. We deliberately make *no* assumptions about their relationship: i could be more talented than j, j could be more talented than i, or both could be equally talented. However, assume that, for any reason, $m_0^A > m_0^B$, so that the evaluator initially believes that, in expectation, the talent of an individual from group A is greater than that of an individual of group B. This belief might be well founded or not.

5.1 Wages

As we have seen, under Bayesian learning, the wages of two individuals i and j will converge to their true talents:

$$w_t^i \xrightarrow{t\to\infty} \eta_i \tag{29}$$

$$w_t^j \xrightarrow{t\to\infty} \eta_j \tag{30}$$

Bayesian markets will correct any misplaced beliefs of any initial precision over time provided enough information -even if the beliefs are initially conditioned on the group characteristic.

On the other hand, under confirmatory biased learning, the wages for those two individuals will converge to:

$$\hat{w}_t^i \xrightarrow{t\to\infty} m_0^A \left(\frac{h_0^A}{h_0^A + h_\varepsilon}\right) + \eta_i \left(\frac{h_\varepsilon}{h_0^A + h_\varepsilon}\right) \tag{31}$$

$$\hat{w}_t^j \xrightarrow{t\to\infty} m_0^B \left(\frac{h_0^B}{h_0^B + h_\varepsilon}\right) + \eta_j \left(\frac{h_\varepsilon}{h_0^B + h_\varepsilon}\right) \tag{32}$$

That is, initial beliefs m_0^A, m_0^B based on group characteristics will remain as a determining factor of the long run wage. More specifically, candidate i's equilibrium wage will be a combination of his individual talent η_i, and the evaluator's conditional prejudice m_0^A over group A, weighted by the "entrenchment" of her prior h_0^A relative to the signal precision. Candidate j's wage might be different from candidate i's because of differences in *any* of those three factors.

5.2 Inter-Individual Wage Gap

An important part of public discussion about gender wage fairness revolves around the wage gap between the two groups. Define the **Inter-Individual Wage Gap** as the difference between the wage of two given individuals. In Bayesian markets:

$$\text{Inter-Individual Wage Gap}_{i,j}^{Bayes} \equiv (w_t^i - w_t^j) \xrightarrow{t\to\infty} (\eta_i - \eta_j) \tag{33}$$

In Bayesian markets, the inter-individual wage gap converges to the inter-individual talent gap.

However, under confirmation bias (assuming for simplicity that initial precisions h_0^g for all groups g are equal), the following proposition holds:

Proposition 11. *In a confirmation biased market with conditional priors m_0^A, m_0^B with $h_0^A = h_0^B = h_0$,*

$$\begin{aligned} \textit{Inter-Individual Wage Gap}_{i,j}^{Interp} &\equiv (\hat{w}_t^i - \hat{w}_t^j) \\ &\xrightarrow{t\to\infty} \frac{h_0}{h_0 + h_\varepsilon}(m_0^A - m_0^B) + \frac{h_\varepsilon}{h_0 + h_\varepsilon}(\eta_i - \eta_j) \end{aligned}$$

When markets are confirmation-biased, with priors conditional on group attributes, the **inter-individual wage gap** is a weighted average of the *inter-individual* talent gap $(\eta_i - \eta_j)$, and the *inter-group* prejudice gap $(m_0^A - m_0^B)$. This distinction is important: considerable part of public discussion seems to rush to conclusions that group differences in wages are mainly the product of an inter-group prejudice gap, dismissing that there might be differences in the composition of individual talents between the compared groups. On the other hand, if markets learn with confirmation biases, a wage gap *does* reflect some degree of a group-based prejudice gap $(m_0^A - m_0^B)$ if $m_0^A \neq m_0^B$ and $h_0 > 0$.

The weights in the average are given by the relative strengths of the prior-to-signal precisions. Relatively high h_0 (stronger *a priori* convictions about initial beliefs) leads to a greater influence of the inter-group prejudice gap $(m_0^A - m_0^B)$. Conversely, higher h_ε (better chances to convey information through less noisy signals) moves the long-run wage gap closer to the inter-individual talent gap $(\eta_i - \eta_j)$, reducing the importance of the inter-group prejudice gap $(m_0^A - m_0^B)$. Unless markets are not affected by confirmation bias, it seems precipitous to entirely attribute a wage gap to either a prejudice gap or a talent gap. The prior-to-signal relative precision will dictate the strength of the correspondence between the observed wage gap and either of those two components.

5.3 Intra-Individual Retribution Bias

What ultimately matters, at least from a standpoint of efficiency, is not the comparison between the wages of two individuals, but the deviations of the individual wage from the individual talent itself. In the limit, this is measured by the **Retribution Bias**, the distance of paid wage from real talent η (since effort ultimately shrinks to zero). The long run Retribution Bias for each individual converges to:

$$\text{Retribution Bias}_i \xrightarrow{t\to\infty} \frac{h_0^A}{h_0^A + h_\varepsilon}(m_0^A - \eta^i) \tag{34}$$

$$\text{Retribution Bias}_j \xrightarrow{t\to\infty} \frac{h_0^B}{h_0^B + h_\varepsilon}(m_0^B - \eta^j) \tag{35}$$

We can see that *even if two individuals are equally talented ($\eta_i = \eta_j$)*, they might face different degrees of retribution bias (wages departing from their own productivities due to confirmation bias) because of differences in: a) group-based prejudices m_0^A, m_0^B, and b) in degrees of "conviction" h_0^A, h_0^B. In short, unless $h_0 = 0$ or $m_0 = \eta$, interpretive learning generates persistent Retribution Biases of differing magnitudes for different individuals, depending on the parameters of the conditional initial beliefs about group-based distributions.

5.4 Retribution Bias Gap

Given that two individuals who belong to two different groups can be *unequally* disadvantaged by different magnitudes of retribution bias, we can consider who is *more* disadvantaged, and to what degree. Define the **Retribution Bias Gap**$(i, j)\equiv$ Retribution Bias$_i$ − Retribution Bias$_j$. This gap measures the relative advantage or disadvantage individual i has due to the existence of confirmation biases relative to other individual j. Again, assume $h_0^A = h_0^B = h_0$ for simplicity. Then:

$$\text{Retribution Bias Gap}(i, j) \xrightarrow{t\to\infty} \frac{h_0}{h_0 + h_\varepsilon}\left[(m_0^A - m_0^B) - (\eta^i - \eta^j)\right] \tag{36}$$

Who is retributed more unfavorably by a confirmatory-biased market? A positive Retribution Bias Gap(i, j) means that individual i, who happens to belong to group A, is retributed more favorably by a confirmatory biased market than individual j, who happens to belong to group B. This relative advantage in retribution biases is determined by three components. First, the difference in initial beliefs, or the prejudice gap $(m_0^A - m_0^B)$. The more favorable the prejudice for group A relative to group B, the more advantaged by the confirmation bias is individual i relative to j, given that i belongs to a group more esteemed than the group to which j belongs. Second, the relative prior to signal precision acts as an amplifying factor that magnifies or shrinks this advantage. The greater the "conviction" of beliefs, the larger the relative advantage of one individual compared to the other becomes. Third, there is an inverse relationship with the inter-individual talent gap $(\eta^i - \eta^j)$, such that the more talented i is relative to j, the more unfavorable is retribution bias for i relative to j. Ceteris paribus, regardless of j's talent, if i is more talented, given initial beliefs m_0^A, a confirmatory biased market will retribute i *less than proportionally* for his talent increase. For each unit of talent increase, his interpretive wage will increase only by $\frac{h_0}{h_0+h_\varepsilon} < 1$. Since i "loses" a fraction $\frac{h_\varepsilon}{h_0+h_\varepsilon}$ of his talent in his retribution, the presence of confirmation bias marginally damages i more than j.

In short, the presence of confirmation bias in a competitive labor market damages relatively more individuals who: a) belong to a group with a worse prejudice (m_0^g), and b) are more talented (η^i), and c) the damage becomes larger as the initial conviction is stronger (h_0^g).

This analysis suggests that if we are to engage in a discussion about the unfairness of confirmation biases in the retribution to labor, a measure to assess the relative unfairness that is more appropriate than the Inter-Individual Wage Gap (widely used in public discussion) would be a measure of Retribution Bias Gap, since it takes into account only the *deviations* from the underlying talent. In other words, the Retribution Bias Gap is orthogonal to the inter-individual talent gap, and therefore controls for any variation in that dimension. The inter-individual wage gap, insofar is correlated with talent gaps, might be a confounding measure of the effect prejudices.

6 Conclusions

In this work, we have posited a model of strategic interaction between an interpretive, confirmatory-biased evaluator and a candidate who can partially influence the information the evaluator gets to see and interpret. We found that, as in the case of a Bayesian evaluator, an agent facing an interpretive evaluator will in fairly broad conditions always exert a strictly positive level of effort to influence the signals. In addition, the optimal level of effort decreases over time, since it results harder to change beliefs when those are based on greater amounts of a previous body of information -regardless of whether learning is interpretive or not.

Nonetheless, a major difference is that while a Bayesian evaluator converges in his beliefs to the true value of η, an interpretive evaluator never converges in her beliefs to the true value of η if $h_0 > 0$. In other words, regardless of the number of signals received, her prejudice m_0 will have a strictly positive weight in the posterior belief, even in the limiting distribution. This has a correspondence to the "perseverance of initial beliefs" pointed out by Anderson, Lepper, and Ross (1980). We have highlighted that, whereas the dynamic inefficiencies of reputation building in Bayesian markets stated by Holmstrom (1999) are self-correcting, confirmatory biased markets, because of that perseverance, feature a new source of inefficiency that is of permanent nature. Because of the self-enforcing interpretation process, the convergence is attained at exponential speed (at rate 2^t), which takes place faster than the Bayesian evaluator's convergece (at a linear rate t). Therefore, not only does the confirmatory-biased evaluator converge to beliefs that are off the true talent; she also reaches that biased stationary state faster.

The two key parameters that govern where the interpretive posterior will be located between the prior m_0 and the true talent η are the prior and signal precisions h_0 and h_ε. A typically interesting situation arises when the evaluator has deeply entrenched beliefs about the type η, highly concentrated around her initial prior m_0 with a low variance σ_0^2 (high h_0). Those conditions are especially difficult for the candidate who hopes to move the evaluator's belief at his true type η, since the evaluator will place a relatively high weight to her prejudice m_0 in her asymptotic belief. One possible solution to reduce the bias and improve the fairness of the evaluation (and hence the candidate's retribution) is to design alternative signaling mechanisms that feature larger signal precisions h_ε, which by making inference clearer, will induce the evaluator to rely more on the information, ultimately moving her beliefs closer to η. However, it remains a very gloomy result that, for any $h_0 > 0$, or $h_\varepsilon < \infty$, the candidate is not able to erase the effect of the initial belief, even with his optimal signalling played infinite times.

Regarding the retribution to talent, we have seen that given that an interpretive evaluator has a permanent bias, there will in turn be persistent biases in the retribution to talent. Specifically, the long run retribution bias will be equal to $\frac{h_0}{h_0+h_\varepsilon}(m_0 - \eta)$. The long-term retribution bias increases in the initial bias $(m_0 - \eta)$ and in the prior to signal precision h_0/h_ε.

We have also provided a framework to analyze differences in retribution biases between individuals who have an observable characteristic (which may correspond to a "group" or social category they belong to). By explicitly identifying the inter-individual talent gap and the initial prejudice differences between groups, this work can hopefully shed some light to the ongoing debates about inter-group differences in retribution to labor. First, we examined how a plain wage gap (a measure often referred to in public discussion) conflates two different factors contributing to wage differences: inter-group prejudice gaps and inter-individual talent gaps. Since this measure does not control for variations in inter-individual talent differences, no conclusions can be made from a mere observation of inter-individual wage differences about the effect of group-based prejudices in confirmation biased markets, unless additional information is considered. We have provided an alternative measure, the Retribution Bias Gap, which considers instead the *deviations* from talent of different individuals, and compares the lenghts of those deviations. By orthogonalizing the measure from the talent dimension, the source of confounding is removed, providing a more robust indicator of retribution deviations from talent due to inter-group differences in prejudices.

If human learning process features confirmation bias and if we are to reduce retribution biases, a mathematical analysis of this problem can shed light to the discussion about possible solutions. To begin with, it seems very challenging to erradicate our initial bias $(m_0 - \eta)$. We all may have initial views and expectations m_0 about unobserved attributes such as ability. Since the true talent η is unobservable in the first place, we will likely not know whether moving our initial belief m_0 would get us closer to η. Nonetheless, we can hope to act on at least two levels, by working on the remaining components of the asymptotic bias. First, we can attempt to reduce h_0, the entrenchment of our initial beliefs. The sole awareness of the possibility and pervasiveness of our cognitive vulnerabilities will make us more willing to grant a higher variance to our initial views. A second way to reduce our biases is by improving the signal precision h_ε. The effects of confirmatory learning can be mitigated as the information we get exposed to is less ambiguous and more precise. However, as we have seen, even if we improve our signalling mechanisms, a final bias will likely remain, whenever $h_0 > 0$. These solutions are far from perfect, and more research will be needed to find more effective ways to minimize confirmation biases in our learning process, and therefore in the market retribution to talent.

Though this work hopefully sheds light and provides an analytical framework for the discussion about retribution to talent and fairness, the model presented here has limitations worth mentioning. First, we have adopted a very specific functional form to model confirmation bias, based on the continuous version of FHJ (2013). We should note that this is a “mild” confirmation bias model, in which, as information grows, beliefs move closer (though not sufficiently) to the true talent. Therefore, the initial bias (and so the degree of polarization between two observers with different initial beliefs) is *reduced* with information. This is at least debatable considering the aforementioned evidence from Social Psychology studies showing that opposing initial beliefs diverge even in the presence of the same body of information, and in light of the recent polarizing trends in political and social worldviews. Our “mild” confirmation bias process, though highly useful to derive insights because of its tractability, may be replaced by other “stronger” confirmation bias models, with polarizing long run equilibria. However, we have found that sustained under-retributions to talent are likely to emerge and persist over time even under mild confirmation biased learning. If anything, stronger confirmation-bias processes will most surely exacerbate this result.

Finally, in this work we have indeed adopted a very important assumption: that the individual talents η_i do not change over time. This is a strong assumption, and it might well be the case that individuals and candidates are able learn and improve their talents over time. If talents change over time, we expect that, because in confirmatory biased markets an increase in talent leads to a *subproportional* increase in the long-run retribution to talent, unfavored individuals will need to “over-increase” their talent to offset the initial group prejudice to achieve the level of retribution that would have been fair in the first place. However, given this latter increase in talent, the retribution would be again lagging behind, so that retribution fairness would not be attained. Additionally, if talent building is a time-consuming process, by the time talent is improved the evaluator would have more entrenched beliefs -with higher precision- given all the belief updating already in place. So, if learning takes time, because time is *exponentially* valuable in confirmatory biased markets, incentives for learning, training, and improving job-related skills might be especially eroded under confirmatory biased evaluation of talent.

Appendix

Proof of Proposition 8

Recall that in the Bayesian case,

$$\alpha_t = \frac{h_\varepsilon}{h_0 + t\,h_\varepsilon} \xrightarrow{t\to\infty} 0$$

which occurs at velocity of convergence t.

On the other hand, in the confirmatory biased case,

$$\alpha_t = \frac{h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon} \xrightarrow{t\to\infty} 0$$

so that the velocity of convergence is 2^t.

We have shown that the interpretive effort decays to zero faster than the Bayesian effort. We now evaluate if there exists any $\bar{t}$ such that $\forall t < \bar{t}, a_t^{**} > a_t^*$. This occurs if and only if:

$$\begin{aligned}
\frac{\hat{\alpha}_t}{\alpha_t} &> 1 \\
\frac{h_0 + t\,h_\varepsilon}{2^t(h_0 + h_\varepsilon) - h_\varepsilon} &> 1 \\
\frac{2^t(h_0 + h_\varepsilon) - h_\varepsilon}{h_0 + t\,h_\varepsilon} &< 1 \\
2^t(h_0 + h_\varepsilon) - h_\varepsilon &< h_0 + t\,h_\varepsilon \\
(2^t - 1)(h_0 + h_\varepsilon) &< t\,h_\varepsilon \\
\frac{(2^t - 1)}{t} &< \frac{h_\varepsilon}{h_0 + h_\varepsilon}
\end{aligned}$$

As we have seen, the interpretive effort decays faster than the Bayesian effort, so we are interested in whether there exists a $\bar{t}$ in the vicinity of $t = 0$. Hence, to make t as small as possible, we take the limit as $t \to 0$:

$$\lim_{t\to 0} \frac{(2^t - 1)}{t} = \ln(2)$$

This gives us the following conditions for the existence of $\bar{t}$:

$$\begin{cases}
\text{If } \frac{h_\varepsilon}{h_0 + h_\varepsilon} > \ln(2),\ \exists\bar{t} \text{ such that} \forall t < \bar{t}, a_t^{**} > a_t^* \\[2ex]
\text{If } \frac{h_\varepsilon}{h_0 + h_\varepsilon} < \ln(2),\ a_t^{**} < a_t^* \forall t
\end{cases}$$

In other words, the signal precision should be sufficiently high relative to the prior precision to induce enough incentives for the candidate to exert effort in the face of a confirmatory biased evaluator.

■

References

Psychology. Social Psychology.

ANDERSON, C., LEPPER M. R., and ROSS, L. (1980). "Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information". *Journal of Personality and Social Psychology*, 39, 6, 1037-1049.

ASCH, S. (1946). "Forming impressions of personality". *Journal of Abnormal and Social Psychology*, 1946, 41, 258-290.

BARON, J. (1988). *Thinking and Deciding.* Cambridge University Press, 4th edition.

DARLEY, J., and GROSS, P. (1983). "A Hypotehsis-Confirming Bias in Labelling Effects". *Journal of Personality and Social Psychology*, 44, 1, 20-33.

HOLLYN, J., and SEIFERT, C. (1994). "Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 6, 1420-1436.

KLAYMAN, J., and HA, Y. W. (1987). "Confirmation, Disconfirmation, and Information in Hypothesis Testing". *Psychological Review*, 94, 2, 211-228.

LEPPER, M. R., ROSS, L., and LAU, R. R. (1986). "Persistence of inaccurate and discredited personal impressions: A field demonstration of attributional perseverance". *Journal of Personality and Social Psychology*, 482-491.

LORD, C., ROSS, L., and LEPPER, M. R. (1979). "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence". *Journal of Personality and Social Psychology.* Vol. 37, No.11, 2098-2109.

MERTON, R. K. (1948). "The self-fulfilling prophecy". *Antioch Review*, 8, 193-210.

MYERS, D., and LAMM, H. (1976). "The Group Polarization Phenomenon". *Psychological Bulletin*, 83, 4, 602 627.

NISBETT, R., and ROSS, L. (1980). *Human inference: strategies and shortcomings of social judgement.* Englewood Cliffs, N.J.: Prentice-Hall, in press.

ROSS, L., LEPPER, M.R., and HUBBARD, M. (1975). "Perseverance in self-perception and social perception: Biased attributional proceses in the debriefing paradigm". *Journal of Personality and Social Psychology*, 1975, 32, 880-892.

TVERSKY, A., and KAHNEMAN, D. (1974). "Judgement under Uncertainty: Heuristics and Biases". *Science*, New Series, 185, 4157, pp. 1124-1131.

WASON, P. C. (1960). "On the failure to eliminate hypotheses in a conceptual task". *The Quarterly Journal of Experimental Psychology*, 12, 129-140.

Economics. Information Economics.

DEWATRIPONT, M., JEWITT, I. and TIROLE, J. (1999), "The Economics of Career Concerns, Part I: Comparing Information Structures", *Review of Economic Studies*, 66, 183-201.

FRYER, R., HARMS P., and JACKSON, M. (2013). "Updating Beliefs when

Evidence is Open to Interpretation: Implications for Bias and Polarization".

GIBBONS, R. (1992), *Game Theory for Applied Economists.* Princeton University Press.

GIBBONS, R. and MURPHY, K.(1992), "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence". *Journal of Political Economy*, 100, 468-505.

GROSSMAN, S. and HART, O. (1983), "An Analysis of the Principal-Agent Problem", *Econometrica*, 51, 7-45.

HOLMSTROM, B. (1999) "Managerial Incentive Problems: A Dynamic Perspective", *The Review of Economic Studies*, Vol. 66, No. 1, Special Issue: Contracts (Jan., 1999), pp. 169-182.

LAFFONT, J.J. and MARTIMORT, D. (2002), *The Theory of Incentives. The Principal-Agent Model.* Princeton University Press.

MILGROM, P. and ROBERTS, J. (1988), "An Economic Approach to Influence Activities", *American Journal of Sociology*, 94, Supplement, 154-179.

MULLAINATHAN, S., and SHAFIR, E. (2013). *Scarcity: Why Having Too Little Means So Much*, Henry Holt and Company.

RABIN, M. and SCHRAG, J. (1999), "First Impressions Matter: A Model of Confirmatory Bias", *Quarterly Journal of Economics.*

International Relations. Political Science.

ALLISON, G. (1971). *Essence of Decision: Explaining the Cuban Missile Crisis.* Boston: Little, Brown, and Company.

BACON, F. (1620). *Novum Organon*, XLVI.

CASHMAN, G. (1993). *What Causes War? An Introduction to Theories of International Conflict.* Chapter 3: "The Individual Level of Analysis: Psychological Explanations for War". New York: Lexington Books.

FUKUYAMA, F. (2017). "The Emergence of a Post-Fact World", *The Year Ahead 2017*, Project Syndicate. https://www.project-syndicate.org/onpoint/the-emergence-of-a-post-fact-world-by-francis-fukuyama-2017-01.

GEORGE, A. (1992). "Adapting to Constraints on Rational Decisionmaking", in ART, R., and JERVIS, R. (eds.) *International Politics: Enduring Concepts and Contemporary Issues*, New York, Harper Collins.

MEARSHEIMER, J. (2001). *The Tragedy of Great Power Politics.* W.W. Norton & Company.

SNYDER, J. (1978). "Rationality at the Brink: The Role of Cognitive Process in Failure Deterrance", *World Politics*, 30:3.

TETLOCK, P. (2005). *Expert Political Judgement. How Good Is It? How Can We Know?* Princeton University Press.

THUCYDIDES (431 B.C.). *History of the Peloponnesian War.* Cambridge: Harvard University Press, 1920. Loeb Classical Library.