Fall 2009

Information Projection: Model and Applications (old version, containing additional material)

Kristof Madarasz, *London School of Economics and Political Science*

Information Projection: Model and Applications[1]

Kristóf Madarász

STICERD, London School of Economics

First Draft: November 2007. Current Draft: November 2009.

[1] I thank George Akerlof, Jean-Pierre Benoit, Colin Camerer, Gary Charness, Vince Crawford, Marina Halac, Daniel Kahneman, Botond Kőszegi, Ulrike Malmendier, Andrea Prat, Marit Rehavi, Joel Sobel, Adam Szeidl and seminar participants at Stanford, Berkeley, London School of Economics, San Diego, London Business School, Oxford, Cambridge, Yahoo! Research Labs, Biased-Belief Workshop at Else, University College London, Iza Bonn, Santa Barbara, Bay Area Algorithmic Game Theory Conference at Google, Hawaii, Iese, Ceu, Gerzensee and in particular Matthew Rabin for helpful comments. All errors are mine.

Abstract

Evidence from psychology, Fischhoff (1975), and economics, Camerer, Loewenstein, and Weber (1989), confirms that people systematically exaggerate the extent to which their private information is available to others. I present a general model of such *information projection*, and apply it to a variety of settings. When assessing an expert's competence using ex-post information, jurors overweigh how much they learn from failed predictions and underweigh how much they learn from successful ones. As a result, they underestimate the competence of experts on average. To defend their reputation, experts are too reluctant to base predictions on ex-ante information that complements, and too eager to base predictions on ex-ante information that substitutes, for what jurors independently learn ex-post. Optimal monitoring is coarser and career incentives are weaker than under Bayesian assumptions. A commitment to asymmetric, rather than symmetric, performance measures can significantly reduce defensive practices. Communication protocols that encourage experts to talk, but restrict the use of messages that complement the speaker's expertise, reduce favoritism and strictly improve welfare.

Keywords: Biased Beliefs, Optimal Monitoring, Defensive Medicine, Asymmetric Career Concerns, Communication Protocols.

1 Introduction

People systematically exaggerate the extent to which their information is available to others, and too often act as if others could have guessed their private information correctly. Learning about the symptoms a patient developed recently, jurors typically exaggerate the likelihood with which a careful physician should have detected cancer earlier. After the failure of an engineering project, such as the tragedy of the Challenger space shuttle in January 1986, investigations exaggerate how easy it would have been ex-ante to avoid the tragedy. Similarly, though they have every incentive to communicate clearly, producers of electronic devices supply user-manuals that are too vague for customers to understand. When advising graduate students, John Cochrane (2005) makes the following observation: "The most important thing in writing is to keep track of what your reader knows and doesn't know. Most Ph.D. students assume far too much. No, we do not have the details of every paper ever written in our heads."[1]

This paper summarizes evidence for and offers a widely applicable model of such information projection. People are aware of informational differences, but exaggerate the extent to which their information is available to others. Information projection will cause evaluators to underestimate experts on average, who in turn engage in specific defensive practices. Organizational solutions that reduce agency costs under standard assumptions, often increase agency costs when people suffer from information projection.

Consider an environment where people observe signals about an underlying state. If Alice suffers from information projection, she exaggerates the probability that the content of her signals is also available to Bob. A key property of this definition is that a biased Alice always exaggerates the value of Bob's information. By projecting information, Alice overestimates how well Bob can do in his own expected utility terms and does so in proportion to how valuable her private information would be to Bob. This identifying property of the model captures a variety of seemingly unrelated social mispredictions, and provides a unified framework to study their consequences empirically and theoretically. It also distinguishes my approach from earlier anchoring-based descriptions of these mispredictions.

To illustrate the consequences of such exaggeration, consider a medical example. A radiologist recommends a treatment based on a noisy radiograph. Suppose radiologists differ in ability; the best ones hardly ever miss a tumor when its visible on the X-ray, bad ones often do. After the treatment is adopted, an evaluator reviews the case to learn about the radiologist's competence. By observing outcomes, evaluators naturally have access to information that was not available ex-ante;

[1] http://faculty.chicagobooth.edu/john.cochrane/research/Papers/phd_paper_writing.pdf

in that interim medical outcomes are realized and new X-rays might have been ordered. A biased evaluator thinks as if such ex-post information had also been available ex-ante. A small tumor is typically difficult to spot on an initial X-ray, but once the location of a major tumor is known, all radiologists have a much better chance of finding the small one on the original X-ray.[2] In this manner, by projecting information, the evaluator becomes too surprised observing a failure and interprets success too much to be the norm. It follows that she underestimates the radiologist's competence on average.

The comparative statics of this logic has consequences to skill assessment in labor markets. When corporate boards decide to increase the scrutiny of CEOs, or governments decide to investigate health professionals in more detail, these principals will consistently be disappointed. Even if reputation risk in theory is now lower, experts' ex-ante incentive to invest in their skill or the relationship is reduced in my model. If lower assessments imply greater CEO turnover, or a greater willingness to regulate health professionals, periods characterized by finer monitoring technologies will result in excess turnover or excess regulation.

Information projection does not imply, however, that evaluators underestimate experts both after a success and after a failure. If the interim medical outcomes *complement* the skillful reading of the X-ray, performance differences due to luck are systematically misattributed to differences in talent; successful agents here receive too much credit. If easy-to-understand ex-post information *substitutes* for the skillful reading of the ex-ante radiographs, differences in skill are misattributed to differences in luck and evaluators might be too forgiving after a failure.

The most severe consequences of information projection happen when, as suggested by the evidence from law and medicine, experts anticipate biased evaluations and respond to them strategically.[3] An otherwise fully-concerned and risk-neutral radiologist distorts the production of tests ex-ante in two opposite directions. If an ex-ante test reveals information that is a substitutes of what evaluators independently learn ex-post, physicians will produce this test even in medically unwarranted situations. Absent the information that only an overly costly and painful biopsy can provide, the radiologist's perfect ex-ante recommendation will often appear incompetent in hindsight. At the same time, physicians will stay away from ordering ex-ante efficient tests if they complement evaluators' ex-post information. A noisy mammogram is often the best way to detect breast cancer at an early stage. Looking at it ex-post, however, typically allows the evaluator to determine with certainty whether a tumor was already developing ex-ante. This ex-post insight is often impossible

[2] In his testimony to the US Senate Committee on Health, Education Labor and Pensions, the radiologist Leonard Berlin (2003) illustrates the room for such ex-post wisdom in medicine. Based on empirical studies conducted in prestigious US medical institutions, he argues that in hindsight as many as 90% of lung cancers and 70% of beast cancers can be observed on radiographs previously read as normal. http://www.fda.gov/ohrms/dockets/ac/03/briefing/3945b1_05_Berlin%20testimony.pdf

[3] Indeed in medicine, the fear from the '*retrospectroscope*' refers exactly to such recognition.

in the absence of the ex-ante mammogram. Ironically, the radiologist will enjoy a higher reputation if he does not produce it ex-ante. Finer performance measurement exacerbates defensive practices.

The above agency conflict does not exist under Bayesian assumptions. Incentives that correct for agency conflicts that arises due to physician risk-aversion or due to the fact that physicians do not internalize the costs or benefits of information production are no effective means to reduce defensive practices. Such incentives will reduce distortions in the volume or in the skill-composition of the ex-ante produced information, but do not directly change physicians' incentives to distort the composition of ex-ante information in the above manner. To reduce defensive medicine, differential incentives for the production of complement and substitute information are necessary in my model.

Comparative static properties of the results also provide a rationale for the asymmetric allocation of monitoring resources. Simple and non-manipulable performance measures that commit to more intensely monitor particular outcomes can alleviate physicians' incentives to engage in defensive practices. I show that such asymmetric career assessments, but not symmetric ones, can significantly reduce the incentives for defensive practices in the biased case without introducing production distortions in the unbiased case.

The predictions of the model are consistent with evidence from medicine. There it has been repeatedly argued that it is doctors' fear of the ex-post exaggerations of the ex-ante accuracy of tests which is the prime motivation for 'defensive medicine' defined as medical practices adopted to minimize false liability, rather than maximize cost-effective health care. David Studdert et al. (2005) show that to protect their reputation a vast majority of physicians engage simultaneously both in over-production and under-production of skill-intensive medical tests. Daniel Kessler and Mark McClellan (1996, 2000) show that weaker career concerns have changed the composition of diagnostic procedures in a way that lowered medical costs but not the over-all quality of care.

In Section 6, I show that similar results hold when considering a setting with moral hazard instead of uncertain reputation. I find that in my setting correctly anticipating biased evaluations under limited liability decreases physicians' incentives to exert care. Similar to the case of adverse selection, finer monitoring will often decrease an agent's incentive to invest in the relationship. The returns to investing in the relationship are decreased. Adopting Bayesian superior incentive schemes that monitor agents more closely often backfire. For example a transition from a noisy strict liability scheme to a more informative negligence scheme decreases pre-caution and overall welfare.

Principals who understand that evaluators suffer from projection bias might be able to restructure incentives and restore the unbiased second-best. Indeed, a principal who anticipates the bias will correctly predict the probability with which evaluators recommend wrong actions. Information projection, however, introduces *noise* into monitoring because information projection causes not only mistaken punishments but also mistaken rewards. As a consequence, the classic trade-off between rent extraction and incentive provision is distorted. Hospitals do best when they use coarser performance measures and induce lower levels of care than the Bayesian optimal. This saves on incentives that are appropriate in the Bayesian case, but are too strong in the biased one.

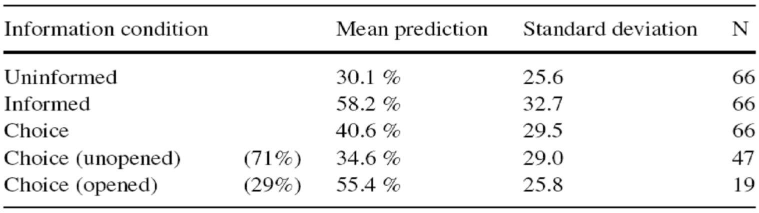
In Section 7, I briefly turn to an application of the model to communication. Information projection causes experts to send messages that are too ambiguous for their audiences to understand. Receivers conform too much to the advice they receive. Protocols that mandate communication, but restrict the use of messages that complement the speaker's expertise, are those that dominate free-form communication. These protocols reduce miscoordination and mistaken favoritism in organizations.

In Section 8, I conclude the paper. I address how the model can be extended to provide a model of limited perspective taking in social judgments, Jean Piaget and Bärbel Inhelder (1967), which might help understand the presence of a representativeness heuristic, as discussed by Amos Tversky and Daniel Kahneman (1974), in this domain. I consider further applications as well as some limitations of my approach.

2 Evidence and Related Literature

This section presents both lab and field evidence from a variety of domains. Evidence on the curse of knowledge, Colin Camerer, George Loewenstein, and Martin Weber (1989), Elizabeth Newton (1990), interpersonal hindsight bias, Baruch Fischhoff (1975), illusion of transparency, Thomas Gilovich et al. (2000) and false-beliefs provides such support in simple social judgements. Studies with legal and medical professionals who solve tasks they are thoroughly familiar with offer further support. The general theme of this section is that people who privately observe pieces of information, underestimate the importance of these pieces shaping their beliefs. Hence when they estimate what others should believe without these pieces, they report estimates that are systematically biased towards their private information. Although individual studies can often be subject to alternative interpretations the sum-total of the studies provides a compelling case for this interpretation.

Informational differences are key in the context of communication. In a striking study, Newton (1990) randomly assigned subjects at Stanford to be tappers or listeners, and presented them with 25 well-known songs. Tappers had to privately pick one of the songs and tap out its rhythm. Listeners then guessed the song based on the rhythm tapped. Out of 120 songs, only three (2.5%) were identified correctly. When tappers were asked to predict these odds after picking their songs however, the mean prediction was around 50%. In a context people should be familiar with, they projected their knowledge of the songs and overestimated the true faction of correct guesses twenty-



Information condition		Mean prediction	Standard deviation	N
Uninformed		30.1 %	25.6	66
Informed		58.2 %	32.7	66
Choice		40.6 %	29.5	66
Choice (unopened)	(71%)	34.6 %	29.0	47
Choice (opened)	(29%)	55.4 %	25.8	19

Figure 1: From: George Loewenstein, Don Moore, and Roberto Weber (2006).

fold. Similarly strong overestimation has been documented in a great number of published studies, for example, Chip Heath and Nancy Staudenmayer (2000), Boaz Keysar and Ann Henly (2002), Justin Kruger et al. (2005).

While information projection is a very plausible explanation of the above results, people may simply overestimate their ability to communicate. Yet the same effect is demonstrated in various other domains without communication. In a fully incentivized study, Loewenstein, Don Moore, and Roberto Weber (2006) presented business students at CMU with logical and visual tasks. In one set of tasks, subjects first saw two pictures that differed in one important detail. Subjects were divided into three groups: uninformed who received no further information, informed who were told the difference, and subjects who could choose to learn the difference for a fee. In all treatments, subjects had to guess the fraction of people in the *uninformed condition* who would correctly identify the difference. Subjects were paid for the accuracy of their predictions.

The true fraction was 20%. As Figure 1 indicates, informed subjects greatly overestimated this fraction relative to the uninformed, and a significant share of subjects paid to learn the difference, which then pushed their estimates further away from the truth (55% versus 30%). Similar results were found on the logical puzzles. Since more information should help at least on average, informed estimates should have been closer to the truth. In this experiment, people not only projected their superior information, but paid for information that biased their judgements and systematically lowered their earnings.[4]

In a similar experiment, raters who knew the solution to a word puzzle estimated whether solvers would figure out the solution, Emily Pronin, Carolyn Puccio, and Lee Ross (2002). Here raters significantly overestimated the likelihood of success (83% as opposed to 21%), and attributed the difference to raters' low skill at puzzles.[5]

Camerer, Loewenstein, and Weber (1989), henceforth CLW, provide further careful evidence. MBA students from Wharton and Chicago traded assets via a double-oral auction. In the first

[4] To control for curiosity, LMW told subjects that they would learn the solution to the logical and visual puzzles at the end of the experiment.

[5] Heath and Staudenmayer (2000), who replicated Newton's experiment, found that over 40% of tappers attributed the surprisingly low success rates of listeners to the listeners' lack of effort while listening to the tapping, but not to the difficulty of the task.

group, traders learned the past performance of the traded companies and returns on trading where determined by the actual earnings of these companies. In the second group, traders also received the actual earnings, but here returns were determined by the market price established by the first group. The results showed that in the second group the market price was biased by 30% towards the actual ex-post earnings. Individual judgements were biased by 60% . Though both largely significant, traders with a smaller bias traded more aggressively, reducing the bias in the market by acting as if they anticipated the bias of others.

The most extensively documented form of information projection is interpersonal *hindsight bias*. The first systematic demonstration of this fact is due to Fischhoff (1975). In a between-subject design, Fischhoff (1975) showed that reporting the outcome of an uncertain historical event changes the perceived ex-ante likelihood of the reported outcome occurring. A large literature following Fischhoff's study has shown that people report systematically biased estimates exactly in this direction of the hindsight effect. Rebecca Guilbault et al. (2004) conduct a recent meta-analysis using 95 studies (83 published and 12 unpublished) and document a very significant average hindsight effect under both objective and subjective uncertainty. Both in the lab and in the field hindsight bias is robust to a great variety of debiasing techniques, e.g., Fischhoff (1982), Lawrence Sana, Norbert Schwartz, and Shevaun Stocker (2002), Erin Harley (2007), including repetition, teaching or explicit warning. Importantly, more ex-post information typically leads to greater absolute mispredictions, e.g., Pamela Hinds (1999).

In the context of performance assessment, Jonathan Baron and John Hersey (1988) demonstrate the hindsight effect. Students at UPenn were asked to rate the quality of thinking that went into ex-ante decisions. Raters saw ex-ante decisions between a sure prize and a risky monetary gamble. A typical choice was getting $100 for sure or facing a 50/50 chance of gaining $220 or $0. Raters also learned however, the ex-post realization of the risky gambles, but were told that realizations were determined by the spin of a balanced roulette wheel after ex-ante decisions were already made. Comparing 160 pairs of ex-ante identical choices, a higher ex-post earning was rated as a more correct ex-ante choice in 60% of cases, as an equally correct choice in 28%, and as a less correct choice in 12%. The same results were true when only forgone, not actual earnings were different.

Information projection is observed in economically much more important choices involving professionals who solve tasks they are thoroughly familiar with. In medicine, Hal Arkes et al. (1981) divided 75 practicing physicians into five groups, gave them the same medical case history, and asked them to assign a probability estimate to each of four possible diagnoses being correct. One group received no additional information, but four "hindsight" groups were told an actual outcome. Nevertheless, all groups were asked to make their assessments based purely on the case history and independently of the diagnosis that turned out to be correct. The hindsight groups that were told that the least likely diagnoses were correct, assigned far greater probability estimates to these diagnoses than did the other groups. Robert Caplan et al. (1991) present similar results using 112 practising anesthesiologists. They show that the difference in ruling ex-ante negligence can be as great as 51% between hindsight and foresight groups. Berlin and Roland Hendrix (1998) and Berlin (2004) summarize similar evidence from radiology. In law, John Anderson et al. (1997) demonstrate the mistake with practicing judges deciding on cases of auditors' liability. Harley (2007) provides an excellent surveys. Jeffrey Rachlinski (1998) discuss normative implications.

A set of other psychological mispredictions indicates that people project various other forms of private information. Gilovich, Victoria Medvec, and Kenneth Savitsky (1998) provide clean evidence that people suffer from what they call the *illusion of transparency* or *spotlight effect* as they greatly overestimate the probability that their emotional and mental states are detected by others, or that their lies, once made, would be discovered.[6] Similarly, Gilovich, Savitsky, and Medvec (2000) show that the average person overestimates the probability that others notice and the probability that others recall her actions and appearances.

The lack of radical informational projection due to the lack of perspective-taking has been emphasized in children since the seminal work of Barbel Inhelder and Jean Piaget (1958) and the false-belief experiments of Simon Baron-Cohen et al. (1985). Excellent recent research by Daniel Bernstein et al. (2004) or Susan Birch and Paul Bloom Paul (2003, 2007) shows that in slightly more complicated versions of the experiments inspired by Piaget's theory, adults exhibit very similar forms of behavior.[7]

2.1 Related Literature

The closest to my paper is CLW (1989) who illustrate their classic experimental findings by employing a partial model of anchored expectations. Bruno Biais and Weber (2008) complete the anchoring approach of CLW to offer a model of intrapersonal hindsight bias. There people correctly remember the variance of their past beliefs, but misremember the mean. In a Gaussian environment this leads to under-reaction to financial news. Biais and Weber test these predictions using psychometric and investment data on investment bankers from London and Frankfurt. In Section 3, I briefly compare anchoring to information projection. Although the two approaches relate to very similar intuitions, and in the case of the experiment of CLW both anchored expectations and information projection explain the results, an anchoring-based account is generally inconsistent with information projection:

[6] Van Boven, Gilovich, and Medvec (2003), study illusion of transparency in bargaining, but their results are harder to interpret.

[7] Susan Birch and Paul Bloom (2003, 2007) show that adults also commit the false belief mistake in tasks that are slightly more complicated that those studied by Baron-Cohen (1985). Daniel Bernstein et al. (2004) show that interpersonal hindsight bias is not significantly diminished from children to adults.

a person who is anchored to his own beliefs would typically violate information projection and vice versa.[8]

In the context of predicting future changes in one's own taste, the phenomenon of projection has been studied by Loewenstein, Ted O'Donoghue, and Matthew Rabin (2003). In contrast to the projection of taste, the projection of information is most relevant in the interpersonal domain and hence it is primarily a social bias. The paper complements a recent literature on limited strategic reasoning in Bayesian games where people predict information differences correctly, but fail to appreciate the extent to which the choices others make are conditioned on the private information others have, Erik Eyster and Rabin (2005), Vincent Crawford and Nagore Iriberri (2007). More broadly, the paper is also related to the growing literature on individual biases in economic decision-making e.g., Camerer (1987) and Rabin and Dimitri Vayanos (2009).

3 Model

Consider an environment where people privately observe signals about the payoff-relevant state $\omega \in \Omega$. There is a finite set of signals N, and a finite set of people M. A signal is a function from the set of states to the set of lotteries over a realization space, $s_j : \Omega \rightarrow \Delta Z$. Signals are interpreted given a common prior σ_0 over the finite set Ω. A person k's information is the set of signals I_k whose realizations she knows. I denote the power set of N by $\mathbb{N}$ and hence $I_k \subset \mathbb{N}$.

To characterize the distribution of information, let $p_k^j \in [0,1]$ denote the initial probability that person k observes the realization of signal s_j. Let us collect these probabilities over signals and across people into a vector $p = \{\{p_k^j\}_{j=1}^N\}_{k=1}^M$. This vector p describes the true distribution of information in this environment. Each sub-vector $p_k = \{p_k^j\}_{j=1}^N$ of p is a probability distribution over $\mathbb{N}$ assigning probability to the event that person k faces a particular information set I_k for all $I_k \subset \mathbb{N}$. In turn, the informational environment can be summarized by the tuple $\Gamma = \{\Omega, \sigma, \{s_j\}_{j=1}^N, p\}$. To conclude the setup, let person k's finite action set be Y_k, and her von-Neumann-Morgenstern utility function $u_k(y, \omega) : Y_k \times \Omega \rightarrow \mathbb{R}$.

3.1 Definition

As long as people have correct perception, the vector p describes people's expectation of how information is distributed. Information projection introduces a *bias* in this perception. A person who suffers from information projection, exaggerates the probability that a signal that is in her information set is also in the information set of others. I introduce a parameter $\rho \in [0,1]$ to express the degree of such information projection.

[8] Prior to my paper, several other papers, with no explicit model, emphasize the importance of hindsight bias for economic applications e.g., Rachlinski (1998) and Camerer and Ulrike Malmendier (2007).

Definition 1 *Person k with information set I_k exhibits information project of degree $\rho > 0$ if her perception that person i's information contains signal j, is given by $p_i^{j,\rho}$ where*

$$p_i^{j,\rho} = (1-\rho)p_i^j + \rho \textit{ if } s_j \in I_k \textit{ and } p_i^{j,\rho} = p_i^j \textit{ if } s_j \notin I_k \textit{ for all } j \in N \textit{ and } i \neq k \tag{1}$$

Like the true p_i, for each information set I_k and ρ, person k's biased perception p_i^ρ defines a probability distribution over $\mathbb{N}$. A biased person differs from an unbiased one in that she exaggerates the probability of those events where others have the information she does and underestimates the probability of those events where others don't have the information she does. In the case of full projection, $\rho = 1$, the biased person believes that all her information is available to others. In the case of partial information projection, $0 < \rho < 1$, she believes that the probability that her information is available to others is between the truth and the full projection case. Finally, when $\rho = 0$, she has correct Bayesian expectations. In effect, as long Alice has has some private information that Bob does not, and not only when she is strictly better informed, her perception is affected.[9]

The degree of projection is uniform in the above definition, an assumption made for notational simplicity only. Generally, the degree of projection is allowed to be heterogenous. Here information projection should be represented by a vector, rather than a scalar. This reflects the fact that different signals might be projected to different degrees. Formally, if ρ_k^j is the degree to which person k projects signal j, then $\rho_k = \{\rho_k^j\}_{j=1}^N$ is person k's generalized degree of projection. All results of the paper extend to heterogenous projection and whenever I refer to an increase in the bias, I mean an increase in *any* component of this vector in a given environment Γ. The claim of the model is only that $\rho_k^j \geq 0$.

The above definition is formulated without explicit reference to time. If $i' \in M$ is the past or future self of person $i \in M$, the definition claims that person k projects her current information onto this past or future self of person i. Finally, while the above definition adopts a simple linear form. In some contexts due to issues of measurability, it might be more appropriate to adopt the definition where $\rho \in [0,1)$ and $p_i^{j,\rho} = (p_i^j + \rho)/(1 + \rho p_i^j)$ if $s_j \in I_k$, and $p_i^{j,\rho} = p_j^i$ otherwise. The key difference here is that if $p_j^i = 0$, then $p_i^{j,\rho} = 0$ for all ρ and $p_i^{j,\rho} < 1$ unless $p_j^i = 1$. More generally, all that matters for the qualitative results and the key property of the model is that $p_i^{j,\rho}$ is continuously increasing in ρ and spans the interval $[p_i^j, 1]$.This fact alone will identify the key consequences of the model. Before I turn to these consequences, let me interrupt the discussion with a simple example.

[9] The model can be extended by allowing the parameters p to depend on the state ω, $p(\omega)$. In this formulation, after observing a set of signals, a Bayesian person forms a posterior estimate $p_i(\omega)$ denoted by $p_i^e(\omega)$. The definition then can be applied to this vector $p_i^e(\omega)$ in the same way as above. The model thus can be interpreted as one where people have heterogenous priors. Importantly though, relative to postulating the existence of heterogenous priors with no theory of the way these priors will be heterogenous, the current model makes clear *directional* predictions on people's conflicting estimates as a function of the true informational environment.

The reader can skip this example without interruption.

3.2 Dinner Example

Consider a dinner invitation from Gremin to Tatiana. Gremin can either prepare fish or meat. If Gremin is kind, his goal is to prepare Tatiana's favorite option. If Gremin is unkind, he only cares about his own taste. While the parties both know their own taste, they are uncertain about the taste and the intentions of the other person. Ex-ante, Tatiana believes that Gremin is equally likely to prefer fish or meat and that independently, Gremin is equally likely to be kind or selfish. Gremin receives some noisy information about Tatiana's taste. In particular, suppose it's common knowledge that this information conveys Tatiana's true preference $\frac{2}{3}$ of the time, and the wrong one $\frac{1}{3}$ of the time. If Tatiana is fully Bayesian, she appreciates this informational asymmetry. If she is fully biased, she believes Gremin knows her taste. The following table summarizes the respective inferences:

	Bayesian posterior, $\rho = 0$	Biased posterior, $\rho = 1$
$\pi_1(\theta_{kind} \mid \text{right dish}) =$	$\frac{2/3+2/3}{2/3+1+2/3} = \frac{4}{7}$	$\frac{1+1}{1+1+1} = \frac{2}{3}$
$\pi_1(\theta_{kind} \mid \text{wrong dish}) =$	$\frac{1/3+1/3}{1/3+1+1/3} = \frac{2}{5}$	$\frac{0}{1} = 0$
$E\pi_1(\theta_{kind}) =$	$\frac{7}{12} * \frac{4}{7} + \frac{5}{12} * \frac{2}{5} = \frac{1}{2}$	$\frac{7}{12} * \frac{2}{3} = \frac{7}{18}$

Tatiana makes two types of inferential mistakes: *over-inference* and *underestimation.* She over-infers kindness when served the meal she likes, and overinfers hostility when served the meal she dislikes. These two do not cancel out, however. Tatiana underestimates the kindness of Gremin on average, $\frac{7}{18} < \frac{1}{2}$. Analogous calculations show that Tatiana underestimates the similarity of her taste to Germin's on average. Given the type-space $\{\theta_{ks}, \theta_{kd}, \theta_{ms}, \theta_{md}\}$, the ex-ante expected Bayesian posterior is $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$. The biased one is $\{\frac{7}{36}, \frac{7}{36}, \frac{7}{36}, \frac{15}{36}\}$.

It is easy to see, that allowing for a sufficient number of repeated interactions, Tatiana and Gremin will depart as foes rather than friends too often. To understand the definition and these result more carefully, let me now turn to the identifying property of the model and its implication to monotone inference problems.

3.3 Projection and the Value of Information

First, I show that if Alice suffers from information projection, she exaggerates the value of information others have. Furthermore, she does so in proportion to how valuable her information would have been to them. Second, I show that in a broad class of monotone inference problems this property implies that Alice underestimates the productive qualities of the actor *on average.* In the rest of this Section, I proceed in a more abstract manner, but illustrate the results and discuss the intuition in Section 4.

Estimating Expected Utility. Consider the expected utility maximization problem of Bob. Let $u^*_{I_B} \in \mathbb{R}$ denote the value of Bob's program when his information is some $I_B \subset \mathbb{N}$. Formally,

$$u^*_{I_B} = \max_{y_b \in Y_b} E[u_b(y_b, \omega) \mid I_B] \tag{2}$$

Note that from Alice's perspective, $u^*_{I_B}$ is a random variable in environment Γ whose distribution is determined by $\underline{p}$. Accordingly, let $f^\rho(u^*) \in \Delta\mathbb{R}$ denote the probability density function, which stands for a ρ-biased Alice's belief about the distribution of $u^*_{I_B}$.[10] Although Alice's exact beliefs $f^\rho(u^*)$ will generally depend on her realized information I_A, the following proposition shows that for all $I_A \subset \mathbb{N}$, a biased Alice overestimates how well Bob can do in expected utility terms.

Proposition 1 *Given vN-M preferences and an environment* Γ*,* $f^\rho(u^*)$ *first-order stochastically dominates* $f^{\rho'}(u^*)$ *if* $\rho < \rho'$*.*

A person who projects private information exaggerates the expected utility of others. The proof of the above result is based on Blackwell's classic theorem on the comparison of information sets. While Blackwell (1953) offers only a partial ordering of information sets, the above proposition shows that given an environment Γ, the misperceptions induced by information projection can be ordered by this criterion. An important corollary of this result robustly identifies the comparative static properties of the model. The more Alice knows, the higher is her estimate of how well Bob can do in his own expected utility terms.

Corollary 1 *Adding information to* I_A *by adding a signal* $s_j \subset N$ *always leads to an increase in* $f^\rho(u^*)$ *in the sense of first-order stochastic dominance if and only if* $\rho > 0$*.*

The above corollary helps understand the consequences of the model. It shows that for any fixed ρ, an increase in I_A leads to an increase in beliefs. Importantly, this fact does not depend on the details of environment Γ. It is true no matter how informational differences are partitioned into signals. Furthermore, note that the increase in beliefs due an increase in I_A is always increasing in ρ and this is true for *all partitioning* of the informational differences in Γ. Note again that this property of the model naturally translates to the case where projection is heterogenous. Adding signals to Alice's information set, or equivalently making her strictly more informed, will always increase her estimate of Bob's expected utility and do so as an increasing function of her bias. Thus the above corollary relies purely on the fact that there are positive informational differences between Alice and Bob.

[10] To be precise, Alice's beliefs should be indexed as $f^\rho_{I_A}(u_{I_B})$. For simplicity, I suppress these indexes.

Inference. Let me now apply the above result to a class of monotone inference problems. Alice is learning about the hidden type θ of Bob by observing some performance realization x from a finite, ordered set X. This performance measure can be understood as a signal of Bob's actions or intentions. The outcome depends on Bob's type which is an element of an ordered set Θ. This type can have many interpretation such as Bob's level of human capital, loyalty, kindness.

The key assumption for the result is that x depends positively both on Bob's type and the value of his information. Formally, let there be a family of real valued conditional density functions, $\{f(x \mid \theta, u^*)\}$, where $f(x \mid \theta, u^*) : \Theta \times \mathbb{R} \to \Delta X$. Adopting the definition of Paul Milgrom (1981), this family satisfies the monotone-likelihood-ratio property given priors $f(\theta) \in \Delta\Theta$ and $f(u^*) \in \Delta\mathbb{R}$ if the following condition is satisfied.

Condition 1 (MLRP) *Let $f(x \mid u^*) = \int_\Theta f(x \mid \theta, u^*) f(\theta) d\theta$, and define $f(x \mid \theta)$ analogously. The family $f(x \mid \theta, u^*)$ satisfies the strict monotone likelihood ratio property in θ and u^* if*

$$f(x \mid u^*) * f(x' \mid u^{*\prime}) - f(x \mid u^{*\prime}) * f(x' \mid u^*) > 0 \tag{3}$$

whenever $x > x'$ and $u^* > u^{*\prime}$, and the analogous condition holds for $\theta \in \Theta$.

Proposition 1 showed that Alice exaggerated the value of Bob's information. In the inference problem where the outcome process satisfies the monotone-likelihood-ration property with respect to u^*, this means that for each type of Bob, Alice expects a higher outcome realization on average than what is warranted. Her expectations about the average realization of x is increasing in her beliefs about the value of Bob's information. i.e., in taking increases of $f^\rho(u^*)$ in the sense of first-order stochastic dominance.

Let $f^\rho(\theta \mid x)$ denote Alice's posterior on θ after observing x when her beliefs about u^* are given by $f^\rho(u^*)$. The next result shows that as long as Condition 1 is satisfied, Alice underestimates Bob's type on average and this underestimation is increasing in the degree of her bias. Corollary 1, then implies that adding signals to Alice's information set, decreases her estimate of Bob's type on average.

Proposition 2 *Suppose the MLRP holds. For all priors $f(\theta)$ and $f(u^*)$, $E_X[f^\rho(\theta \mid x)]$ first-order stochastically dominates $E_X[f^{\rho'}(\theta \mid x)]$ whenever $\rho < \rho'$ for expectations taken with respect to the true distribution of signals.*

Above, I assumed that a higher outcome is good news about quality θ. Since the results depend only on the monotonicity assumption, it follows that when a higher outcome is bad news about θ – in the sense of Milgrom (1981) – a biased observer will overestimate θ on average, and her expected beliefs about θ will be too high. Finally, when x is neutral about θ, no average misestimation is implied by information projection.

3.4 Substitute and Complement Information

Increasing Alice's information increases her beliefs about Bob's utility. Key implications of the model depend, however, not simply on the value of Alice's information, but on how it relates to Bob's information. Given an environment Γ, Alice's exaggeration of Bob's utility is increasing in the value of her private information to Bob.

To pin down the comparative static predictions of the model, a distinction between substitute and complement information turns out to be key. Two images of a bone fracture, one on an X-ray and another on an MRI, are substitutes, if any one of these two is sufficient to establish the fracture. Knowledge of the location and the knowledge of the type of the patient's t tumor a patient has, however, are complements whenever only the combination of these two allows one to identify the best treatment. Importantly, these relationships depend on the objective of the decision-maker.

Holding constant an objective function, two signals are substitutes if knowing both is less valuable than the sum of knowing each separately. They are complements if the opposite holds. If again $u^*(s)$ is the value of a person's expected utility maximization problem given information s, two signals are substitutes if u^* is submodular in them, and complements if u^* is supermodular in them.[11]

Definition 2 *Given $u(\omega, y)$, two signals s_l and s_j are substitutes if $u^*(s_l \cup s_j) - u^*(s_j) < u^*(s_l) - u^*(\emptyset)$ and complements if $u^*(s_l \cup s_j) - u^*(s_j) > u^*(s_l) - u^*(\emptyset)$.*

To conclude the setup, I introduce the possibility that people might correctly anticipate the bias of others. Let the probability density function $\psi_{i,k}(\rho) \in \Delta\ [0, 1]$ describe the beliefs of person i concerning the extent to which person $k \neq i$ projects information. If ψ_i is not concentrated on 0, person i believes that there is a non-zero probability that person k is biased.

3.5 Discussion

The results above do not depend on details of the environment Γ. Specific restrictions, motivated by economic or psychological considerations, might be imposed on the distribution of information, the partitioning of informational differences into distinct signals or on the players' expected utility preferences. These will enrich the set of predictions, but will not violate the results. Given the same actual information differences, a shift in a biased person's belief after an increase in her private information is always an increasing function of her bias.

One binding restriction on Γ is when people's true information is strictly ordered. In such a context, CLW (1989) offer an anchoring-based explanation of the curse of knowledge: a strictly

[11] On the presence of non-concavities in the value of information see e.g., Roy Radner and Joseph Stiglitz (1984).

better informed Alice perceives the *mean* expectation of a lesser informed Bob to be the convex combination of her mean expectation and Bob's true mean expectation. This approach leaves other moments of Alice's belief unspecified and does not address the case where information is not strictly ordered. Even within this class, the example below shows that the exaggeration of the *proximity of two means* is not a measure of informational closeness. Hence no-matter how one completed this anchoring-based account, anchored expectations violate information projection and vice versa.

Example 1 *Let there be three people with strictly ordered information about the return on an asset. 1.* Bob *is uninformed and has a uniform prior over* $[0,1]$. *2.* The White Rabbit *is better informed and receives valuable information by learning that the return is either* 0 *or* $\frac{3}{4}$ *with equal probability.* Alice *learns that the true return is* $\frac{3}{4}$*. The distance between* Alice's *and the least informed* Bob's *mean belief is* $\frac{1}{4}$*. The distance between the* White Rabbit's *and* Alice's *mean beliefs is larger* $\frac{3}{8}$*.*

Anchored expectations by Alice imply neither that she overestimates nor that she underestimates of the value of Bob's information. In fact, by projecting information, Alice might *exaggerate the distance* between her expectation and the expectation of Bob. This often happens when she projects complement information. The current model thus offers not simply a more general framework, but behavioral and welfare predictions that one could not derive and would contradict the predictions one would obtain after any completing of the anchoring based approach.

4 Performance Evaluation

Let's now turn to the main application of the model. Consider a *supervisor* who evaluates the performance of an *agent* whose task is to process and act upon information available to him. When evaluating agents ex-post, supervisors typically have access to information that was not available ex-ante. This creates room for information projection. Section 4 considers the impact of projection bias on inference problems aimed at assessing the agent's talent and allocating resources accordingly. Section 5 introduces implicit career incentives, and considers how agents might respond to biased evaluations. Section 6 extends the setup to the case where explicit incentive contracts are necessary to reduce the cost of moral hazard.

4.1 Setup

First, the radiologist receives an ambiguous *X-ray* about the medical condition of the patient; that is a noisy signal s_0 about ω. Then he adopts a treatment y_a. This later leads either to a success x_S or to a failure x_F.

Competence. The probability that the radiologist understands the *X-ray* depends on his competence**.** Radiologists differ in their competence. Let $\theta \in [0,1]$ denote the radiologist's type which expresses the probability that he understands s_0. More generally, for any fixed set of ex-ante signals I_a, a more competent type always faces a higher probability of understanding any signal in I_a than a lower type. If he does understand a signal, the radiologist's beliefs about ω remain unchanged after reading it.

Technology. The treatment y_a and the medical condition of the patient jointly determine the probability that a success happens. An unrestricted technology matrix A collects these probabilities for all action-state combinations. I assume that all elements of A are strictly between 0 and 1.[12] The radiologist adopts a treatment y_a^* to maximize $u_a(\omega, y)$ given by the probability of a success:

$$y_a^* \in \arg\max y' A \sigma_1 \tag{4}$$

where σ_1 denotes the radiologist's belief over Ω and y' is a $|Y_a|$ dimensional non-negative indicator vector with exactly one positive element. In short, given A, the condition of the patient, the ex-ante available information and the radiologist's competence jointly determine the outcome process: $\pi(x_S \mid \theta, s_0, \omega)$.

Assessment. The evaluator's goal is to assess the radiologist's competence. Since she cannot observe θ directly, she can only update her prior $\pi_0(\theta)$ by observing the agent's performance. To form an assessment $\pi_1(\theta)$ the evaluator observes x along with novel ex-post information s_1 about the patient.

Two important observations are in place 1. Since $\pi(x_S \mid \theta, s_0, \omega)$ depends on ω, observing a success or a failure alone provides novel information about the patient's condition. Outcome information is also information about the task that would have been useful ex-ante. The failure of a medical treatment typically provides novel information about the type of cancer the patient already had ex-ante. In addition, since over time the patient develops new symptoms and new radiographs are also ordered, the ex-post information is further enriched. 2. Since $\pi(x_S \mid \theta, s_0, \omega)$ depends both on θ and ω, the more the evaluator knows about the patient the more valuable her inference about the radiologist's competence should be. If there are no costs to learning ex-post information, under Bayesian assumptions, the evaluator is best-off by always learning the true state ω ex-post.[13]

[12] This assumption plays no role in the results, except that it guarantees that all observations are on the equilibrium path i.e. for all s_0 success and failure has both positive probabilities.

[13] **Example**: To illustrate this property consider a specific task where $Y_a = \Omega$ and let $\lim A =$

$$\begin{matrix} 1 & 0 & 0 \\ 1 & 0.499 & 0 \\ 1 & 0 & 0.5 \end{matrix}$$

and let $\pi_0(\theta) = 1$ for all θ. The prior σ_0 on Ω is fully symmetric, and $\sigma_1$when incorporating $s_0$is given by $(\frac{1}{3}, \frac{3}{6}, \frac{1}{6})$.

First, note that learning that $x = x_F$or $x = x_S$already reveals novel information about ω. Second, note that the evaluator's inference when observing x and $s_0$only is that $\pi_1(\theta \mid x_S) = \frac{6+2\theta}{7}$. In case the evaluator also observes $s_1 = \omega$, her inference is more precise. In particular, $\pi_1(\theta \mid x_S, \omega_1) = 1$, $\pi_1(\theta \mid x_S, \omega_2) = 2\theta$ and $\pi_1(\theta \mid x_S, \omega_3) = 2 - 2\theta$. It is straightforward to calculate the bias assessments.

Career Concerns. For the remainder of this section, it suffices to assume that the radiologist prefers a success to a failure. To further motivate his choice, however, suppose that the radiologist has career concerns: for extrinsic of intrinsic reasons he prefers a higher assessment to a lower one. Formally, let $u_a(\omega, y) = E_{y,\omega}[b(\pi_1(\theta))]$ where $b(\pi_1(\theta)) : \Delta[0,1] \to \mathbb{R}$ and $b(\pi_1(\theta)) \geq b(\pi_1'(\theta))$ whenever $\pi_1(\theta)$ fosd $\pi_1'(\theta)$. One possible specification of $b(\pi_1(\theta))$ is determine the second-period compensation or promotion of the agent based on his expected competence. In this section, as long as assessments are higher after a success than after a failure, y_a^* is given by Eq. (4). In the efficient Bayesian Nash Equilibrium of the game, the agent takes y_a^* and the evaluator forms π_1 according to Bayes rule. As in Holmström (1982) or Holmström and Harris (1982), I assume that the radiologist faces the same uncertainty about his type as the evaluator.[14]

Job Assignment and Skill Acquisition. Skill assessment in organizations serve two distinct roles: (i), they aid the efficient allocation of human resources in internal and external labor markets, and (ii), they provide incentives for physicians to privately invest in their skill. For the second role, the setup could be supplemented with an un-modelled ex-ante stage, where the physician privately decides how much to invest in skill. The individual return on such investment will depend on the quality of performance assessments ex-post. For either of these goals, under Bayesian assumptions, efficiency is increasing in the precision of the evaluator's assessment.[15]

This frictionless learning model of wage dynamics is similar to many in the literature, for example those studied by Henry Farber and Robert Gibbons (1996) and Gibbons and Michael Waldman (1999a). Unlike in signal-jamming models with moral hazard as introduced by Bengt Holmström (1982) and Mathias Dewatripont, Ian Jewitt, and Jean Tirole (1999), the shape of the career concerns, b, and the details of the signalling environment, A, play no role in the current analysis. The lack of immediate moral hazard considerations and that the evaluator can fully observes s_0 distinguishes the setup from the setup of David Scharfstein and Jeremy Stein (1990) and Canice Prendergast (1993). As we will see, the source and the nature of the inefficiency also differs.

4.2 Skill Assessment

Given her perception of the agent's ex-ante information and behavior, the evaluator updates her prior on θ via Bayes' rule. Information projection introduces a bias in this perception. A bayesian evaluator understands that she has information about the patient that was not available before and thus adopts the right perspective on the ex-ante situation. A biased evaluator in contrast projects the novel information revealed along with the realization of x and thinks as if s_1 was also available ex-

[14] Importantly, this assumption plays no role in Section 4. The qualitative results of Proposition 6 will also extend to the many contexts where the radiologist has superior information about his own competence.

[15] For a discussion on the relation between human capital investment and organizational design, starting with Gary Becker (1962), see the survey of Gibbons and Waldman (1999b).

ante. Let π_1^ρ denote the posterior of a ρ-biased supervisor. The following result applies Proposition 2 to the case of binary outcomes.

Proposition 3 *For all $\pi_0(\theta)$, $E_{\omega,x}[\pi_1^0(\theta)] = \pi_0(\theta)$ and $E_{\omega,x}[\pi_1^\rho(\theta)]$ first-order stochastically dominates $E_{\omega,x}[\pi_1^{\rho'}(\theta)]$ whenever $\rho' \geq \rho$ where expectations are taken with respect to the true distribution of signals.*

A biased evaluator overestimates the probability that the radiologist's prediction should have led to a success. She is too surprised observing a failure. She thus puts too much weight on the information revealed by a failure and too little on the information revealed by a success. Since the likelihood of a failure is decreasing in the radiologist's competence, the supervisor underestimates this competence *on average*. Importantly, because the above result is true for all continuous priors underestimation after repeated performance sampling holds a fortiori.[16]

More Scrutiny → Lower Assessments. The comparative static result below shows that the more useful the ex-post information would have been ex-ante, the more skeptical the supervisor will be on average. Given the martingale property of Bayesian beliefs, a finer partition of the ex-post information only helps to form more precise estimates. Under information projection, however, finer performance measures lead to predictably lower posteriors. In this manner, when boards decide to increase the scrutiny of CEOs, or a government decides to investigate the activity of social workers in greater detail, the reputation of these experts suffers.

Corollary 2 *Let $g = E_\theta[u_a^*(s_0 \cup s_1) - u_a^*(s_0)]$, then $E_{\omega,x}[\pi_1^\rho]$ is decreasing in g in the sense of fosd if $\rho > 0$.*

4.3 Relative Performance Evaluation

Information projection distorts absolute performance measures. As a result, the ex-ante incentives for experts to privately invest in their competence is decreased. Importantly, the model has implications not only to absolute, but to the dynamics of relative performance evaluation.

Favoritism. Suppose the supervisor evaluates two workers on independent tasks. Let the information gap for the first task be g_1 and for the second g_2. Suppose $g_1 > g_2$. For instance, there might be ex-ante perfect information on the first task, but only ex-post on the second. Given two equally competent workers, a biased evaluator ranks the one assigned to the first task higher than the one assigned to the second.[17] In the classic account of favoritism in organizations, as introduced

[16] If subscript t refers to the evaluator's assessment after t rounds of observation, under auxiliary measurability assumptions it can be ensured that biased estimates $\pi_t^\rho(\theta)$ comprise a *supermartingale* for all $\rho > 0$. Here, $\lim_{t\to\infty} \pi_t^\rho(\theta \mid \widehat{\theta})$ exists almost everywhere, and in addition, $\lim_{t\to\infty} \pi_t^\rho(\theta \mid \widehat{\theta}) < \lim_{t\to\infty} \pi_t^0(\theta \mid \widehat{\theta})$ for a.a. $\widehat{\theta} > 0$.

[17] Alan Durell (1999) provides related laboratory evidence. In his study, employees were randomly assigned to easy or hard word puzzles. Employers observed the performance of the employees on these tasks and had to predict their future performances. Employers correctly predicted the future performance of those initially assigned to easy tasks, but underestimated it for people initially assigned to hard tasks.

by Prendergast and Topel (1996), distorted rankings are due to an exogenous preference that the evaluator has for some but not all workers. Here instead, systematic favoritism arises as the result of the biased learning process. Furthermore, it is an unintended outcome endogenously determined by how information is distributed within the organization.

4.4 Luck or Talent

Underestimation on average does *not* imply that the radiologist's reputation will be too low either after a success or a failure. Conditional assessments depend on the *kind* of information projected. If the projected information had been more useful for high types than for low types, the evaluator over-infers skill from performance. If the reverse is true, she under-infers skill from performance.

Proposition 4 *Suppose* $u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)$ *is increasing in* θ*, then* $E_\omega[\pi_1^\rho(\theta \mid x_S)]$ *is increasing in* ρ *in the sense of fosd for all* $\pi_0(\theta)$*. Suppose* $u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)$ *is decreasing in* θ*, then* $E_\omega[\pi_1^\rho(\theta \mid x_S)]$ *is decreasing in* ρ *in the sense of fosd for all* $\pi_0(\theta)$*.*

Consider the case where knowing the ex-post market outcomes would have helped only those CEOs who had the competence to understand the ex-ante information about their firms' assets. Biased board-members then mistakenly attribute performance differences to skill rather than luck. Even if only the combination of the ex-ante and the ex-post information has value, understanding the assets alone does not, board members develop an *illusory confidence* in the competence of CEOs after a successful performance.

The opposite misattribution happens when easy-to-understand ex-post information substitutes for the hard-to-read ex-ante information. A proof of a result might turn out to be simple, yet identifying the simple proof often requires great competence. Readers not recognizing how hard the problem was ex-ante, will not appreciate a successful solution enough. If under-inference is a stronger force than underestimation, evaluators will also be too forgiving after a failure.

4.5 Asymmetric Career Assessments I

Limiting the evaluator's access to ex-post information will reduce distortions. Importantly, however, it is typically impossible to separate the observation of x from the observation of some s_1. At the same time, suppressing all, or almost all, ex-post information eliminates the room for any inference biased or Bayesian. Hence it seems that it is impossible to provide positive and undistorted career assessments in the biased case. This is true, however, only if we restrict attention to the symmetric case. I now turn to why asymmetric performance assessments, assessments generated after selectively suppressing ex-post information, can greatly improve the evaluator's inference in the biased case, without distorting long term career incentives in the Bayesian case.

The idea here is that the evaluator's exposure to ex-post information depends on the outcome of the task. To model asymmetric performance measures, suppose a pre-programmed machine designs the evaluator's information set. The machine has two characteristics: it is *non-manipulable* and *simple.* These assumptions mean that the machine cannot lie, but can either (i) suppress all ex-post information, let's call this the blank outcome, or (ii) show all ex-post information. The machine is simple when the output of the machine only depends on x. A simple machine thus does not need to interpret A, s_0 or s_1. A machine is non-manipulable if it never lies.

Definition 3 *A simple asymmetric performance measure is described by a parameter pair* $\{m_S, m_F\}$ *where* $m_S, m_F \in [0,1]$. *It inputs the realization of* $\{A, s_0, x, s_1\}$. *If* $x = x_S$, *it outputs* $\{A, s_0, x_S, s_1\}$ *w.p.* m_S *and* $\{A, s_0\}$ *w.p.* $(1 - m_S)$. *If* $x = x_F$, *it outputs* $\{A, s_0, x_F, s_1\}$ *w.p.* m_F *and* $\{A, s_0\}$ *w.p.* $(1 - m_F)$.

The argument below relies on the realistic assumption that when simply inferring the outcome x, rather than directly observing it, the evaluator has less exposure to ex-post information. To simplify the analysis, I consider the limit case and assume that inferring x carries no ex-post information about ω. I maintain the assumption that observing x is still only possible along with observing the realization of novel productive s_1 at the same time.

Condition 2 *Suppose that* $\pi(\omega \mid x_F, s_0) = \pi(\omega \mid x_S, s_0)$ *for all* ω *and* s_0.

To further justify Condition 2, outputting $\{A\ ,s_0\}$ could be interpreted as outputting the technology A and the ex-ante information $s_0(\omega)$ but not its realization in Z. This typically further limits inference about ω upon inferring x. Importantly, as long as there is no updating about ω without the observation of the realized s_1, the results do not depend on whether the realization of s_0 is observed or not.

Since the machine cannot lie, the expected assessment of a Bayesian evaluator is independent of the value of $\{m_S, m_F\}$.By knowing m_S and m_F, the evaluator knows the extent to which the blank outcome is good news or bad news about the radiologist's competence. The noise in her assessment will depend on m_S and m_F, but her estimates will always be correct.

Lemma 1 *In the Bayesian case,* $\rho = 0$, *the ex-ante expected posterior* $E_{\omega,x}[\pi_1(\theta)] = \pi_0(\theta)$ *is independent of* m_S *and* m_F *for all* π_0. *Welfare is maximal with unlimited inference,* $m_S^* = m_F^* = 1$.

If $\rho > 0$, symmetric machines cause underestimation. Key to our analysis is that information projection implies that assessments are distorted in *different* ways after a success and after a failure. This implies that asymmetric machines will change expected inference. While more monitoring always means lower expected assessments, more monitoring after a failure or more monitoring after a success does not imply the same.

To focus on the simplest possible case, suppose that the additional value of s_1 is the same for all types. By Proposition 4, there is no over- or under-inference. Here only underestimation is present, and assessments are correct after a success and too low after a failure. Letting the evaluator observe the ex-post information only when $x = x_S$ implies correct assessments both conditionally and on average.

Proposition 5 *Suppose $u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)$ is constant in θ and Condition 2 holds. For all $\rho > 0$ and π_0, $E_{\omega,x}[\pi_1^{\rho}(\theta)]$ is decreasing in m_F and unchanging in m_S. $E_{\omega,x}[\pi_1^{\rho}(\theta)] = \pi_0(\theta)$ if and only if $m_S \geq 0$ and $m_F = 0$.*

Above more focus on success than on failure improved assessments. This institution combined limited inference with asymmetric performance assessment. Limited inference decreases short-term efficiency in the Bayesian case, but under information projection, it helps debias assessments. Similar result holds when over-inference or under-inference is also present. I return to a more extensive discussion of asymmetric performance measures and the interpretation of these results when discussing incentives for information production in the next section.

5 The Supply of Information

I now turn to the *key* application of the paper and consider the case where experts anticipate biased evaluations.[18] When experts respond strategically, an agency problem not present under Bayesian assumptions arises. I describe the consequences of this conflict and suggest specific ways to mitigate them. Note first though that in my model, conditional on accepting a task, both the radiologist who anticipates the bias, and the one who does not, adopt the same treatment and maximize the ex-ante probability of an ex-post success.

Lemma 2 *Suppose $u_a(\omega, y) = b(\pi_1^{\rho}(\theta))$, then for all $\psi(\rho)$, the agent's best-response is given by Eq. (4).*

Even though the treatment choice is independent of $\psi(\rho)$ the task choice of the radiologist depends on whether she anticipates biased evaluations or not.

[18] As mentioned in Section 2, evidence shows that people often anticipate information projection-based mispredictions. Furthermore, people appear to display a strong asymmetry when thinking about the fact that others might be biased versus assessing the fact that they themselves are biased. For an illustrative study, see Pronin, Gilovich and Ross (2004).

5.1 Defensive Practices

Suppose the radiologist has no direct choice over what task to undertake, but has discretion over what ex-ante information to order on his task. A radiologist who fears information projection will have non-standard preferences over such information. To understand these, suppose the radiologist can decide whether, on top of s_0, to produce an additional radiograph, s_0', ex-ante or not. The social value of producing s_0' is given by $\mathbf{a} \in \mathbb{R}$. It is valuable to produce s_0' iff $a \geq 0$. Benefits of production include the additional knowledge gained, costs include the alternative use of medical resources, delay in treatment, increased pain or radiation. The radiologist privately observes a and decides to produce s_0' before taking action y_a.[19]

To assume away all direct agency conflicts, I assume both that the radiologist fully internalizes a and that he is risk-neutral over assessments π_1. I relax both of these assumptions in the discussion below. I assume that the production choice of the expert is observable to the evaluator. The agent's augmented objective is now:

$$u_a(y, a, \omega) = \chi a + b(\pi_1^{\rho}(\theta)) \tag{5}$$

where χ equals 1 if s_0' is produced and 0 otherwise and by the assumption of risk-neutrality.

Let m be the probability that an evaluation occurs after task completion. For now assume that m is independent of the outcome. In the Bayesian case, the agent always produces s_0' whenever it is socially optimal: his ex-ante expected reputation does not depend on what the set of ex-ante signals is. In contrast, an agent who fears the underestimation triggered by information projection *over-produces* tests that are *substitutes* of, and *under-produces* tests that are *complements* of the information the evaluator independently learns ex-post.

Proposition 6 *Suppose the supervisor observes the production of s_0'. For all π_0, s_0,s_1 the agent's best response is given by a cut-off strategy $a(m,\rho)$ where s_0' is produced if and only if $a \geq a(\rho, m)$. Furthermore,*

1. *For all ρ and m, $a(0, m) = a(\rho, 0) = 0$.*
2. *If s_0' and s_1 are substitutes, $a(\rho, m)$ is increasing in m and $\rho > 0$.*
3. *If s_0' and s_1 are complements, $a(\rho, m)$ is decreasing in m and $\rho > 0$.*

The common force driving these opposite responses is the desire to reduce the information gap between the ex-ante and the ex-post stages. An expert who bases his prediction on an overly costly and medically unwarranted MRI enjoys a higher ex-post reputation if this test provides information that evaluators will independently learn ex-post. More surprisingly, the opposite happens when

[19] Importantly, the qualitative results of Proposition 6 will continue to hold when the amount of inference about θ is held constant and the radiologist has superior information about his own type.

the ex-ante test complements the ex-post information. Consider a social worker who can make a valuable, but still ambiguous, phone call to a foster family. Suppose evaluators will always learn whether physical harm has happened to the child, but this will only be informative of child abuse once combined with the information that only an ex-ante call can provide. To limit such ex-post insights, the social-worker is better-off not making the phone call ex-ante.

In short, as performance assessment happens more frequently, the *composition* of ex-ante information shifts from complement to substitute signals. This shift reduces the quality of the medicine practiced. Furthermore, if defensive practices cannot fully eliminate underestimation, the average reputation of physicians also drops.

5.2 Discussion and Evidence

Risk aversion versus Fear of Information Projection. I assumed risk-neutrality over assessments, but the same qualitative results hold under risk-aversion. This is true because information projection affects the mean rather than the spread of the reputational lottery. In fact, risk-aversion will typically amplify both the over-production of substitute and the under-production of complement *skill-intensive* information. As noted by Holmström (1982), risk averse agents have a preference to limit the production of skill-intensive information. In contrast, a risk-averse radiologist who fears underestimation will display differential responses, and production inefficiency arises even when optimal risk-sharing is not an issue and the amount of inference about his skill is held constant.

Production Incentives Radiologists might not fully internalize the costs or the benefits of producing s_0'. This poses a moral hazard problem where respectively stronger and weaker production incentives are required. Importantly, stronger or weaker production incentives affect the volume but not the composition of the information produced. The radiologist still enjoys a higher expected utility by distorting production in the above manner. Similarly, inflating or deflating the evaluator's assessments will not limit defensive practices because for any fixed inflation or deflation policy the radiologist will again distort the ex-ante composition of tests. Differential incentives for the production of substitute and the production of complement information is necessary.

Defensive Medicine. The above results are supported by stylized evidence on defensive medicine. Studdert et al. (2005) interviewed physicians in six high-risk medical fields in the state of Pennsylvania. Of the 824 physicians interviewed, 93% reported that they engaged in defensive medicine. 90% of the respondents reported that they engage in assurance behavior: they order diagnostic tests in medically unwarranted situations. The fraction of physicians who reported frequent avoidance of various efficient technologies was greater than the fraction that reported that they rarely or never avoid efficient practices. As an example, 54% of radiologist reported that they often avoid ordering medically efficient mammograms and 36% reported of the radiologist reported ordering unnecessary MRIs.

Supportive of my setup, Kessler and Mclellan (1996) argue that the motivation for defensive medicine is not pecuniary loss, but rather fear of reputational loss, and show that reducing liability pressure significantly reduces medical expenditures without worsening mortality or medical complications.[20] Supportive of the mechanism of information projection, Kessler and McClellan (2000) find that the main effect of cost reduction is on diagnostic rather than on therapeutic practices. Although defensive practices are sometimes attributed to physicians' fear of random judicial judgements, if the cause of false liability is not random judgement but information projection, as often informally argued in the medical literature, e.g., Berlin and Hendrix (1998), Berlin (2004), increased efficiency should operate through an observable change in the composition of the diagnostic practices as specified by Proposition 5.

5.3 Asymmetric Career Assessments II

Above, I described the consequences of the agency conflict caused by information projection under symmetric career assessments. Let's now turn to asymmetric assessments. As we have seen, in the fully Bayesian case, expected asymmetries leave expected reputations unchanged and hence there production incentives are not affected. Asymmetric assessments, however, offer important additional insights in the biased case as they can both exacerbate and alleviate workers' incentives to engage in defensive practices.

To simplify the presentation, I assume that given s_0, the additional value of having s_1 is the same for all types. Relaxing this condition will complicate notation without changing the key insights.

Condition 3 *$u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)$ is constant in θ.*

Consider first the case where the additional value of having s_1 is also the same for all types given $s_0 \cup s_0'$. Here, increasing m_F again leads to more assurance and avoidance behavior just as we observed after an increase in m before. The following corollary of Proposition 5 shows that suppressing ex-post data after a failure, but revealing it after a success, restores efficient information production and also allowing for positive career incentives at the same time. If the evaluator only observes the ex-ante data when the outcome is a failure, $m_F = 0$, her perspective on the tasks is correct both when the machines outputs the blank state and when it outputs the full information state (ii). As long as $m_S > 0$ the evaluator can still learn about the agent's competence.

Corollary 3 *Suppose $u_a^*(s_0 \cup s_0' \cup s_1 \mid \theta) - u_a^*(s_0 \cup s_1 \mid \theta)$ is constant in θ. Then $a(m, \rho)$ is increasing*

[20] As Kessler and McClellan (1996) argue since virtually all fully insured against the financial costs of malpractice such as damages and legal defense expenses, defensive practices result from the fact that physicians "may employ costly precautionary treatments in order to avoid nonfinancial penalties such as fear of reputational harm, decreased self-esteem from adverse publicity".

in m_F if s_0' and s_1 are complements, and decreasing in m_F if they are substitutes. Furthermore, for all ρ and s_0', $a(m, \rho) = 0$ if and only if $m_S^* \geq 0$ and $m_F^* = 0$.

The above logic generalizes to cases where projected information might have more additional value to some types than others. Importantly though, typically different asymmetries will be necessary for decreasing avoidance and for decreasing assurance behavior. The key in both cases is to ensure that the level of the expected reputation does not vary with the set of ex-ante information.

Consider first avoidance behavior and focus on the case where the projected information would have been more useful for high types. Here producing the complement s_0' causes assessments to be too high after a success but too low on average.Given our assumptions, focusing on success alone will overshoot and lead to the overproduction of s_0'. To decrease incentives some focus on failures will also be necessary. The machine that can restore efficiency will still allow only for limited inference, because more attention has to be devoted to success than to failure. The next proposition also shows that a similar result is true in the case of avoidance behavior and under-inference.

Proposition 7 *Suppose $u_a^*(s_0 \cup s_0' \cup s_1 \mid \theta) - u_a^*(s_0 \cup s_1 \mid \theta)$ is increasing in θ. If s_0' is a complement of s_1, then $a(m, \rho)$ is decreasing in m_S and increasing in m_F for all $\rho > 0$. Furthermore, for each ρ there exists positive $m_S^*(\rho) > m_F^*(\rho) > 0$ such that $a(m^*(\rho), \rho) = 0$. The same is true if $u_a^*(s_0 \cup s_0' \cup s_1 \mid \theta)$ $-$ $u_a^*(s_0 \cup s_1 \mid \theta)$ is decreasing in θ and s_0' is a substitute of s_1.*

Two facts are important to note here. First, carefully designed asymmetric performance measures can improve information production. Second, the comparative static implications of asymmetric assessments depend crucially on whether s_0' is a substitute or a complement of the ex-post information.

Above, career concerns were biased towards success when determining the best way to alleviate defensive practices. It does not follow that behaviorally optimal asymmetries will always exhibit this feature. Limited inference helps if it inflates expected assessments conditional on avoidance behavior and deflates assessments conditional on assurance behavior. Indeed in the case of avoidance behavior, where production of s_0' would increase underestimation, the most effective asymmetric performance measures inflate biased assessments. In the case of assurance behavior, where production of s_0' would decrease underestimation, the most effective asymmetric performance measures deflate biased assessments. For example, in the case of assurance behavior if more information would have helped lower types more than higher ones, assessment asymmetries that are biased towards failure will mitigate incentives for over-production. often be welfare improving.

Organizational Design. Simple non-manipulable machines will not always be able to fully eliminate production inefficiencies. For example, in the presence of severe under-inference and complement signals, there is no m other than $m_S^* = m_F^*$ such that $a(m, 1) = 0$ can be implemented.[21] Selectively suppressing information, however, can always improve the efficiency of production. In the Bayesian case, this implies efficient production but weaker career incentives on the short run. In short, non-manipulable and limited inference is often the compromise necessary to limit distortions when some evaluators are biased while maintaining production efficiency in the unbiased case. The model makes environment specific predictions on whether a greater focus on success or failures is necessary.

6 Moral Hazard

Let's now turn from a setting with reputation to one with moral hazard. In the language of the medical example, assume that it is now the radiologist choice of care in evaluating ex-ante radiographs rather than his ability which determines performance. If the radiologist fails to fully internalize the benefits of careful information processing, a classic agency conflict arises. An ex-ante commitment to evaluate the radiologist ex-post is necessary to provide incentives. The purpose of the discussion here is to show the extent to which the insights from a reputational setting translate to one with explicit moral hazard.

I show that as in the case of career incentives and the return to investment in skill, the radiologist's return on exerting care is reduced when ex-ante negligence is judged by biased evaluators ex-post. Furthermore, more precise monitoring often backfires and reduces care. Since at best information projection introduces noise into evaluations, optimal monitoring will be coarser and performance incentives will be weaker than in the Bayesian optimum.

Formally, after receiving s_0, but before selecting y_a, the radiologist privately decides how much care, $e \in \mathbb{R}^+$, to exert. This determines $p(e)$ the probability that the radiologist understands s_0. I assume that there are decreasing returns to care, $p'(e) > 0$ and $p''(e) < 0$, and the limit conditions hold, $\lim_{e \to 0} p'(e) = \infty$ and $\lim_{e \to \infty} p'(e) = 0$. To focus on the key insights, I assume a simple technology where if $y_a = \omega$ the chance of a successful treatment is k and if $y \neq \omega$ this reduces to $z < k$. Furthermore, the ex-ante probability that s_0 conveys the correct state is h. Thus upon understanding the X-ray, the radiologist's chance of a success is $hk + (1 - h)z$. If he does not understand the X-ray, the probability that he takes the right action reduces to some number d.

Both the radiologist and the principal are risk neutral. The radiologist's payoff is his wage minus his effort, $u_a(w, e) = w - e$. The principal's payoff is the 1 if the $x = x_S$, 0 if $x = x_F$ net the wage paid to the radiologist: $v(x, w) = x - w$. Finally, the radiologist is protected by ex-post limited liability and can not receive a negative wage. The incentive setup is similar to the efficiency wage

[21] More complicated non-manipulable machines that condition choices on observables other than x might enhance production efficiency in these cases.

analysis of Carl Shapiro and Stiglitz (1984).[22]

6.1 Bayesian Monitoring

The socially optimal first-best care e_{fb} equates marginal cost and marginal benefit and is defined implicitly by $qp'(e_{fb}) = 1$, where $q = (h - d)(k - z)$ is the productivity gain from processing information. To save on notation, let the four technology parameters be summarized by vector $\underline{q}$. To eliminate unstable equilibria, assume that $p'''(e) \leq 0$.

Consider first the case where the evaluator only observes x. Consistent with the more general results of Innes (1990), the optimal incentive scheme here is a strict liability one which offers $w_F = 0$ after a failure, and an efficiency wage w_S after a success. The level of the efficiency wage, $w_{strict}(\underline{q})$, and the optimal care, $e_{strict}(\underline{q})$, is determined by the solution of the principal's problem:

$$\max_{e,w_S} v(x,w) = [p(e)q + dk + (1-d)z](1 - w_S) \tag{6a}$$

$$\text{such that } e_{strict}(\underline{q}, w) = \arg\max_{e}[p(e)q + dk + (1-d)z]w_S - e \tag{6b}$$

Suppose now that instead of x, the evaluator can decide to observe the ex-ante diagnosis y_a and the ex-ante information s_0.[23] Given s_0 and y_a, it is now possible to write a more efficient *negligence contract*.This rewards the expert if his action matches the ex-ante information, and punishes him otherwise. This scheme filters out noise, due to the fact that $k < 1$ and $z > 0$, and ties compensation closer to the unobservable care as expressed by the new incentive constraint faced by the agent:

$$e_{negl}(\underline{q}, w) = \arg\max_{e} p(e)(1-d)w_S + dw_S - e \tag{7}$$

For any given efficiency wage, increasing the observability of the radiologist's activity increases the radiologist's incentives to exercise care. The trade-off between providing incentives and extracting a positive rent from the radiologist is improved. Let the optimal solution of the principal's problem under this negligence contract be $e_{negl}(\underline{q})$ and $w_{negl}(\underline{q})$. The negligence contract with monitoring induces higher care and provides higher welfare.

Lemma 3 *For all* $\underline{q}$, $e_{strict}(\underline{q}) < e_{negl}(\underline{q}) < e_{fb}(\underline{q})$ *and* $Ev(e_{negl}) > Ev(e_{strict})$.

6.2 Biased Evaluations

The presence of information projection biases an evaluator's interpretation of the ex-ante $X-$ray towards the ex-post state. This causes her to commit two errors: a type I and a type II. If ex-

[22] See also the classic work of David Sappington (1983) and Robert Innes (1990).

[23] Note that $\{y_a, s_0\}$ is not always a sufficient statistic for $\{y_a, s_0, x\}$ and hence efficiency could be improved both in the Bayesian case and in the case where the principal correctly predicts the evaluator's bias. Considering this possibility, does not change the qualitative results, but would slightly complicate notation.

post the X-ray has a different correct interpretation than ex-ante, an correct ex-ante interpretation is misclassified as an incorrect one. This leads to mistaken punishment. At the same time, an incorrect ex-ante interpretation is sometimes classified as a correct one. This leads to mistaken reward. Importantly, the probability that the biased evaluator commits the type I error increases as the function of the care exerted by the radiologist.

Formally, the probability that a type I error occurs is $p(e)(1-h)$. No care implies no mistaken punishments, the highest possible care implies that the frequency of mistaken punishment converges to the amount of the ex-ante uncertainty. A radiologist who anticipates biased evaluations will anticipate both of these errors. His best response to a negligence scheme with any wage w_S is given by:

$$e^{\rho}_{negl}(\underline{q}, w) = \arg\max_{e} p(e)(h-d)w_S + dw_S - e \tag{8}$$

Note, that the return on care is diminished relative to the unbiased case. Two related effects are present. First, unless $p(e) = 0$, the frequency with which the agent is mistakenly punished is higher than the probability that he is mistakenly rewarded. This will diminish his earnings given any fixed wage w_S. Second, reward and punishment is less closely correlated with effort. Thus, a physician who anticipates the bias lowers the care he exerts in response.

Proposition 8 *The care under negligence and biased evaluations $e^{\rho}_{negl}(\underline{q}, w) < e_{negl}(\underline{q}, w)$ for all $w > 0$. Furthermore, $e_{negl}(\underline{q}, w)$ is constant and $e^{\rho}_{negl}(\underline{q}, w)$ is strictly increasing in h.*

Key to the above logic is that by misperceiving the ex-ante situation the quality of the evaluator's judgement suffers.[24] This reduces the return to exerting care and the efficiency gain from adopting the negligence scheme is reduced. A corollary of this result is that if the amount of uncertainty at the ex-ante stage is sufficiently high, the transition from strict liability to the Bayesian-optimal negligence actually *backfires.* When in theory physicians are expected to increase care, they find it best to decrease it in practice. A Pareto-inferior outcome with lower levels of employment results.

Corollary 4 *If $h < h^*$, then $e^{\rho}_{negl}(\underline{q}, w_{negl}) < e_{strict}(\underline{q}, w_{strict})$ for all $\underline{q}$.*

The logic of the above result is that in the Bayesian case monitoring improves the trade-off between incentive provision and rent extraction and this allows the principal to induce greater care at a lower wage. In this biased case, the Bayesian optimal wage might be too low because the returns to exerting costly care are diminished. If the problem were simply that the radiologist is punished

[24] In a setting with no moral-hazard and perfect observability, Rachlinski (1998) informally argues that under negligence workers will take excessive precautions to avoid ex-post blame. This reasoning will clearly depend, however, on the associated cost of additional care and the benefit of avoiding mistaken rewards or punishment.

too often, there might be a simple way to restore the second best. A principal who anticipated the evaluator's bias could simply increase the efficiency wage until he restores the unbiased second-best. This, however, is not true because besides the fact that punishment occurs too often there is an additional effect.

A principal who knows the evaluator's bias perfectly predicts the probability that the evaluator submits a mistaken report. The principal only knows this probability, but not whether the actual report is right or wrong. When biased evaluations also lead to false rewards, $d > 0$, information projection introduces *noise* in the monitoring process. This distorts the trade-off between incentive provision and rent extraction and decreases the return to monitoring. Recognizing this, the principal decides to adopt monitoring less often, and even when he does so, he induces lower effort than under Bayesian conditions. This way, the principal saves on overall pay to the radiologist which is appropriate under Bayesian monitoring but too high under information projection.

Proposition 9 *The solution to the principal's problem when the evaluator's bias is common knowledge between the principal and the physician is given by* $e_{bnegl}(\underline{q})$ *and is induced by* $w_{bnegl} = [(h-b)p'(e_{bnegl}(\underline{q}))]^{-1}$.

1. *Suppose* $d = 0$. *For all* $\underline{q}$, *it follows that* $e_{bnegl}(\underline{q}) = e_{negl}(\underline{q})$ *with* $w_{bnegl} = \frac{1}{h}w_{negl}$.
2. *Suppose* $d > 0$. *For all* $\underline{q}$, *it follows that* $e_{strict}(\underline{q}) < e_{bnegl}(\underline{q}) < e_{negl}(\underline{q})$.

Organizational Design. Information projection increases agency costs and changes the Bayesian ranking of various performance measures. Classic results by Holmström (1979) and Jewitt (1997) imply that under Bayesian assumptions the efficiency of incentive provision increases in the informativeness of the performance measure. Under information projection, less informative monitoring technologies will often produce better results. In a world with biased evaluations, even if principal's are aware of these biases, less rather than more monitoring might be necessary to achieve the best possible outcome. Specifically, the sufficient statistic theorem of Holmström (1979) will typically fail in my model. To establish the ranking of incentive schemes, one here needs to consider the extent to which different schemes are *projection proof.*[25] When separating the ex-ante and ex-post interpretation of the evidence is key, the performance measures is not projection proof. Future research can address these issues in more detail.

7 Exchange of Information

I now turn to the second application of my the model. Evidence shows that information projection plays a crucial role in everyday communication. Indeed communication is that activity whose primary

[25] For example, the principal might be able to choose between: (i) a somewhat noisy performance measure conditioned purely on the fact whether the event $y_a = s_0$ is true or not, but not on s_0 or y_a, or (ii) perfectly observing both y_a and s_0 directly.

goal is to bridge informational differences. To establish the first-order effects of my model to this context, I focus on pure coordination problems where the interests of the sender and the receiver are perfectly aligned. The sender sends a message, y_s and the receiver responds y_r with the common objective that the receiver's action matches the state ω as often as possible.

7.1 Expert Talk

A computer manual that is perfectly informative for an electrical engineer is often meaningless for the average user. In this manner, the value of a message is determined not by its content alone, but by the relation between its content and what the audience knows and does not know to begin with. To capture such complementarity in the simplest way, let there be three signals: s_1 the technical language, s_2 a specific technical term, s_3 a noisy lay description. Suppose that the true state ω equals $\varpi_1 * \varpi_2$ where $\varpi_1, \varpi_2 \in \{-1, 1\}$ and there is a symmetric prior on ϖ_1 and ϖ_2. The sender's private information is $s_1 = \varpi_1$ and $s_2 = \varpi_2$. These two pieces of information are perfect complements, one is only valuable if the other is known. The lay description is an imperfect substitute of these two, and is given by $\Pr(s_3 = \omega \mid \omega) = h < 1$.

Communicating the technical language is prohibitively costly and the sender is faced with only three options: (i) send the technical term, (ii) send the lay description, or (iii) remain silent. The cost of sending a message is c, and remaining silent is free.[26] The table below summarizes a ρ-biased perception of payoffs corresponding to each of these choices:

silence	expert term	lay term
$\rho^2 + \frac{1}{2}(1-\rho^2)$	$\rho + \frac{1}{2}(1-\rho) - c$	$\rho^2 + h(1-\rho^2) - c$

An unbiased sender chooses the lay description if $h - c > \frac{1}{2}$ and remains silent otherwise. A biased sender deviates in two fundamental ways. First, she prefers remaining silent over sending the lay description too often. Second, she prefers sending the expert term over the lay term or remaining silent too often. Overall, she under-communicates information that substitutes her expertise and over-communicates information that complements her expertise. If the precision of the lay description is high, she remains silent too often. If the reverse is true, she communicates too often, but sends the wrong message which conveys little effective information.

Proposition 10 *If $\rho < \frac{h-1/2}{1-h}$, the advisor sends the lay term when $(h - \frac{1}{2}) > \frac{c}{1-\rho^2}$, and stays silent otherwise. If $\rho > \frac{h-1/2}{1-h}$, the advisor sends the expert term when $\frac{1}{2} > \frac{c}{\rho-\rho^2}$, and stays silent otherwise.*

[26] To simplify the analysis, I assume that the advisor knows the properties of s_3, but not its realization. The insights extend to the case where this assumption is relaxed.

Optimal Protocols. Consistent with earlier intuition, expert talk is too ambiguous when intended for lay audiences. More importantly, the model implies that simply mandating communication does not restore efficiency. In fact when experts are sufficiently biased, $\rho > \frac{h-1/2}{1-h}$, such a policy will backfire. Both adding and removing communication options can strictly lower welfare. To restore efficient information transmission, a protocol that restricts the use of messages that complement the speaker's background knowledge, but mandates the use of messages that substitute for this background is necessary.

Favoritism. The usefulness of protocols is further enhanced if we consider the parties inference about each other after the event of miscoordination $y_r \neq \omega$. Let's focus on the sender. Since a biased sender exaggerates the value of the receiver's information, she too often explains miscoordination by the receiver's lack of attentiveness or loyalty. Formally, if θ is the receiver's type, which stands for the probability that he attends to the sender's message, then a biased sender underestimates θ. For all fixed h and c, $E_\omega[\pi_1^\rho(\theta)]$ is again decreasing in ρ in the sense of first-order stochastic dominance. Furthermore, this underestimation is greater the greater is the difference in their background.

The analysis above suggests that the best people to proof-read a computer manual might be those who know little about computers. Similarly, textbooks should be proof-read by students not only by professors. To limit miscoordination, managers should not only be required to talk to sub-ordinates regularly, but be given specific guidelines on the set of admissible answers to the questions that might arise. In the absence of communication protocols, disadvantaged workers, such as high-skilled immigrants, might refuse to join corporate hierarchies due to the fear of underestimation, and instead choose less productive self-employment, such as opening a restaurant.

7.2 Credulous Listeners

Information projection also affects the way receivers learn from the advice they receive. To learn from the sender's message effectively, a receiver has to adapt the information content of this message to his own circumstances. A biased receiver will make a systematic mistake here. Suppose now that the receiver has some private information about the state. Before taking his action y_r, he listens to the sender who now recommends the action she thinks is best: y_e. By projecting his private information, the receiver will wrongly assume, however, that the sender's recommendation already incorporates his private information. He will thus fail to accommodate the advice, and instead of learning from it, he will imitate it too closely. If the receiver's private information is sufficiently important, even if there are no frictions, parties will be ex-ante better-off without the opportunity to communicate.[27]

[27] For proofs and a more detailed analysis, see Madarász (2007). The analysis there describes how information projection can result in the emergence of *continuous and overconfident herds* that would not arise under Bayesian assumptions.

8 Conclusion

The presence of the curse-of-knowledge, CLW (1989), and hindsight bias, Fischhoff (1975), in social judgements has been widely recognized. The aim of this paper has been to enrich the economic analysis of social inference by modelling a broad class of mispredictions people display in this context. Section 2 provided evidence in support of this model. Applications to agency and communication settings demonstrated the model's potential relevance for economics.

Information projection introduced an agency problem not present under standard Bayesian assumptions. In a frictionless learning environment, evaluators learned about workers' ability. In this context, the model provided a unified explanation of the types of assurance and avoidance behavior that medical observers have long attributed to defensive practices rather than cost-effective health care. To protect their reputation, fully concerned and risk-neutral workers over-produced ex-ante tests that complement the information evaluators independently learn ex-post, and under-produced ex-ante tests that substitute for that information. Comparative static predictions of the model offered new insights on the type of institutions that can help reduce these inefficiencies and showed that it is increased observability that will often exacerbate these practices.

Various tests will help identify the model. Information projection can be tested on disaggregated choice data, for example by combining choices over information sets and choices over outcomes, as in the design of Loewenstein et al. (2006). Results on belief-updating allow one to test the model in dynamic inference problems. The comparative static results of the paper help identify the presence and the significance of information projection in economic data more generally.

Importantly, when testing the model a significant limitation of my approach should be noted. Similarly to other economic models, while the model allows for heterogeneity in information projection, it does not pin down such heterogeneity. Certain pieces of information can be projected to a greater extent than others, but the model is limited by the fact that absent additional data, it cannot predict heterogeneity in a systematic manner. As argued before, however, the key implications of the model and the results of the paper, however, do not depend on the nature of such heterogeneity. As shown in Section 3, a biased person always exaggerates the value of the information available to another party. Furthermore, in all environments with the same information differences, a shift in a biased person's belief after an increase in her private information is always an increasing function of her bias.

Information projection is likely to matter for organizational problems not covered in this paper. Nonetheless, I believe that the results are suggestive of what will hold in related contexts. As Proposition 2 shows the main implications of Section 4 and 5 will hold with any finite number of outcomes. The results of Section 6 will also extend to more complicated tasks and to the presence of risk-averse agents. Future research can explore both empirical and theoretical consequences of the model to wage dynamics, careers or hierarchies in organizational contributing to a literature studied under Bayesian assumptions. For example, it would be useful to test whether in contrast to most statistical accounts, finer performance measures and unrestricted communication may in fact increase discrimination at the work-place. Such unintended discrimination may again be self-confirmatory and will also discourage human capital investments. The solution to alter these outcomes will be different. [28]

The model applies to various other social inference problems with asymmetric information. The dinner example of Section 3, for example, can be extended to study how information projection shapes friendships and might be a force which contributes to mistaken hostility and social conflict. Here people might choose the wrong social learning environments and mistakenly come to attribute taste differences to differences in social intentions while also developing false perceptions of hostility. Interventions that re-shape the structure of learning could greatly reduce false hostility and inefficient conflict.

Although the paper focused on interpersonal information projection, it is possible to extend the model to the case of intrapersonal information projection onto one's own future selves. The model predicts that people will be overconfident about their prospective memory for those pieces of information that they currently know, but less so for information they know they will learn in the future.[29] Exploring this link can shed new light on various puzzles in intertemporal choice, and provide novel predictions on the role of deadlines and reminders in mitigating self-control problems.

Finally, one possible theory of the source of information projection is limited perspective-taking. In line with this interpretation, one can extend the model is to consider the problem of ignorance projection where people underestimate the probability with which the signals whose realizations she does not know are unavailable to others.[30] The model can help understand limited perspective taking in Bayesian games of bargaining or voting. In contrast to cursedness where a person believes that others do not condition their choices on their private information, as introduced by Eyster and Rabin (2005), under information projection a person mistakenly believe that others do in fact condition their behavior on her private information. Future research can explore the contexts where cursedness and projection bias have similar and contexts where they have different implications.

[28] Similarly, the model might help understand the extent to which hindsight bias accounts for the anomalous evidence on CEO turnover, as suggested by Dirk Jenteer and Fadi Kanaan (2009), and politician turnover, as suggested by Justin Wolfers (2007).

[29] For evidence on over-confidence in prospective memory see for example Keith Erickson (2009).

[30] Although I believe that ignorance projection will not play a significant role in the contexts studied in this paper, the experimental paradigms of Fischhoff (1975), Newton (1990) and CLW (1989) do not allow to test for ignorance projection. Also anchoring-based accounts do not allow to formally distinguish it from information projection.

9 Appendix

Proof of Proposition 1. To prove this proposition, I first show that information projection shifts probabilistic weight from less to more informative information sets. Here informativeness is taken in the sense of Blackwell (1953). Note that given any two information sets I_i, $I_k \subset \mathbb{N}$, $I_i \cup I_k$ is always weakly more informative than I_i.

Let the posterior on ω induced by information I_i be σ_i and for $I_i \cup I_k$, σ_{i+k}. Since both Ω and Z are finite, σ_i and σ_{i+k} are finite and we can collect the realization of these posteriors as a function of all the possible realizations of the signals in N into matrices Σ_i and Σ_{i+k} respectively. By the law of iterated expectations: $E[\sigma_{i+k} \mid I_i] = \sigma_i$ for all I_i and I_k. Hence there exists a Markov-matrix T, i.e., a non-negative matrix with columns summing up to 1, such that $\Sigma_i = T\Sigma_{i+k}$.

The next step of the proof follows from Blackwell (1953): If $\Sigma_i = T\Sigma_{i+k}$ where T is a Markov-matrix, then for any fixed von-Neumann Morgenstern utility function, $u_i(y,\omega)$, $E_{\omega,p}[u^*_{I_i}] \leq E_{\omega,p}[u^*_{I_i \cup I_k}]$ where $u^*_{I_i}$ is defined by Eq.(2) as the maximum of $u(y,\omega)$ given information I_i.

Fix person k's information set I_k. The Bayesian and the fully biased perception of person $i's$ information are given by two probability distributions p^0_i and p^1_i over $\mathbb{N}$. The transition from p^1_i to p^0_i can be generated by re-allocating probabilistic weight from less informative to more informative information sets given Definition 1. Since a more informative information set implies greater expected utility, it then follows that $f^1(u^*_i)$ first-order stochastically dominates $f^0(u^*_i)$. Since $f^\rho(u^*_i)$ is a probabilistic mixture of $f^1(u^*_i)$ and $f^0(u^*_i)$ where the probabilistic weight is continuously moved from $f^0(u^*_i)$ to $f^1(u^*_i)$ as ρ increases the proposition follows. Corollary 1 follows again from the fact that projecting a more informative information set I'_k instead of I_k means shifting probabilistic weight to information sets with even greater expected utility. QED. .

Proof of Proposition 2. As stated, we show the claim for the finite outcome case. For simplicity, let X also denote the cardinality of the finite and ordered set X and index the elements of X by l in ascending order.

By the law of conditional probability, given any set of priors $f(\theta) \in \Delta\Theta$,and $f(u) \in \Delta\mathbb{R}$, ex-ante $E_{X,U}[f(\theta \mid x,u)] = f(\theta)$. Consider now the case where $\rho > 0$ and let again denote the ρ-biased inference by superscript ρ. The ex-ante expected beliefs of a ρ-biased observer are given by:

$$E_{X,U}[f^\rho(\theta \mid x,u)] = \sum_{l=1}^{X} f^\rho(\theta \mid x_l) f^0(x_i) = \sum_{l=1}^{X} \frac{f^\rho(x_l \mid \theta) f(\theta)}{f^\rho(x_l)} f^0(x_l) \tag{9}$$

where on the RHS of the above equation I suppressed the expectation operator with respect to u.

Maintaining this simplified notation, let's fix θ and re-write the above equality as:

$$E_{X,U}[f^{\rho}(\theta \mid x, u)] = f(\theta)\left[\sum_{l=1}^{X} f^{\rho}(x_l \mid \theta)\frac{f^0(x_l)}{f^{\rho}(x_l)}\right] = f(\theta)[\lambda^{\rho}(\theta)] \tag{10}$$

Here $\lambda^{\rho}(\theta) = \left[\sum_{l=1}^{X} f^{\rho}(x_l \mid \theta)\frac{f^0(x_l)}{f^{\rho}(x_l)}\right]$ is a short-hand for the term in the square-brackets.

Note first that $\sum_{l=1}^{X}[f^{\rho}(x_l \mid \theta) - f^{\rho}(x_l \mid \theta')] = 0$ and hence when $\rho = 0$, $\lambda^0(\theta) = 1$ for all θ. I now show that $\lambda^{\rho}(\theta)$ is decreasing in θ for all ρ. By virtue of the strict MLRP of $f(x \mid u, \theta)$ with respect to u, $\frac{f^0(x_l)}{f^{\rho}(x_l)}$ is decreasing in l. Taking any two θ, $\theta' \in \Theta$ such that $\theta > \theta'$, by virtue of the strict MLRP of $f(x \mid u, \theta)$ with respect to θ:

$$\sum_{l=1}^{L}[f^{\rho}(x_l \mid \theta) - f^{\rho}(x_l \mid \theta')] < 0 \text{ for any } L < X$$

Consider now $\lambda^{\rho}(\theta) - \lambda^{\rho}(\theta')$. Because $\frac{f^0(x_l)}{f^{\rho}(x_l)}$ is weakly decreasing in l, for the first $X - 1$ elements in this summation of $\lambda^{\rho}(\theta) - \lambda^{\rho}(\theta')$ there exists some number $Z > 1$ such that

$$\sum_{i=1}^{L}[f^{\rho}(x_l \mid \theta) - f^{\rho}(x_l \mid \theta')]\frac{f^0(x_l)}{f^{\rho}(x_l)} = Z\sum_{i=1}^{L}[f^{\rho}(x_l \mid \theta) - f^{\rho}(x_l \mid \theta')] < 0$$

Consider now the terms associated with x_X. Here, $f^{\rho}(x_X \mid \theta) - f^{\rho}(x_X \mid \theta') \geq 0$ and $\frac{f^0(x_X)}{f^{\rho}(x_X)} < 1$. The fact that $\lambda^{\rho}(\theta) < \lambda^{\rho}(\theta')$ then follows from the fact that $Z > \frac{f^0(x_X)}{f^{\rho}(x_X)}$. Intuitively, early elements in $\lambda^{\rho}(\theta) - \lambda^{\rho}(\theta')$, low l, are over-weighted and late elements, high l, are under-weighted.

Given that $\lambda^{\rho}(\theta)$ is decreasing in θ, for any prior $\theta \in \Delta\Theta$ and it follows that:

$$\int_{\theta < \theta^*} f(\theta)\lambda^{\rho}(\theta)d\theta \geq \int_{\theta < \theta^*} f(\theta)\lambda^0(\theta)d\theta \tag{11}$$

which means that $E_{X,U}[f^0(\theta \mid x, u)]$ first order stochastically dominates $E_{X,U}[f^{\rho}(\theta \mid x, u)]$.

To show that the same relation holds for $\rho < \rho'$ note that by Proposition 1, $\frac{f^{\rho'}(x_l)}{f^{\rho'}(x_{l'})} > \frac{f^{\rho}(x_l)}{f^{\rho}(x_{l'})}$ whenever $x_l > x_{l'}$. Combined with the fact that $f(x \mid \theta, u)$ satisfies the MLRP in u the proposition follows. QED. .

Proof of Proposition 3. Note that the probability of success is given by $\pi(x_S \mid \omega, I_a, \theta)$ where $I_a \subset \mathbb{N}$. is the agent's ex-ante set of signals. Taking the ex-ante expectations over Ω, it follows that $E_{\omega}[\pi(x_S \mid \omega, I_a, \theta)]$ is increasing in θ. Furthermore, this expression is also increasing by taking strict supersets of I_a from $\mathbb{N}$.

Let p_1 be the probability that s_1 is in I_a. Holding everything else constant, $E_{\omega}[\pi(x_S \mid \omega, I_a, \theta)]$

satisfies the MLRP in θ and p_1. The result then follows from the proof of Proposition 2 by assuming that $X = 2$. QED. .

Proof of Corollary 2. First, note that $u_a^*(I_a \mid \theta) \propto E_\omega[\pi(x_S \mid \omega, I_a, \theta)]$ for any fixed θ. Furthermore, holding $E_\theta[u_a^*(s_0 \mid \theta)]$ constant, $E_{\omega,\theta}[\pi^\rho(x_S \mid \omega, I_a, \theta)]$ is increasing in $E_\theta[u_a^*(s_0 \cup s_1 \mid \theta)]$, hence $E_{\omega,\theta}[\pi^\rho(x_S \mid \omega, \theta)]/E_\omega[\pi^0(x_S \mid \omega, \theta)]$ is increasing in g. The result then follows again from Proposition 2. QED. .

Proof of Proposition 4. Suppose $u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)$ is increasing in θ. Holding $u_a^*(s_0 \mid \theta)$ constant for all θ, it follows that $E_\omega[\pi^\rho(x_S \mid \theta, \omega)] \;/\; E_\omega[\pi^0(x_S \mid \theta, \omega)]$ is increasing in θ whenever $\rho > 0$. We now show that the following inequality is satisfied for all $\theta^* < 1$:

$$\begin{aligned} E_{\omega,\theta}[\pi^\rho(x_S \mid \omega, \theta)] \;\;/\;\; E_{\omega,\theta}[\pi^0(x_S \mid \omega, \theta)] \;\; > \\ \textstyle\int_0^{\theta^*} E_\omega[\pi^\rho(x_S \mid \theta)]\pi_0(\theta)d\theta \;/\; \int_0^{\theta^*} E_\omega[\pi^0(x_S \mid \theta)]\pi_0(\theta)d\theta \end{aligned} \tag{12}$$

This is true because if there is a $\theta^* < 1$ where this inequality is violated, then $E_\omega[\pi^\rho(x_S \mid \theta, \omega)] \;/\; E_\omega[\pi^0(x_S \mid \theta, \omega)]$ could not be increasing there.

Re-arranging the above inequality, we get that the definition of first-order stochastic dominance of $E_{\omega,\theta}[\pi_1^\rho(\theta \mid x_S, \omega)]$ over $E_{\omega,\theta}[\pi_1^0(\theta \mid x_S, \omega)]$. Note finally, that if $E_\omega[\pi^\rho(x_S \mid \omega,\, \theta)] \;/\; E_\omega[\pi^0(x_S \mid \omega,\, \theta)]$ is increasing in θ, then $E_\omega[\pi^\rho(x_S \mid \omega{,}\theta)] \;/\; E_\omega[\pi^{\rho'}(x_S \mid \omega{,}\theta)]$ is also increasing in θ whenever $\rho > \rho'$.

The proof for the case where $u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)$ is decreasing in θ is analogous. QED. .

Proof of Lemma 1. Fix m_S and m_F. Since s_0 and ω appear in all relevant terms, I suppress these from the notation. The evaluator's expected posterior is $E_{\omega,x}[\pi_1(\theta \mid x, m_S, m_F)] = E_\omega[E_\theta[\pi(x_S \mid \theta)][m_S\pi_1(\theta \mid x_S, s_1) + (1 - m_S)\pi_1(\theta \mid x_S)] + E_\theta[\pi(x_F \mid \theta)][m_F\pi_1(\theta \mid x_F, s_1) + (1 - m_F)\pi_1(\theta \mid x_F)]]$. By the law of iterated expectation, we get that $E_\omega[\pi_1(\theta \mid x_S, s_1)] = E_\omega[\pi_1(\theta \mid x_S)]$. Similarly, for the case where $x = x_F$. The claim then follows from the law of total probability $E_\omega[E_\theta[\pi(x_S \mid \theta)]\pi_1(\theta \mid x_S) + E_\theta[\pi(x_F \mid \theta)]\pi_1(\theta \mid x_F)] = \pi_0(\theta)$. QED. .

Proof of Proposition 5. As before, to simplify notation, I again suppress s_0 and ω from the notation. The expected posterior of a ρ-biased evaluator equals $E_{\omega,x}[\pi_1^\rho(\theta \mid m_S, m_F)] = E_\omega[E_\theta[\pi(x_S \mid$

$\theta)][m_S\pi_1^\rho(\theta \mid x_S, s_1) + (1 - m_S)\pi_1^0(\theta \mid x_S)]+ E_\theta[\pi(x_F \mid \theta)][m_F\pi_1^\rho(\theta \mid x_F, s_1)+ (1 - m_F)\pi_1^0(\theta \mid x_F)]]$. Given our assumptions, $E_\omega[\pi_1^\rho(\theta \mid x_S, s_1)] = E_\omega[\pi_1(\theta \mid x_S)]$. Hence form Proposition 3 it follows that $E_\omega[\pi_1(\theta \mid x_F)]$ first-order stochastically dominates $E_\omega[\pi_1^\rho(\theta \mid x_F, s_1)]$. By setting $m_F = 0$, the ρ-biased evaluator's expected assessments reduces to $E_\omega[E_\theta[\pi(x_S \mid \theta)]\pi_1(\theta \mid x_S)+E_\theta[\pi(x_F \mid \theta)]\pi_1(\theta \mid x_F)] = \pi_0(\theta)$ for all ρ. QED. .

Proof of Proposition 6. Consider the information gap between the ex-ante and the ex-post stages in the two different production scenarios. When s_0' is produced, the information gap is $g(s_0 \cup s_0')$ $= E_\theta[u_a^*(s_1 \cup s_0 \cup s_0' \mid \theta) - u_a^*(s_0 \cup s_0' \mid \theta)]$. When s_0' is not produced the information gap is $g(s_0)$ $= E_\theta[u_a^*(s_0 \cup s_1 \mid \theta) - u_a^*(s_0 \mid \theta)]$. In these expressions, $u_a^*(s)$ is again defined as the agent's ex-ante probability of producing success when the set of signals he has access to is s.

If given A and s_0, the two signals s_0' and s_1 are substitutes, then $g(s_0 \cup s_0') < g(s_0)$. If two signals s_0' and s_1 are complements, then $g(s_0 \cup s_0') > g(s_0)$. Fix the ex-post signal s_1. Let the ρ-biased evaluator's expected assessment when the set of ex-ante signals is s be $E_{\omega,\theta}[\pi_1^\rho(\theta \mid s)]$. The agent decides to produce s_0' if the following inequality holds:

$$b(E_{\omega,\theta}[\pi_1^\rho(\theta \mid \omega, s_0 \cup s_0')]) - b(E_{\omega,\theta}[\pi_1^\rho(\theta \mid \omega, s_0)]) + a \geq 0 \tag{13}$$

Suppose that monitoring always takes place, $m = 1$ and set $a = 0$. A sufficient condition for the above inequality to hold strict for all ρ is that $g(s_0 \cup s_0') > g(s_0)$. This follows from Corollary 2. Similarly, when $g(s_0 \cup s_0') < g(s_0)$, the above expression is strictly violated for all $\rho > 0$.

Note that for $m = 0$ the LHS of the above inequality equals a. Similarly, for $\rho = 0$ the LHS also equals a by the law of iterated expectations. Hence $a(0, 0) = 0$ is true. By continuity, it follows that the cut-off $a(m, \rho)$ is decreasing in m if $g(s_0 \cup s_0') < g(s_0)$ and $\rho > 0$. Similarly, $a(m, \rho)$ is increasing in m if $g(s_0 \cup s_0') > g(s_0)$ and $\rho > 0$. QED. .

Proof of Corollary 3. Given the assumption that there is no over- or under-inference, $E_\omega[\pi_1^\rho(\theta \mid s, x_S)]$ is the same for all s. Hence $a(m, \rho)$ is constant in m_S. Given the underestimation after a failure, $a(m, \rho)$ is increasing in m_F if $g(s_0 \cup s_0') > g(s_0)$ and decreasing in m_F if $g(s_0 \cup s_0') < g(s_0)$. When $m_F = 0$, however, $E_\omega[\pi_1^\rho(\theta \mid s, x_F)] = E_\omega[\pi_1(\theta \mid s, x_F)]$ because s_1 is not observed. It follows that for $m_S \geq 0$ and $m_F = 0$ $a(m, \rho) = 0$ for all ρ. QED. .

Proof of Proposition 7. To prove this proposition, I specify conditions on m_S and m_F such that $E_{\omega,x}[\pi_1^\rho(\theta \mid m_S, m_F)]$ is independent of the production of s_0'. This will ensure that $a(m, \rho) = 0$.

To simplify the notation, I drop the expectation operator E_ω and the conditioning on s_0. The reader is reminded, however, that all the expressions below hold from the ex-ante perspective. Also while below the distributions stand for the scalars $b(\pi) \in \Re$ where again $b(\pi) > b(\pi')$ if π fosd π', to save on notation I drop the function b from the notation.

If only s_0 is produced the evaluator's ex-ante expected assessment is

$$E_x[\pi_1^\rho(\theta \mid m_S, m_F)] =$$
$$= E_\theta[\pi(x_S \mid \theta)][m_S\pi_1^\rho(\theta \mid x_S) + (1 - m_S)\pi_1(\theta \mid x_S)] + E_\theta[\pi(x_F \mid \theta)][m_F\pi_1^\rho(\theta \mid x_F) + (1 - m_F)\pi_1(\theta \mid x_F)] \tag{14}$$

An analogous expression holds for $E_x[\pi_1^\rho(\theta \mid m_S, m_F, s_0')]$, the evaluator's ex-ante expected assessment when s_0' is also produced. These two expected assessments are the same, if the following condition holds:

$$m_F[E_x\pi_1^\rho(\theta) - E_x\pi_1^\rho(\theta \mid s_0')] = E_\theta[\pi(x_S \mid s_0')][m_S - m_F][\pi_1^\rho(\theta \mid x_S, s_0') - \pi_1(\theta \mid x_S, s_0')] \tag{15}$$

In deriving the above equation, I used the fact that if no under or over-inference is present when only s_0 is produced, then $\pi_1^\rho(\theta \mid x_S) = \pi_1(\theta \mid x_S)$ for all ρ.

If s_0' is a complement of s_1, then $E_x\pi_1^\rho(\theta)$ fosd $E_x\pi_1^\rho(\theta \mid s_0')$ and the LHS of the above equality is positive. Since $u_a^*(s_0 \cup s_0' \cup s_1 \mid \theta) - u_a^*(s_0 \cup s_1 \mid \theta)$ is increasing in θ, the RHS is also positive. Hence $m_S > m_F$ has to be satisfied. Furthermore, if $m_S > m_F$, then $m_F/(m_S - m_F) \in [0, \infty)$. Hence there always exists $m_S^*(\rho) > m_F^*(\rho)$ such that the above equality is satisfied. Finally, it is easy to see that the expression $a(m, \rho)$ is increasing in m_S since the RHS is increasing in m_S. Similarly, since the LHS is increasing in m_F and the RHS is decreasing in m_F, $a(m, \rho)$ is decreasing in m_F.

When s_0' is a substitute of s_1, then the LHS of the above equation is negative and given under-inference the RHS is also negative. The result then follows from the previous discussion. QED.

.

Proof of Lemma 4. Let's first derive the optimal strict liability, The principal's problem yields the following Lagrangian: $\mathcal{L}(w_S, e, \mu) = (p(e)q + bk + (1 - b)z)(1 - w_S) + \mu(p'(e)qw_S - 1)$. Solving for μ and substituting $w_{strict} = 1/p'(e)q$ the optimal effort level e_{strict} satisfies:

$$qp' = 1 - \frac{p''(p + (bk + (1-b)\dot{z})/q)}{(p')^2} = 1 - \frac{p''(p + b/(h-b) + z/q)}{(p')^2} \tag{16}$$

Given that $p''' \leq 0$, there is a unique solution increasing in k and h.

Let's now derive the optimal negligence contract. The principal's maximization problem now yields the following Lagrangian: $\mathcal{L}(w_S, e, \mu) = (p(e)q + bk + (1-b)z) - (p(e)(1-b) + b)w_S + \mu(p'(e)(1-b)w_S - 1)$. Solving for μ and substituting $w_{negl} = 1/p'(1-b)$ the optimal equilibrium effort level e_{negl} satisfies:

$$p'q = 1 - \frac{p''(p + b/(1-b))}{(p')^2} \tag{17}$$

The fact that $e_{strict} < e_{negl}$ follows from that fact that $(bk + (1-b)z)\ /\ (h-b)(k-z) > b/(1-b)$ because $b(k-z) + z(1-b) > bh(k-z)$ and $p''' \leq 0$. To show that the principal's welfare is greater under the negligence contract note that since $p'(e_{strict}) > 1$ and $p'(e_{negl}) > 1$, and $b/(h-b) + z/q > b/(1-b)$, it follows that $Ev(e_{strict}) < Ev(e_{negl})$. QED. .

Proof of Proposition 8. Let's fix an efficiency wage w_S. The effort choice under biased monitoring, $e^{\rho}_{negl}(\underline{q}, w)$ satisfies $p'(h-b)w_S = 1$. The effort choice under unbiased monitoring, $e_{negl}(\underline{q}, w)$ satisfies $p'(1-b)w_S = 1$. The result then follows from $p'' < 0$. QED. .

Proof of Corollary 6. Consider $e_{strict}(\underline{q})$ it is always true that $p'(e_{strict}(\underline{q})) < \infty$. Hence there exists h^* such that if $h < h^*$ then $p'(e_{negl}(\underline{q}))/h > p'(e_{strict}(\underline{q}))$. This implies that $e^{\rho}_{negl}(\underline{q}, w_{negl}) < e_{strict}(\underline{q}, w_{strict})$ if $h < h^*$. QED. .

Proof of Proposition 9. Consider the principal's problem when the agent's action is given by $e^{\rho}_{negl}(\underline{q}, w)$. Here the principal's Lagrangian is given by: $\mathcal{L}(w_S, e, \mu) = p(e)q + (bk + (1-b)z) - p(e)(h-b)w_S - bw_S + \mu(p'(e)(h-b)w_S - 1)$. Solving for μ and substituting $w_{bnegl} = 1/p'(h-b)$ we get that $e_{bnegl}(\underline{q})$ satisfies:

$$p'q = 1 - \frac{p''(p + b/(h-b))}{(p')^2} \tag{18}$$

It follows that $e_{bnegl} < e_{negl}$ as long as $h < 1$. QED. .

References

[1] Anderson, John, Marianne Jennings, Jordan Lowe, and Philip Reckers. 1997. "The Mitigation of Hindsight Bias in Judges' Evaluation of Auditor Decisions." *Auditing: A Journal of Practice and Theory*, Vol. 16, 20–39.

[2] Arkes, Hal, Paul Saville, Robert Wortman, and Alan Harkness. 1981. "Hindsight Bias Among Physicians Weighing The Likelihood of Diagnoses." *Journal of Applied Psychology*, Vol. 66, 252 – 254.

[3] Baron, Jonathan and John Hershey. 1988. "Outcome Bias in Decision Evaluation." *Journal of Personality and Social Psychology*, Vol. 54, 569-579.

[4] Baron-Cohen, Simon, Alan Leslie, and Uta Frith. 1985. "Does the Autistic Child Have a 'Theory of Mind'?" *Cognition*, Vol. 21, 37–46.

[5] Becker, Gary. 1962. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy,* Vol. 70, 9- 49.

[6] Bernstein, Daniel, Cristina Atance, Geoffrey Loftus, and Andrew Meltzoff. 2004. "We Saw It All Along: Visual Hindsight Bias in Children and Adults." *Psychological Science,* Vol. 15, 264 - 267.

[7] Biais, Bruno and Martin Weber. 2008. "Hindsight bias and Investment Performance." Working Paper *IDEI Toulouse.*

[8] Birch, Susan and Bloom Paul. 2007. "The Curse of Knowledge in Reasoning About False Beliefs." *Psychological Science,* Vol. 18, 382 -386.

[9] Berlin, Leonard and Roland Hendrix. 1998. "Perceptual Errors and Negligence." *American Journal of Roentgenology*, Vol. 170, 863-867.

[10] Berlin, Leonard. 2004. ""Malpractice Issues in Radiology: Outcome Bias." *American Journal of Roentgenology*, Vol.183, 557-560.

[11] Blackwell, David. 1953. "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics*, Vol. 24, 265-272.

[12] Camerer, Colin. 1987. "Do Biases in Probability Judgment Matter in Markets? Experimental Evidence." *American Economic Review.* Vol. 77, 981-997.

[13] Camerer, Colin, George Loewenstein, and Martin Weber. 1989. "The Curse of Knowledge in Economic Settings: An Experimental Analysis." *Journal of Political Economy,* Vol. 97, 1234-1254.

[14] Camerer, Colin. 1995. "Individual Decision Making." in John Kagel and Alvin Roth eds. *The Handbook of Experimental Economics.* Princeton University Press, Princeton.

[15] Camerer, Colin and Ulrike Malmendier. 2007. "Behavioral Economics of Organizations." in: P. Diamond and H. Vartiainen (eds.), *Behavioral Economics and Its Applications*. Princeton University Press, Princeton.

[16] Caplan, Robert, Karen Posner, and Frederick Cheney. 1991. "Effect of Outcome on Physician Judgments of Appropriateness of Care." *Journal of the American Medical Association,* Vol. 265, 1957-1960.

[17] Crawford, Vincent and Nagore Iriberri. 2007. "Level-k Auctions: Can a Non-Equilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica,* Vol. 75, 1721-1770.

[18] Dewatripont, Mathias, Ian Jewitt, and Jean Tirole. 1999. "The Economics of Career Concerns, Part I: Comparing Information Structures." *Review of Economic Studies*, Vol. 66, 183-98.

[19] Durell, Alan. 1999. "Attribution in Performance Evaluation." Ph.D Thesis. Harvard University, Cambridge MA.

[20] Erickson, Keith. 2009. "Forgetting that We Forget." *Journal of the European Economic Association,* forthcoming.

[21] Eyster, Erik and Matthew Rabin. 2005. "Cursed Equilibrium." *Econometrica*, Vol. 73, No. 5., 1623-1672.

[22] Farber, Henry and Robert Gibbons. 1996 "Learning and Wage Dynamics." *Quarterly Journal of Economics*, Vol. 61, 1007–1047.

[23] Fischhoff, Baruch. 1975. "Hindsight / foresight: The Effect of Outcome Knowledge On Judgement Under Uncertainty." *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 1, 288-299.

[24] Gibbons, Robert and Michael Waldman. 1999a. "A Theory of Wage and Promotion Dynamics Inside Firms." *Quarterly Journal of Economics,* Vol. 114, 1321-1358.

[25] Gibbons, Robert and Michael Waldman. 1999b. "Careers in Organizations: Theory and Evidence." in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics,* Vol. 3, Chapter 36. Elsevier: Amsterdam.

[26] Gilovich, Thomas, Kenneth Savitsky, and Victoria Medvec. 1998. "The Illusion of Transparency: Biased Assessment of Other's Ability to Read our Emotional States." *Journal of Personality and Social Psychology*, Vol. 76, 743-753.

[27] Gilovich Thomas, Victoria Medvec and Kenneth Savitsky. 2000. "The Spotlight Effect in Social Judgment: An Egocentric Bias in Estimates of the Salience of One's Own Actions and Appearance." *Journal of Personality and Social Psychology,* Vol. 78, 211-222.

[28] Guilbault, Rebecca, Fred Bryant, Jennifer Brockway, and Emil Posavac. 2004. "A Meta-Analysis of Research on Hindsight Bias." *Basic and Applied Social Psychology,* Vol. 26, 103-117.

[29] Harley, Erin. 2007. "Hindsight Bias in Legal Decision Making." *Social Cognition,* Vol. 25, 48-63.

[30] Harris, Milton and Bengt Holmström. 1982. "A Theory of Wage Dynamics." *Review of Economic Studies,* Vol. 49, 315-333.

[31] Heath, Chip and Nancy Staudenmayer. 2000. "Coordination Neglect: How Lay Theories of Organizing Complicate Coordination in Organizations." *Research in Organizational Behavior*, Vol. 22, 153-191.

[32] Heath, Chip and Dan Heath. 2007. *Made to Stick: Why Some Ideas Survive and Others Die.* Random House.

[33] Hinds, Pamela. 1999. "The Curse of Expertise: The Effects of Expertise and Debiasing Methods on Predictions of Novice Performance." *Journal of Experimental Psychology,* Vol. 5, No. 2., 205-221.

[34] Holmström, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics*, Vol. 13, 324-40.

[35] Holmström, Bengt. 1982. "Managerial Incentive Problems–A Dynamic Perspective." published in *Review of Economic Studies*, 1999. Vol. 66, 169-82.

[36] Innes, Robert. 1990. "Limited Liability and Incentive Contracting with Ex-ante Action Choices." *Journal of Economic Theory,* Vol. 52, 45-67.

[37] Jenter, Dirk and Fadi Kanaan. 2009. "CEO Turnover and Relative Performance Evaluation" Working Paper, Stanford GSB, Stanford CA.

[38] Keysar, Boaz and Ann Henly. 2002. "Speakers' Overestimation of Their Effectiveness." *Psychological Science*, Vol. 13, 207–212.

[39] Jewitt, Ian. 1997. "Information and Principal-Agent Problems." Working Paper, Oxford.

[40] Kessler, Daniel and Mark McClellan. 1996. "Do Doctors Practice Defense Medicine?" *Quarterly Journal of Economics,* Vol. 111, 353-390.

[41] Kessler, Daniel and Mark McClellan. 2000. "How Liability Law Affects Medical Productivity." *NBER Working Papers* No. 7533.

[42] Kruger, Justin, Epley Nicholas, Jason Parker and Zhi-Wen Ng. 2005. "Egocentrism over E-mail: Can People Communicate as well as They Think?" *Journal of Personality and Social Psychology,* Vol. 89, 925-936.

[43] Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics,* Vol. 4, 1209-1248.

[44] Loewenstein, George, Don Moore, and Roberto Weber. 2006. "Misperceiving the Value of Information in Predicting the Performance of Others." *Experimental Economics,* Vol. 3, 281-295.

[45] Madarász, Kristóf. 2007. "Information Projection: Model and Applications." Working Paper, UC Berkeley, Berkeley, CA.

[46] Milgrom, Paul. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics,* Vol. 12, 380-391.

[47] Newton, Elizabeth. 1990. "Overconfidence in the Communication of Intent: Heard and Unheard melodies." Ph.D. Thesis. Stanford University, Stanford, CA.

[48] Piaget, Jean and Bärbel Inhelder. 1967. *The Child's Conception of Space*. New York, Norton.

[49] Prendergast, Canice. 1993. "A Theory of Yes Men." *American Economic Review*, Vol. 83, 757-770.

[50] Prendergast, Canice and Robert Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy*, Vol. 104, 958-978.

[51] Pronin, Emily, Puccio Carolyn, and Lee Ross. 2002. "Understanding Misunderstanding: Social Psychological Perspectives." in T. Gilovich, P. Griffin and D. Kahneman eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge.

[52] Pronin, Emily, Thomas Gilovich, and Lee Ross. 2004. "Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self versus Others." *Psychological Review*, Vol. 111, 781-799.

[53] Rabin, Matthew and Dimitri Vayanos. 2009. "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Review of Economic Studies,* forthcoming.

[54] Radner, Roy and Joseph Stiglitz. 1984. "A Nonconcavity in the Value of Information." in M. Boyer and R. Khilstrom eds. *Bayesian Models of Economic Theory*. Elsevier: Amsterdam.

[55] Rachlinski, Jeffrey. 1998. "A Positive Psychological Theory of Judging in Hindsight." *The University of Chicago Law Review*, Vol. 65, 571-625.

[56] Sanna, Lawrence, Norbert Schwarz, and Shevaun Stocker. 2002. "When Debiasing Backfires: Accessible Content and Accessibility Experiences in Debiasing Hindsight." *Journal of Experimental Psychology,* Vol. 28, 497–502.

[57] Sappington, David. 1983. "Limited Liability Contracts between Principal and Agent." *Journal of Economic Theory*, Vol. 29, 1-21.

[58] Shapiro, Carl and Joseph Stiglitz. 1984. "Equilibrium Unemployment as a Worker Discipline Device." *American Economic Review,* Vol. 74, 433-444.

[59] Scharfstein, David and Jeremy Stein. 1990. "Herd Behavior and Investment." *American Economic Review*, Vol. 80, 465-479.

[60] Studdert, David, Michelle Mello, William Sage, Catherine DesRoches, Jordon Peugh, Kinga Zapert, and Troyen Brennan. 2005. "Defensive Medicine Among High-Risk Specialist Physicians in a Volatile Malpractice Environment." *Journal of the American Medical Association,* Vol. 293, 2609-2617.

[61] Tversky, Amos and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, Vol. 185, 1124-113.

[62] Viscusi, Kip and Richard Zeckhauser. 2005. "Recollection Bias and the Combat of Terrorism." *Journal of Legal Studies,* Vol. 34, 27-55.

[63] Van Boven, Leaf, Gilovich Thomas, and Victoria Medvec. 2003. "The Illusion of Transparency in Negotiations." *Negotiation Journal,* April, 117-131.

[64] Wolfers, Justin. 2007. "Are Voters Rational? Evidence from Gubernatorial Elections." Working Paper Wharton School, UPenn, PA.