MOTIVATED MISSPECIFICATION

MINGZI NIU*

NOVEMBER 7, 2023

Click Here for the Latest Version

ABSTRACT: In a wide range of interactions, a party can benefit from manipulating another party's perception. I propose a model of perception manipulation to investigate how the type of misperception is determined in a principal-agent framework. I take an individual's perception to be their learning model to interpret observations. In my model, the principal benefits from a project whose expected output depends on its quality, the agent's ability, and the agent's effort. The principal controls the agent's perception of the project quality. Subsequently, the agent exerts effort and updates belief about her own ability upon output observations. Overselling the project quality stimulates effort in the short run but makes the agent misinfer her own ability, which potentially backfires and lowers the agent's effort in the long run. I show that if project quality and the agent's ability cannot be empirically disentangled using output observations, then the agent's long-run effort is robust to manipulation. I then characterize sufficient conditions on the production function under which the principal oversells or undersells project quality. The key intuition is that the principal's optimal manipulation strategy downplays factors that affect project output independent of the agent's effort. This makes the agent misperceive that the output is highly sensitive to her effort, and therefore stimulates more effort by the agent. My work thus provides a novel approach to induce effort (perception manipulation), complementary to the usual monetary or informational incentives studied in principal-agent theory. I apply my results to understand manipulation in mentorship and abusive relationships.

KEYWORDS: Perception Manipulation, Misspecified Learning, Motivated Belief, Berk-Nash Equilibrium, Attribution Error, Expectation Management, Blame-Shifting

DEPARTMENT OF ECONOMICS, RICE UNIVERSITY, MINGZI.NIU@RICE.EDU

*I am particularly indebted to Mallesh Pai for his continued guidance and strong support. I am grateful to Nina Bobkova, Hülya Eraslan, and Rakesh Vohra for their valuable suggestions. I also thank Cuimin Ba, Simon Board, Matteo Camboni, Rahul Deb, Francesc Dilmé, Mira Frick, Peter Hartley, Botond Kőszegi, Matt Mitchell, Isabelle Perrigne, and Xun Tang, as well as conference participants at Texas Theory Camp, for their helpful comments.

1. INTRODUCTION

Stimulating motivation is integral to effective management and leadership. How can a manager motivate a worker to exert effort? The manager can use monetary incentives — for example, the manager can give a bonus to the worker for good performance, as in the standard moral hazard model (e.g., Holmström, 1979). The manager can also reveal information over time to induce incremental effort — for example, the manager can commit to providing further feedback once the worker completes a certain workload, as in the information design literature (e.g., Ely and Szydlowski, 2020). In this paper, I underscore an alternative channel to incentivize effort — through *perception manipulation*. That is, the manager can influence how the worker perceives the work at hand. This channel of perception manipulation is common in practice (see *Section 6.1* for a review). Under a distorted perception, the agent can be motivated to exert excessive effort voluntarily even without further monetary or informational transfers.

To fix ideas, consider the following project management scenario. A principal (he) has a project that requires long-term work, and he delegates this project to an agent (she). The principal benefits from the project output, and the expected output depends on three factors: the project quality (i.e., the value or potential of the project), the agent's ability, and the agent's effort. The principal knows the true quality of his project and is trusted by the agent.[1] If the principal tells the agent that the project has high quality, the agent will be motivated to work harder on the project, which is beneficial to the principal. This may incentivize the principal to oversell the quality of his project, making the agent perceive a higher project quality than reality.

However, is overselling project quality always a good strategy for the principal? In the short run, the answer is indeed yes. By overselling project quality, the principal sets up an unrealistically high expectation on the return to the agent's effort and thus makes her misperceive that the project deserves more effort. Nevertheless, overselling can backfire in the long run. Over time, the agent would notice that, given the effort that she exerts, the actual output is not as high as expected. This can trigger the agent to doubt about her own ability; she may misattribute the disappointing output to her low ability to do this job.[2] This low perceived ability can potentially demotivate the agent to exert effort in the long run. Therefore, the **question** of this paper is: taking this long-run learning effect into account, should the principal shift the agent's perception (of the project quality) upwards or downwards? Specifically, under what circumstance, should the principal oversell his

[1] In the baseline model, I assume that the agent fully trusts the principal. See *Remark 1* for a short discussion on this assumption of gullible agent. I relax this assumption in *Section 5.2*.

[2] Factors that the principal manipulates and the agent infers are open to interpretation. For example, the principal may directly manipulate how the agent perceives her own ability as an authority in ability evaluation and leave the agent to infer about the project quality. See *Section 1.2* for more applications.

project so that the agent perceives a project quality higher than its true value? On the flip side, is there any situation where underselling the project quality and making the agent confident about her own ability is nevertheless more advisable for the principal?

This investigation into manipulation strategy informs how individual perception is steered to benefit a particular party yet persists over time. Broadly speaking, individual perception can be considered as the learning model that the individual uses to interpret observations. A misperception thus corresponds to the case where the individual engages in learning under a *misspecified* model, which fails to accommodate the objective data-generating process and misleads the individual to consistently interpret experiences incorrectly. In my model, the principal can influence the agent's perception. In other words, the principal's manipulation strategy directly shapes the agent's learning model. Therefore, an alternative way to frame the question of this paper is: *what is the environment that breeds a certain type of misperception?* To answer this question, I build on a flourishing literature on misspecified learning (e.g., Esponda and Pouzo, 2016; Bohren and Hauser, 2021; Fudenberg, Lanzani and Strack, 2021, see *Section 6.1* for a review). It has been used to explain the persistence of overconfidence (Heidhues, Kőszegi and Strack, 2018), the fragility of social learning (Frick, Iijima and Ishii, 2020), and the recurrence of populism (Levy, Razin and Young, 2022). The majority of this literature takes a certain type of model misspecification as exogenously given and analyzes its implications. My contribution is to investigate how the type of misspecification is endogenously determined in a principal-agent framework.

I show that the principal's manipulation strategy (and thus the agent's model misspecification) depends on two aspects. The first is his time preference; namely, how much the principal cares about his long-run welfare. The second is the nature of the project; namely, how different factors (i.e., project quality, the agent's own ability and effort) contribute to the output. For example, some projects heavily rely on project quality in the sense that all the effort is dedicated to addressing a particular preregistered question. The value of the question (i.e., project quality) is thus decisive in determining the output from it. By contrast, the agent's ability is central to some projects in the sense that the project output is primarily driven by the agent's intelligence. This paper demonstrates that such distinction in the project nature matters for the principal's manipulation strategy (*Proposition 1*). I show in *Section 4.1* that, for the first type of project, the principal should oversell the project regardless of his time preference. In contrast, in the latter case, a patient principal tends to undersell the project.

The key mechanism that drives the principal's manipulation strategy is as follows. For the sake of his long-run welfare, the principal should downplay external contribution that is not contingent on the agent's effort by manipulating the agent's perceived project quality. As such, the principal makes the agent *internalize* contributing forces beyond her control and thus misperceive that the outcome is highly sensitive to her effort. This manipulation inflates the agent's perceived return to her effort and enhances her perceived responsibility for the outcome. With a higher perceived agency over the project output, the agent is then empowered to exert excessive effort in favor of the principal.

1.1. *Overview of Model and Main Results*

I consider a discrete-time infinite-horizon model in which a principal and an agent benefit from a project and observe its output in each period. The output depends on the agent's effort, her ability, the project quality, and a mean-zero normal shock. Only the agent bears the effort cost, and the marginal cost of effort is strictly increasing. Only the principal knows the project quality, and both players are uncertain about the agent's ability. The principal manipulates the agent's perceived project quality at the start of the game. Fixing her perceived quality of the project, the agent then (i) exerts effort that is myopically optimal given her current belief about her ability, and (ii) updates her belief about her ability accordingly in a Bayesian manner upon observing the project's output in each period. Over time, the agent's perceived precision of her ability assessment grows, and the learning process converges to a stable state in which the agent precisely expects the mean output.

The key tension in the model resides in the potentially opposite immediate and long-run effects of manipulation. The immediate effect is straightforward: the agent exerts higher effort if the principal enhances her perceived project quality (i.e., overselling). However, the long-run effect is more subtle, resulting from two conflicting forces — if the project is oversold, the agent ends up with (i) an overly high perceived project quality which motivates effort, and (ii) an unrealistically low assessment of her own ability which demotivates effort. Therefore, it is unclear a priori whether or not overselling the project can induce a higher stable effort from the agent.

To study the long-run effect of manipulation, I first provide a complete information benchmark in which the agent knows her ability and the true project quality. I then show that, if the principal tells the agent the true project quality, or put differently, if the agent learns about her ability under the correctly specified model, then the long-run stable state replicates this first-best benchmark. To examine whether the principal can stimulate long-run effort above the agent's first-best effort, I distinguish three classes of production functions. In the first class, there exists a *sufficient statistic* that summarizes the combined contribution of the project quality and the agent's ability to the output (*Definition 2*). In other words, it is impossible to disentangle the separate impacts of these two factors on output even when the agent experiments with varying levels of effort. In the second class, the agent's ability has a separate direct value on top of its instrumental value in stimulating effort. I call this case *"ability-laden production"* (*Definition 3*). In the third class, project quality has a separate direct value, and I call this case *"quality-laden production"* (*Definition 4*).

I show that if the production features a sufficient statistic, then manipulation does not affect long-run effort. Under any manipulation, the agent always exerts her first-best effort in the long run. Therefore, the agent's long-run choice and welfare are robust to her model misspecification. In this case, the principal oversells the project if he cares about his immediate welfare; otherwise, he is indifferent to manipulation strategies. Conversely, if the production is ability-laden, then overselling stimulates the agent's stable effort as well as her immediate effort above her first-best effort. Therefore, the principal prefers to oversell the project regardless of his time preference. For quality-laden production, underselling results in a higher stable effort relative to the agent's first-best effort. Thus, manipulation generates opposite immediate and long-run effects. In this case, the principal undersells the project if he is sufficiently patient (*Theorem 1*).

The logic of these results is as follows. Suppose the production features a sufficient statistic in the sense that the mean output only depends on the agent's effort and this sufficient statistic. Then in the stable state, fixing her effort, the agent correctly learns the sufficient statistic. In this case, the influence of a misspecified project quality is completely neutralized by the induced assessment of ability. Therefore, although the agent misperceives the project quality and her own ability, she correctly perceives their combined effect and thus the true marginal return to her effort. This is why, under any manipulation, the agent always exerts her first-best effort in the long run. For ability-laden production, the agent's ability has a separate direct contribution to output which does not depend on the agent's effort. Under overselling, the perceived ability is driven below its true value. Thus the agent underestimates this separate value of ability. To explain her observations, the agent ends up perceiving an inflated return to effort and thus exerts excessive effort relative to her first-best effort. The same logic applies to quality-laden production where the project quality has a separate value; in this case, underselling stimulates long-run effort.

Here, I provide the **main takeaway** from these results. To stimulate the long-run stable effort, the principal tends to downplay the external forces independent of the agent's effort. This can be done either by directly understating its value (e.g., underselling when the project quality has a separate value) or by exaggerating the contribution of other factors (e.g., overselling when the agent's own ability has a separate value). By doing so, the principal distorts how the agent interprets the observed outcomes. The agent is then misled to attribute the output more to her effort. Since her decision is now perceived to be more influential in changing outcomes, the agent is held more responsible for the success and failure of the project. This is the key to motivating effort.

1.2. *Applications*

The decision-relevant factors to the output, namely, the project quality and the agent's own ability, are versatile in interpretation. The essential part of this perception manipulation model is that: (i) the observed outcome depends on several decision-relevant factors and the agent's effort; (ii) the principal benefits from the agent's effort and can manipulate the agent's perception of one factor; and (iii) this manipulation inevitably affects the agent's learning about other factors to account for the observed outcome.

Section 4.2 demonstrates how the model fits into an analysis of abusive relationships. I show that emotional abuse can be used to gain power in a professional or intimate relationship. In this context, the project quality refers to an individual's personal quality (e.g., sense of humor, quality of judgment, etc.), the agent's ability refers to how fit the partnership is, and the output refers to her everyday well-being. Since one's personal quality has a separate value on top of her benefit from the relationship, based on the results of this paper, a manipulator tends to undersell in this case. This corresponds to using emotional abuse tactics (e.g., overprotection, blame-shifting, neglecting, and degrading comments) to undermine the individual's perception of her personal quality. The individual is then misled to perceive an unrealistically high return from the relationship, which justifies her devotion and sacrifice to the relationship.

Furthermore, individuals can engage in self-manipulation through mechanisms such as selective memory and wishful thinking. Although I use a principal-agent relationship as the primary scenario of the model, the same analysis can be easily carried over to self-manipulation as in the literature on motivated belief (see *Section 6.1* for a review).

Outline. The rest of the paper unfolds as follows. I set up the baseline model and characterize the agent's effort choices in *Section 2*. In *Section 3*, I illustrate the key tradeoff between immediate and long-run welfare for the principal and characterize how his manipulation strategy is determined. I then apply the main results of the model to mentorship, abusive relationships, and trust-building in *Section 4*. In *Section 5*, I extend the model to cases where (1) there is no noise in output observation and (2) the agent does not blindly trust the principal. I conclude in *Section 6* with a review of related literature and several avenues for future work.

2. Model

I consider a principal-agent model in discrete time, infinite horizon: $t = 0, 1, 2, 3 \ldots$. The principal needs the agent to exert effort on a project over time. The true project quality is $Q \in [\underline{q}, \bar{q}]$ and the true ability of the agent is $A \in \mathbb{R}$. The principal has a one-shot move at the start of the game $t = 0$, influencing the agent's perceived project quality $\tilde{q} \in [\underline{q}, \bar{q}]$. Under this perceived project quality, the agent exerts effort $e_t \geq 0$ in each period

$t = 1, 2, 3 \dots$ to yield the project output

$$y_t = Y(e_t, Q, A) + \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ is a random noise that is identically and independently distributed across periods.

Information Structure. Both the principal and the agent are uncertain about the agent's ability and they share the same prior belief $\pi_0 \sim N(\mu_0, \sigma_0^2)$. The information asymmetry is that only the principal knows the true project quality Q.

Fixing her perceived project quality as $\tilde{q}$, the agent then exerts effort according to her belief about her own ability and updates this belief in each period upon output observations in a Bayesian manner.

Stage-Game Payoffs. Both players benefit from the output but only the agent exerts effort to produce the output. The cost of effort e is $c(e)$. The principal's stage-game payoff is $u_t^P = y_t$, and the agent's stage-game payoff is $u_t^A = y_t - c(e_t)$.[3] Throughout the paper, I assume $Y(\cdot)$ and $c(\cdot)$ are twice continuously differentiable, and maintain the following standard assumptions on the production function $Y(\cdot)$ and the cost function $c(\cdot)$.

ASSUMPTION 1 (Three Contributors)**.** $Y_e, Y_q, Y_a > 0$.

This asserts that effort, project quality, and the agent's ability all contribute to the production.

ASSUMPTION 2 (Marginal Incentives)**.** $Y_{ee} \leq 0$, $c_{ee} > 0$, $c_e(0) = 0$, $\lim_{e \to \infty} c_e = \infty$.

This assumption asserts decreasing marginal product and increasing marginal cost of effort. It guarantees a unique myopically optimal effort for any q and a.

ASSUMPTION 3 (Flexibly Varying Output)**.** *For any* $e \geq 0$, $q \in [\underline{q}, \bar{q}]$, $\lim_{a \to -\infty} Y(e, q, a) = -\infty$, $\lim_{a \to \infty} Y(e, q, a) = \infty$.

This assumption states that the mean output varies flexibly with the agent's ability. It makes attribution error possible: for any $e \geq 0$, $(q, q', a) \in [\underline{q}, \bar{q}]^2 \times \mathbb{R}$, there exists $a' \in \mathbb{R}$ such that $Y(e, q, a) = Y(e, q', a')$.

ASSUMPTION 4 (Effort and Project Quality as Complements)**.** $Y_{eq} > 0$.

[3] I assume here the project output is a public good. The main results of the paper hold as long as u_t^P and u_t^A are strictly increasing in y_t. For example, the results are robust to any monotone sharing rule such as $u_t^P = \lambda y_t$ and $u_t^A = (1 - \lambda) y_t$ where $\lambda \in (0, 1)$. I make the public-good assumption for ease of exposition and focus on the channel of perception manipulation in inducing effort exempt from designing the compensation rule to provide monetary incentives.

Assumption 4 captures the synergy between the agent's effort and project quality: a higher project quality increases the marginal return to effort and thus motivates the agent to work harder.

Payoffs and Strategies. Following the misspecified learning literature, I assume that the agent is myopic. Denote $\pi_{t-1} \in \Delta(\mathbb{R})$ as the agent's (prior) belief about her own ability at time t. The agent exerts effort $e_t \geq 0$ in period t to maximize her expected stage-game payoff:

$$\begin{aligned} U_t^A(e_t; \tilde{q}, \pi_{t-1}) &= E[y_t - c(e_t) | \tilde{q}, \pi_{t-1}] \\ &= \int_{a \in \mathbb{R}} Y(e_t, \tilde{q}, a) \mathrm{d}\pi_{t-1} - c(e_t). \end{aligned} \tag{1}$$

Denote $e^*(\tilde{q}, \pi)$ as the agent's myopically optimal effort when her perceived quality of the project is $\tilde{q}$ and her belief on her own ability is π.

The principal weighs the tradeoff between the short-run and long-run impacts of perception manipulation. To characterize the payoff for the forward-looking principal, the standard approach is to discount his expected stage-game payoffs to period $t = 0$ and sum them up. However, to do this, one must track down the agent's effort trajectory over periods. Since the agent engages in active learning, the data-generating process is endogenous and evolves in each period, which makes tracing effort technically intractable. To see this, recall that the agent's effort depends on her current belief about her own ability. This belief is updated as a function of the output in the previous period (by Bayesian rule), which is in turn a random variable adapted to the agent's effort in the previous period. Starting from period $t = 2$, the agent's effort follows an endogenous stochastic process that evolves depending on the output realization in the previous period, and the calculation becomes increasingly more complicated as period t grows.

For tractability, I assume that the principal only cares about two states — the immediate state ($t = 1$) and the long-run stable state ($t = \infty$). The principal's immediate payoff is given by his expected stage-game payoff in period 1 as follows:

$$\begin{aligned} U_1(\tilde{q}; Q, \pi_0) &= E[y_1 | Q, \pi_0, \tilde{q}] \\ &= \int_{a \in \mathbb{R}} Y(e_1, Q, a) \mathrm{d}\pi_0. \end{aligned} \tag{2}$$

To describe the long-run stable state and characterize the principle's long-run payoff, I draw on the notion of *Berk-Nash Equilibrium* (Esponda and Pouzo, 2016). A Berk-Nash Equilibrium in the current game is a pair of effort strategy and belief on ability such that:

(i) the effort strategy is consistent with the agent's perception — that is, the chosen effort is optimal under the belief on ability and the perceived project quality;

(ii) the belief is confined to a set of ability levels that make the perceived output distribution "closest" to the true output distribution.

Here, the "distance" between two distributions is measured by the Kullback-Leibler divergence. Formally, for any effort $e \in \mathbb{R}_+$, the true output distribution $f(y|e)$ is a normal density with mean $Y(e, Q, A)$ and variance σ_ε^2. The perceived output distribution $f_{a'}(y|e)$ with any perceived ability $a' \in \mathbb{R}$ is a normal density with mean $Y(e, \tilde{q}, a')$ and variance σ_ε^2. The Kullback-Leibler divergence between the true output distribution and the perceived output distribution, given any (mixed) effort strategy $\xi \in \Delta(\mathbb{R}_+)$ and perceived ability a', is thus

$$K_\xi(f, f_{a'}) \equiv \sum_{e \in \mathbb{R}_+} E_{f(\cdot|e)} \left[\ln \frac{f(y|e)}{f_{a'}(y|e)} \right] \xi(e), \tag{3}$$

where the expectation is taken over y following the true output distribution $f(\cdot|e)$ under the effort e.

Given the effort strategy ξ, the set of ability levels that minimize the "distance" between the true output distribution and the perceived output distribution is

$$\Theta(\xi) \equiv \arg\min_{a' \in \mathbb{R}} K_\xi(f, f_{a'}).$$

A Berk-Nash Equilibrium for the agent's learning is defined as follows.

DEFINITION 1 (Berk-Nash Equilibrium). *A Berk-Nash Equilibrium specifies a pair of an effort strategy and a belief about the agent's ability (ξ_∞, π_∞) such that*

(i) the effort strategy $\xi_\infty \in \Delta(\mathbb{R}_+)$ is optimal under the belief π_∞ and the agent's perceived project quality $\tilde{q}$, that is, if $\hat{e}$ is in the support of ξ_∞, then

$$\hat{e} \in \operatorname*{argmax}_{e' \in \mathbb{R}_+} \int_{a \in \mathbb{R}} Y(e', \tilde{q}, a) \mathrm{d}\pi_\infty - c(e');$$

(ii) the belief about about ability $\pi_\infty \in \Delta(\Theta(e_\infty))$, that is, if $\hat{a}$ is in the support of π_∞, then

$$\hat{a} \in \arg\min_{a' \in \mathbb{R}} K_{\xi_\infty}(f, f_{a'}).$$

The principal's long-run payoff is given by his expected stage-game payoff in the Berk Nash Equilibrium

$$\begin{aligned} U_\infty(\tilde{q}; Q, \pi_0) &= E[y_\infty | Q, \pi_0, q] \\ &= \int_{a \in \mathbb{R}} \int_{e \in \mathbb{R}_+} Y(e, Q, a) \mathrm{d}\xi_\infty \mathrm{d}\pi_0. \end{aligned} \tag{4}$$

Let $\gamma \in [0,1]$ be the weight that the principal assigns to his long-run payoff. The principal chooses the agent's perceived project quality $\tilde{q} \in [\underline{q}, \bar{q}]$ to maximize his payoff

$$U^P(\tilde{q}; Q, \pi_0) = (1-\gamma)U_1(\tilde{q}; Q, \pi_0) + \gamma U_\infty(\tilde{q}; Q, \pi_0), \tag{5}$$

where $U_1(\cdot)$ and $U_\infty(\cdot)$ are given by (2) and (4), respectively.

The principal's manipulation strategy is to choose the agent's perceived project quality $\tilde{q} \in [\underline{q}, \bar{q}]$. I call (i) the principal is truth-telling if the agent's perceived project quality matches the true project quality, i.e., $\tilde{q} = Q$; (ii) the principal is overselling if he makes the agent perceive an unrealistically high project quality, i.e., $\tilde{q} > Q$; and (iii) the principal is underselling if he makes the agent perceive an unrealistically low project quality, $\tilde{q} < Q$.

REMARK 1 (Vulnerable to Manipulation)**.** *A natural manipulation case occurs when the agent is naive and takes the principal's claim at face value. A sophisticated agent would discount information provided from a biased source. However, as long as the agent cannot entirely dismiss the principal's information, she is subject to the principal's manipulation and thus the basic driving forces illustrated in this paper remain (see Section 5.2 for an extension to manipulating a strategic agent). Therefore, naivete serves as a useful benchmark and only makes the forces involved stark. Furthermore, the agent's naivete is a good approximation of reality in a large class of interactions. The agent may lack experience, knowledge, or power as opposed to the principal as an authority, an established expert, or a well-accepted social norm. The disadvantaged agent is probably unaware of the principal's strategic motives, rendering them susceptible to manipulation. In this context, naivete is a feature of the model. I maintain the assumption that the agent is vulnerable to manipulation and I grant the principal the power to shape the agent's perception in a particular aspect. I then focus on examining the profitable direction for the principal's manipulation.*

2.1. *Preliminary Results*

To analyze the principal's manipulation strategy, I first characterize the agent's immediate effort and long-run stable effort. In each period, the agent optimizes effort, trading off her perceived return to effort against its cost. By the first-order condition for maximizing (1), the myopically optimal effort $e^*(\tilde{q}, \pi)$ solves

$$\int_{a \in \mathbb{R}} [Y_e(e, \tilde{q}, a) - c_e(e)] \mathrm{d}\pi(a) = 0. \tag{6}$$

By *Assumption 2*, the solution of (6) exists, and is positive and unique, for any perceived project quality $\tilde{q}$ and belief about the agent's ability π. Therefore, the agent always plays pure strategy and exerts effort in the project. Moreover, since $Y_{eq} > 0$, the optimal effort $e^*(\tilde{q}, \pi)$ strictly increases in the perceived project quality $\tilde{q}$.

Fix agent's perceived project quality $\tilde{q} \in [\underline{q}, \bar{q}]$. The agent's immediate effort

$$e_1 = e^*(\tilde{q}, \pi_0) \tag{7}$$

is determined by her perceived project quality $\tilde{q}$ and prior belief on her own ability π_0. Additionally, a higher perceived project quality $\tilde{q}$ corresponds to a higher marginal return to effort. Since the marginal cost of effort is increasing, a higher perceived project quality $\tilde{q}$ stimulates the agent's immediate effort.

The following lemma establishes that any perceived project quality $\tilde{q}$ can pair with some perceived ability $\hat{a}$ to explain the mean output produced by a given effort.[4] In particular, this induced ability is driven downwards to compensate for a higher perceived project quality in explaining output observations.

Lemma 1 (Attribution Error)**.** *For any $(Q, \tilde{q}, A, e) \in [\underline{q}, \bar{q}]^2 \times \mathbb{R} \times \mathbb{R}_+$, there exists a unique $\hat{a} \in \mathbb{R}$ such that*

$$Y(e, \tilde{q}, \hat{a}) = Y(e, Q, A). \tag{8}$$

Additionally, $\hat{a}$ strictly decreases in $\tilde{q}$, and $\hat{a} = A$ if $\tilde{q} = Q$.

Since we have known the agent only plays pure strategy, the stable effort strategy must degenerate at some effort level, $e_\infty > 0$. By the Kullback-Leibler divergence defined in (3), for any perceived ability level $a' \in \mathbb{R}$, we have

$$\begin{aligned} K_{e_\infty}(f, f_{a'}) &= E_{f(\cdot|e_\infty)}\left[\frac{(y - Y(e_\infty, \tilde{q}, a'))^2 - (y - Y(e_\infty, Q, A))^2}{2\sigma_\varepsilon^2}\right] \\ &= \frac{[Y(e_\infty, \tilde{q}, a') - Y(e_\infty, Q, A)]^2}{2\sigma_\varepsilon^2}. \end{aligned} \tag{9}$$

By *Lemma 1*, there exists a unique $\tilde{a} \in \mathbb{R}$ such that the agent's expected output matches the actual mean output, i.e.,

$$Y(e_\infty, \tilde{q}, \tilde{a}) = Y(e_\infty, Q, A). \tag{10}$$

Therefore, the ability $\tilde{a}$ uniquely minimizes the Kullback–Leibler divergence given by (9),

$$\Theta(\xi_\infty) = \Theta(\delta_{e_\infty}) = \{\tilde{a}\}.$$

The stable belief is thus degenerate at $\tilde{a}$, i.e., $\pi_\infty = \delta_{\tilde{a}}$, and the stable effort is

$$e_\infty = e^*(\tilde{q}, \tilde{a}),$$

where $\tilde{a}$ is given by (10).

[4]All proofs are omitted in the main text and can be found in the appendix.

In the stable state, the agent's belief is self-fulfilling: the agent perceives the project quality is $\tilde{q}$ and concludes with certainty that her ability is $\tilde{a}$; under such perception, she exerts the myopically optimal effort $e_\infty = e^*(\tilde{q}, \tilde{a})$, which yields an output distribution $N(Y(e_\infty, Q, A), \sigma_\varepsilon^2)$ that exactly matches her perceived output distribution $N(Y(e_\infty, \tilde{q}, \tilde{a}), \sigma_\varepsilon^2)$.

REMARK 2 (Degenerate Stable Belief)**.** *In this model, Berk Nash equilibrium boils down to self-confirming equilibrium (Fudenberg and Levine, 1993). Furthermore, since the stable belief is degenerate — that is, the agent concludes with certainty that her ability is $\tilde{a}$ — this stable state is robust to perturbation. In general, the asymptotic belief for learning under a misspecified model only minimizes the Kullback-Leibler divergence between the perceived distribution and the actual distribution (under regularity conditions, see Berk (1966)). The asymptotic belief does not necessarily make these two distributions identical and it may not be degenerate.*

3. IMMEDIATE EFFECT VS LONG-RUN EFFECT

3.1. *Benchmark*

Now we are in a position to analyze the principal's manipulation strategy. To study the profitable direction of manipulation, I first provide a benchmark in which the agent knows her true ability A and the true project quality Q. Denote e^{FB} as the agent's optimal effort under complete information, which is given by

$$e^{FB} = e^*(Q, A).$$

The following lemma shows that if the principal is truth-telling, then the long-run stable state replicates the complete information benchmark; that is, the agent eventually learns the truth and exerts her first-best effort under a correctly specified model.

LEMMA 2 (Correct Learning)**.** *For any $(Q, A, \pi_0) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \Delta(\mathbb{R})$, if $\tilde{q} = Q$, then $\tilde{a} = A$ and $e_\infty = e^{FB}$.*

Can the principal manipulate the agent's perceived project quality $\tilde{q}$ and earn a profit above his truth-telling strategy? To address this concern, the principal needs to consider both the immediate effect and the long-run effect of the manipulation. I will show that, if the principal is sufficiently impatient, he always chooses to oversell the project; in comparison, if the principal is sufficiently forward-looking, his manipulation strategy is determined by the shape of the production function $Y(\cdot)$. For the latter case, I provide sufficient conditions on $Y(\cdot)$ that justify different directions of the manipulation strategy. These main results hold true for any product quality Q, agent's ability A, and the prior on ability π_0.

3.2. *Immediate Effect*

The immediate effect of manipulation is straightforward. Since $e_1 = e^*(\tilde{q}, \pi_0)$ strictly increases in $\tilde{q}$, overselling the quality of the project stimulates the agent's immediate effort and thus enhances the principal's immediate expected payoff.

3.3. *Long-Run Effect*

To see how the stable effort $e_\infty = e^*(\tilde{q}, \tilde{a})$ varies when the principal manipulates $\tilde{q}$, observe that

$$e'_\infty(\tilde{q}) = \frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{q}} + \frac{\partial e^*(\tilde{q}, \tilde{a})}{\partial \tilde{a}} \frac{\partial \tilde{a}}{\partial \tilde{q}}. \tag{11}$$

The first term corresponds to the direct effect of perceived project quality in stimulating effort, and it follows from $Y_{eq} > 0$ that $\partial e^*/\partial \tilde{q} > 0$. The second term corresponds to the indirect effect of $\tilde{q}$ in distorting the agent's learning about her own ability. We know by *Lemma 1* that the agent's perceived ability decreases in the perceived project quality i.e., $\partial \tilde{a}/\partial \tilde{q} < 0$. The question remains how the agent's optimal effort varies with her perceived ability. Apparently, if $Y_{ea} \leq 0$, then a lower perceived ability incentivizes the agent to work harder, i.e., $\partial e^*/\partial \tilde{a} < 0$. In this case, $e'_\infty(\tilde{q}) > 0$, and the principal thus benefits from overselling in the long run as well. The following lemma states this result.[5]

Lemma 3 (Self-Reinforcing Learning). *If $Y_{ea} \leq 0$, the stable effort e_∞ strictly increases in the perceived project quality $\tilde{q}$. In particular, $e_\infty > e^{FB}$ if the principal oversells the project $\tilde{q} > Q$.*

However, if instead $Y_{ea} > 0$ – that is, the return to effort strictly increases in the agent's ability — it is not obvious a priori whether, or which direction of, manipulation can promote the long-run effort. This is where the shape of the production function matters.

3.3.1. *Three Classes of Production.* It is easy to see that any production function takes the form $Y(e, q, a) = V(e, S(q, a), R(q, a))$ where $V_S > 0$, $V_{eS} > 0$ and $V_{eR} \geq 0$.[6] Depending on how $R(\cdot)$ is specified, I distinguish three classes of production as follows.

Definition 2 (Sufficient Statistic). *I call production featuring a sufficient statistic if the production function takes the form $Y(e, q, a) = V(e, S(q, a))$ where $V_S > 0$, $V_{eS} > 0$.*

Here $S(q, a)$ is a sufficient statistic of all fixed factors beyond the agent's control — the project quality q and the agent's ability a. It summarizes the combined contribution of

[5] Heidhues, Kőszegi and Strack (2018) analyze this case in which $sgn(Y_{ea}) \neq sgn(Y_{eq})$ for an overconfident agent (fixing $\tilde{q} > Q$ in my setting). They show that the agent engages in self-defeating learning: being able to adjust her actions based on what she learns drives the agent further away from her optimal choice. This corresponds to $e^*(\tilde{q}, \tilde{a}) > e^*(\tilde{q}, A) > e^*(Q, A)$ in my setting.

[6] To see this, one can take $S(q, a) = q$ and $R(q, a) = a$.

the quality q and ability a to output. I call $S(q,a)$ the stimulating factor since the marginal return to effort strictly increases in S ($V_{eS} > 0$). Here is a simple example of production featuring a sufficient statistic.

EXAMPLE 1 (Sufficient Statistic)**.** *$Y(e,q,a) = qae$.*

In this example, the contribution of project quality q and the agent's ability a can be summarized by the stimulating factor $S(q,a) = qa$, and the production function can be rewritten as

$$Y(e,q,a) = V(e,S(q,a)) = S(q,a)e.$$

DEFINITION 3 (Ability-Laden Production)**.** *I call production ability-laden if the production function takes the form $Y(e,q,a) = V(e,S(q,a),a)$ where $V_S > 0, V_a > 0$, $V_{eS} > 0$, and $V_{ea} \leq 0$.*

For ability-laden production, ability affects the output through two channels:

(i) the instrumental value in enhancing effort through the stimulating factor $S(q,a)$;
(ii) the separate direct value contributing to the output, manifested by $V_a > 0$.

Observe that if $Y_{ea} \leq 0$, then the project falls into the ability-laden production by taking $S(q,a) = q$; that is, there is no instrumental value of ability. Here I give another example of ability-laden production in which $Y_{ea} > 0$.

EXAMPLE 2 (Ability-Laden Production)**.** *$Y = qae + a$.*

In this example, the output features two additively separable terms: the instrumental value of ability a is manifested by the stimulating factor $S(q,a) = qa$ in the first term, and the separate direct value of a is manifested by the second term. The production function can be rewritten as

$$Y(e,q,a) = V(e,S(q,a),a) = S(q,a)e + a.$$

DEFINITION 4 (Quality-Laden Production)**.** *I call production quality-laden if the production function takes the form $Y(e,q,a) = V(e,S(q,a),q)$ where $V_S > 0, V_q > 0$, $V_{eS} > 0$, and $V_{eq} \leq 0$.*

Quality-laden production resembles ability-laden production except that now product quality q has a separate direct value to output instead of the agent's ability a. Likewise, I provide an example of quality-laden production.

EXAMPLE 3 (Quality-Laden Production)**.** *$Y = qae + q$.*

The three classes of production functions provided above are mutually exclusive but not exhaustive. They present sufficient conditions on primitives of production function such that the long-run effect of manipulation is definite, as shown in the following proposition. Specifically, if the production features a sufficient statistic, the impact of manipulation on effort vanishes in the long run. In contrast, ability-laden production encourages overselling whereas quality-laden production encourages underselling.

Proposition 1 (Long-Run Effect of Manipulation). *For any* $(Q, A, \pi_0) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \Delta(\mathbb{R})$, *we have:*

(i) If the production features a sufficient statistic, then $e_\infty = e^{FB}$ *for any* $\tilde{q} \in [\underline{q}, \bar{q}]$.
(ii) If the production is ability-laden, then $(e_\infty - e^{FB})(\tilde{q} - Q) > 0$ *for any* $\tilde{q} \neq Q$.
(iii) If the production is quality-laden, then $(e_\infty - e^{FB})(\tilde{q} - Q) < 0$ *for any* $\tilde{q} \neq Q$.

As stated earlier, if $Y_{ea} \leq 0$, the production is ability-laden. Therefore, part (ii) of *Proposition 1* nests *Lemma 3* as a special case. In what follows, I first illustrate the key intuition by a simple example, which nicely knits together the results in *Proposition 1*. I then provide the main mechanism for more general cases where (i) manipulation has no impact on long-run effort, and (ii) manipulation can affect long-run effort.

3.3.2. *An Illustrative Example.* Consider a production function that is linear in the agent's effort, i.e.,

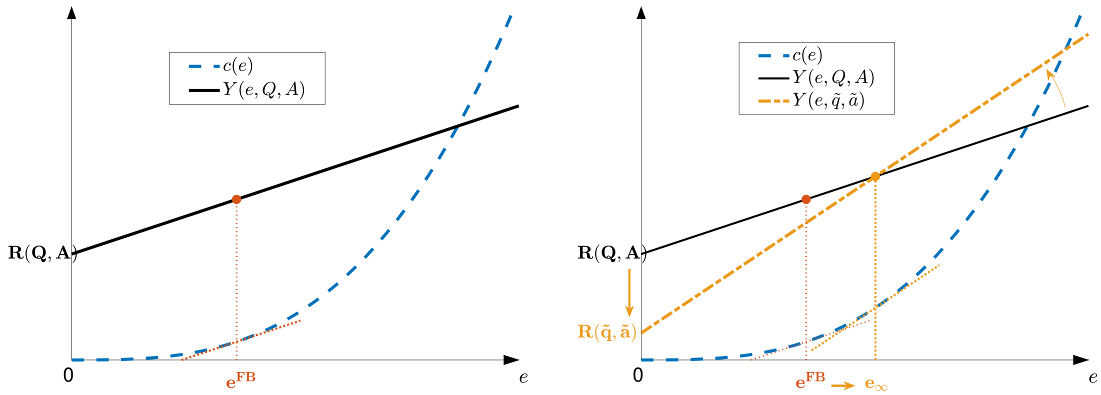
$$Y(e, q, a) = S(q, a)e + R(q, a), \tag{12}$$

where $S(q, a) > 0$ by *Assumption 1* and $S_q(q, a) > 0$ by *Assumption 4*. *Figure 1a* depicts the benchmark case when the agent knows the true project quality Q and her true ability A. In this case, the agent exerts the first-best effort e^{FB}, which equalizes the marginal benefit of effort and its marginal cost.

The principal wants to stimulate the agent's long-run effort e_∞ above e^{FB}. In other words, the principal needs to know how to change the agent's perceived production function $Y(e, \tilde{q}, \tilde{a})$, where $\tilde{a}$ is given by (10), so that $e_\infty = e^*(\tilde{q}, \tilde{a}) > e^{FB}$.

Note first that, since the marginal cost of effort is strictly increasing, the agent needs to perceive a higher marginal return to her effort to justify a higher effort, i.e., $S(\tilde{q}, \tilde{a}) > S(Q, A)$. In the graph, it means the perceived output line $y = Y(e, \tilde{q}, \tilde{a})$ must be steeper than the true output line $y = Y(e, Q, A)$. Second, in the stable state, the perceived mean output needs to match the actual mean output as in (10). Therefore, the perceived output line must cross the true output line at $e_\infty = e^*(\tilde{q}, \tilde{a}) > 0$.

How to satisfy these two conditions simultaneously? The only answer in the graph is to draw down the intercept of the perceived output line and tilt this line counterclockwise (see *Figure 1b*). Now the perceived output line $y = Y(e, \tilde{q}, \tilde{a})$ intersects with the true output line at a point where the marginal cost of effort (i.e., the slope of the cost curve) equals an overly high marginal perceived return of effort (i.e., the slope of the perceived output line



(a) Complete Information Benchmark

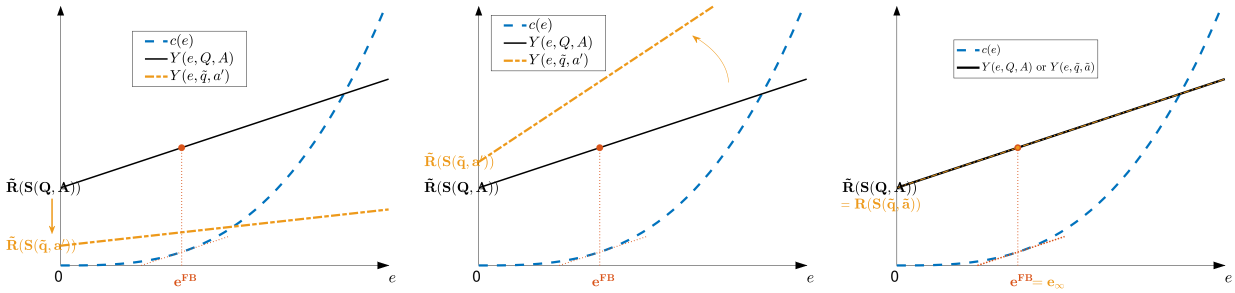
(b) Stimulate Long-Run Effort

FIGURE 1. Linear Production Function

is higher than that of the true output line). Therefore, in order to boost long-run effort, the principal should downplay $R(\tilde{q},\tilde{a}) < R(Q,A)$. This induces an attribution error to account for the output observations, making the agent perceive a higher return to effort $S(\tilde{q},\tilde{a}) > S(Q,A)$.

The economic meaning of this manipulation strategy is as follows. By (12), the output is generated from two parts: one relies on the agent's effort $S(q,a)e$, and the other is independent of the agent's effort $R(q,a)$. The manipulation strategy specified above indicates that the principal should make the agent underestimate the part that is unaffected by her choices, i.e., $R(\tilde{q},\tilde{a}) < R(Q,A)$. As a result, the agent *internalizes* the contribution beyond her control to explain her output observations and perceives a higher agency over the output. In other words, the agent is misled to believe that the output is highly sensitive to her effort and thus she has a higher responsibility for the observed output. As such, the principal successfully shifts the blame from external forces to the agent, making the agent perceive a higher control over the output — this is the key to stimulating effort.

Now the remaining question is: how should the principal manipulate the single factor, namely the perceived project quality $\tilde{q}$, to downplay $R(\tilde{q},\tilde{a})$? The answer depends on what factor is included in $R(\tilde{q},\tilde{a})$. We know that $(\tilde{q}-Q)(\tilde{a}-A) < 0$ by *Lemma 1*. Thus if what remains in the intercept is ability, i.e., $R(q,a) = \tilde{R}(a)$ where $\tilde{R}'(a) > 0$, then the principal should oversell the project $\tilde{q} > Q$ to make the agent perceive a lower ability $\tilde{a} < A$ and thus a lower external influence $R(\tilde{q},\tilde{a}) = \tilde{R}(\tilde{a}) < \tilde{R}(A) = R(Q,A)$ This case belongs to the ability-laden production. If what remains in the intercept is quality (i.e., $R(q,a) = \tilde{R}(q)$ where $\tilde{R}'(q) > 0$), as a case of the quality-laden production, then the principal should undersell the project $\tilde{q} < Q$ so as to lower the agent's perceived external influence $R(\tilde{q},\tilde{a}) = \tilde{R}(\tilde{q}) < \tilde{R}(Q) = R(Q,A)$.



(a) Lower Intercept (b) Higher Slope (c) Stable State

FIGURE 2. Linear Production Function Featuring a Sufficient Statistic

However, if the intercept is strictly increasing in the slope, i.e., $R(q,a) = \tilde{R}(S(q,a))$ where $\tilde{R}'(S) > 0$, then no manipulation strategy can lift the stable effort above e^{FB}. To see this, note that now a higher (or lower) slope of the perceived output line corresponds to a higher (or lower) perceived intercept. In this case, there is no intersection between the perceived output line and the true output line over $e \geq 0$, which violates the condition for the stable state. Therefore, when $R(q,a) = \tilde{R}(S(q,a))$, the perceived output line must coincide with the true output line $Y(e,\tilde{q},\tilde{a}) = Y(e,Q,A)$ and the agent attains her first best by exerting effort e^{FB}. This case belongs to the production featuring a sufficient statistic $S(q,a)$.

In sum, by this simple example, I show that manipulation has no long-run impact on the agent's choice for the production featuring a sufficient statistic, ability-laden production encourages overselling whereas quality-laden production encourages underselling. The overarching principle governing these results is that the principal always intends to downplay the external forces and inflate the agent's perceived return to effort. This blame-shifting mechanism is at the core of the principal's manipulation strategy.

3.3.3. *No Long-Run Impact of Manipulation.* Now I explain *Proposition 1* for general production functions. I first show that, for any production featuring a sufficient statistic — that is, when all fixed factors can boil down to one contributing factor — the agent ultimately exerts her first-best effort, regardless of her prior belief on her ability π_0 and the principal's manipulation strategy $\tilde{q}$. In other words, any manipulation strategy of the principal has no lasting impact on the agent's choices.

The logic behind part (i) of *Proposition 1* is as follows. Suppose $Y(e,q,a) = V(e,S(q,a))$ and $V_S > 0$. In the long-run stable state, given her effort, the agent must have an accurate estimation of the stimulating factor, that is, $S(\tilde{q},\tilde{a}) = S(Q,A)$. Thus, a misspecified assessment of the project quality is fully compensated by its induced assessment of the agent's ability. Overall, the agent acts under the correct incentive and thus exerts her first-best effort; the principal's manipulation yields no long-run effect on the agent's choices.

Example 1 (revisited). In the stable state of this example, the agent fixes her effort $e_\infty > 0$, and the stable belief $\tilde{a}$ must satisfy

$$\tilde{q}\tilde{a} = QA$$

by (10). Thus, the agent ultimately learns the stimulating factor correctly. If the principal oversells the project (i.e., $\tilde{q} > Q$), the agent infers that she has a low ability $\tilde{a} < A$. This misinference completely erases the long-run effect of overselling. Since the agent's effort is pinned down by equalizing the perceived marginal return to effort and the marginal cost of effort, we have

$$c_e(e_\infty) = Y_e(e_\infty, \tilde{q}, \tilde{a}) = \tilde{q}\tilde{a} = QA = Y_e(e^{FB}, Q, A) = c_e(e^{FB}).$$

It follows from $c_{ee} < 0$ that the agent attains her first best in the stable state (i.e., $e_\infty = e^{FB}$).

3.3.4. *Long-Run Impact of Manipulation.* Parts (ii) and (iii) of *Proposition 1* characterize cases in which the principal can stimulate the agent's long-run effort by manipulating her perceived project quality $\tilde{q}$. Specifically, overselling increases long-run effort for ability-laden production, whereas underselling increases long-run effort for quality-laden production.

To see the intuition, let's first walk through the logic for ability-laden production. Recall in the stable state, the agent's perceived mean output matches the actual mean output, i.e.,

$$\begin{aligned} Y(e_\infty, \tilde{q}, \tilde{a}) &= V(e_\infty, S(\tilde{q}, \tilde{a}), \tilde{a}) \\ &= Y(e_\infty, Q, A) = V(e_\infty, S(Q, A), A), \end{aligned}$$

where the first equality and the third equality follow from the definition of ability-laden production. Under overselling, the perceived separate value of ability is unrealistically low (i.e., $\tilde{a} < A$), which leads to a higher perceived stimulating factor $S(\tilde{q}, \tilde{a}) > S(Q, A)$ to account for the observations in the stable state. Since $V_{eS} > 0$, the marginal benefit of effort $Y_e = V_e$ increases in the stimulating factor S. Therefore, overselling heightens the agent's incentive to exert effort in the stable state.

Example 2 (revisited). In this example, confidence in her own ability incentivizes the agent to exert more effort in the project since $S_a > 0$. Therefore, absent from the learning effect, the principal prefers the agent to perceive a higher ability of herself. However, in a long-term relationship, a forward-looking principal takes into account that the agent adjusts her belief about her own ability upon output observations. In this vein, the principal follows the logic above, and prefers instead the agent to underestimate her ability for the sake of boosting long-run effort. The principal thus oversells the project to lower the agent's perceived ability.

On the flip side, when the production is quality-laden, overselling induces a lower long-run effort and underselling becomes advisable to stimulate long-run effort. Analogous to the arguments for the ability-laden production, I rewrite the condition for the stable state as follows:

$$\begin{aligned} Y(e_\infty, \tilde{q}, \tilde{a}) &= V(e_\infty, S(\tilde{q}, \tilde{a}), \tilde{q}) \\ &= Y(e_\infty, Q, A) = V(e_\infty, S(Q, A), Q), \end{aligned}$$

where the first equality and the third equality follow from the definition of quality-laden production. Under overselling, the project quality is perceived to be unrealistically high (i.e., $\tilde{q} > Q$), which leads to a lower stimulating factor $S(\tilde{q}, \tilde{a}) < S(Q, A)$ to satisfy the condition of the stable state. Since the marginal benefit of effort increases in the stimulating factor S, the stable effort is lower than the first-best effort if the principal oversells the project.

3.4. *Weighing Immediate and Long-Run Effects*

To sum up, if the principal only cares about his immediate welfare (i.e., $\gamma = 0$), he would definitely oversell the quality of the project; however, if the principal only cares about his long-run welfare in the stable state (i.e., $\gamma = 1$), his manipulation strategy depends on the production function. For more general cases, the principal's manipulation strategy depends on (i) the principal's time preference, indicated by the weight γ he assigns to the long-run welfare, and (ii) the nature of the project, manifested by the shape of $Y(\cdot)$. The following theorem characterizes how the principal's manipulation strategy is endogenously determined.

THEOREM 1. *Fix any $(Q, A, \pi_0) \in (\underline{q}, \bar{q}) \times \mathbb{R} \times \Delta(\mathbb{R})$.*

- *(i) For the production featuring a sufficient statistic, the principal oversells as long as he cares about his immediate welfare (i.e., $\gamma < 1$). If he only cares about long-run welfare (i.e., $\gamma = 1$), the principal does not profit from any manipulation.*
- *(ii) For the ability-laden production, the principal oversells the project.*
- *(iii) For the quality-laden production, there exists a threshold $\hat{\gamma} \in (0, 1)$ such that: the principal oversells the project if he focuses on immediate welfare $\gamma < \hat{\gamma}$; he undersells the project if he focuses on long-run welfare $\gamma > \hat{\gamma}$; and he tells the true project quality if $\gamma = \hat{\gamma}$.*

By *Theorem 1*, when the principal only cares about long-run welfare ($\gamma = 1$), he tends to be truth-telling or even undersells the project. In light of this result, integrity is optimal for an informed principal who aims to promote his long-run welfare. However, a concern about immediate welfare tilts the scale toward opportunistic choices: once $\gamma < 1$, the principal almost always oversells the project except when the project is quality-laden and the principal is sufficiently patient. This partially explains why conventional wisdom suggests honesty in long-term relations whereas overselling is prevalent in practice.[7]

4. Applications

4.1. *Mentorship*

A natural fit to the model is the mentor-mentee relationships (such as lab advisor versus graduate student) in which the findings (y) from a research project are expected to depend on the value (q) of the research question, the intelligence (a) of the mentee and/or her coworkers, and mentee's effort (e) into the project. Additionally, the marginal return to the mentee's effort increases with her intelligence and the value of the research question. I consider the two types of projects introduced at the start of the paper to examine how manipulation strategy varies with the nature of the project.

Quality-Driven Project. For some projects, all work is directed toward a specific pre-registered question, and thus the findings are primarily determined by the value of this question (project quality). I model the production function of this type of project as follows:

$$Y(e,q,a) = q(e + a + \psi a e), \tag{13}$$

where $\psi \geq 0$ stands for the synergy of the agent's ability and effort. Here, the value q of the research question is a decisive factor in yielding scientific findings.

Note that the production function (13) can be rewritten as

$$Y(e,q,a) = V(e, S(q,a), a) = S(q,a)\left(e + \frac{a}{1+\psi a}\right)$$

where $S(q,a) = q(1+\psi a)$ refers to the intrinsic motivation to study. In light of *Theorem 1*, the mentor should oversell the value of the research project even if he only cares about the long-run research output. Such manipulation makes the mentee underestimate her own ability. However, this humility enhances the mentee's appreciation of the return to effort $S(\tilde{q},\tilde{a}) > S(Q,A)$, thus stimulating her effort in the long run.

Ability-Driven Project. For another type of project, the agent's ability is a dominant factor in determining research findings. I assume

$$Y(e,q,a) = a(e + q + \psi q e), \tag{14}$$

where $\psi > 0$. One can easily see that this case switches the role of project quality and the agent's ability compared to the quality-driven project described above. The production

[7] Competition can be another force to encourage overselling. See *Appendix A.1* for further discussion.

function (14) takes the form

$$Y(e,q,a) = V(e, S(q,a), q) = S(q,a)\left(e + \frac{q}{1+\psi q}\right).$$

By *Theorem 1*, an impatient mentor would oversell the value of the research project to induce a higher immediate effort from the agent. In contrast, a sufficiently patient mentor undersells the value of the research project to stimulate the agent's effort that is sustainable in the long run. For example, the mentor can emphasize the difficulty of the project so that the agent does not blame herself for not producing many findings for the project yet. By underselling the project, the principal raises the agent's perception of her own ability. The boosted confidence enhances the mentee's perceived return to her effort $S(\tilde{q}, \tilde{a}) > S(Q, A)$, thus stimulating her stable effort and promoting research advances.

REMARK 3. *A project may fall between purely quality-driven production and purely ability-driven production. The precise nature of a project is determined by its production function, which needs to be identified by empirical studies. Given the production function, this paper can then be used to illustrate the nature of the project and derive its implications on the manipulation strategy.*

4.2. *Abusive Relationship*

Emotional abuse (such as gaslighting) in professional or intimate relationships can be profoundly damaging, sometimes leading to depression and anxiety that endure for decades (Abramson, 2014; Stern, 2018). The manipulator can perpetuate structural inequalities (such as gender stereotypes in domestic violence) against the victim (Sweet, 2019). In this section, I apply the model to the context of emotional abuse and illustrate how it can be used to gain power in a relationship.

I use the following functional form to analyze an abusive relationship:

$$Y(e,q,a) = S(q,a)e + q.$$

An individual's expected well-being (Y) hinges on her personal quality (q) and her benefit from a relationship — which in turn depends on the individual's effort e in this relationship and her attachment $S(\cdot)$ to it. The relational attachment conveys the marginal value that the individual derives from her effort and it is a function of the individual's personal quality (q) and how fit the partnership is (a).

A manipulative partner induces the individual to doubt her own ability through emotional abuse tactics such as overprotection, blame-shifting, neglecting, isolating, and degrading comments. Emotional abuse erodes the individual's self-esteem and undermines the individual perception of her personal quality; she would rationalize the abuses as if they are, at least partly, due to her weakness ("*I am not good enough to deserve better*"). The individual then misattributes her well-being to the value of this relationship and becomes highly sensitive to it. This increased attachment to the relationship justifies sacrifice and devotion. As such, the manipulative partner can blamelessly exploit more benefits from the individual, entrenching inequality in a relationship.

4.3. *Credibility*

One major concern with self-serving manipulation is its lasting damage to trust (Doney and Cannon, 1997; Thorp, 2020). The basic idea is simple: if one sets others up to false hope by overpromising, then the realized under-delivery would frustrate them from further cooperation.

I consider long-term seller/buyer relationships (e.g., upstream/downstream firms in a supply chain) as a leading application and use *Example 1* to model the trust-building process:

$$Y(e, q, a) = qae.$$

A buyer purchases a product repeatedly from a seller, and the product quality is privately known by the seller. The seller maintains a claim about the product quality. In this context, q stands for the product quality; $a > 0$ stands for the trustworthiness/business integrity of the seller (with $A = 1$ being honesty); e stands for consumption; $S(q, a) = qa$ captures the marginal utility of the consumption. If the seller exaggerates product quality (i.e., $\tilde{q} > Q$), the buyer would lower the assessment of the seller's trustworthiness over time (i.e., $\tilde{a} < A = 1$), i.e., trust is compromised. Eventually, the buyer learns the real product quality (i.e., $\tilde{q}\tilde{a} = S(\tilde{q}\tilde{a}) = S(Q, A) = Q$). Therefore, the immediate effect of manipulation is washed away over time and there is no lasting benefit of overselling.

Furthermore, if the seller values trustworthiness (e.g., trustworthiness may facilitate selling other products), then overselling eventually harms the seller since it ruins the seller's reputation of trustworthiness. To develop product loyalty or strengthen long-term cooperation, truth-telling or even "under-promise, over-delivery" is an advisable management strategy.

The same logic underlies a wide variety of interactions including:

Political Propaganda. A politician claims the effect of a policy ($\tilde{q}$) to earn public support (e) in the form of campaign funding and votes, and the public observes the policy's performances over time. If the politician overpromises the policy effect ($\tilde{q} > Q$), then his credibility perceived by the public wanes ($\tilde{a} < 1$) and the public learns the actual policy effect $Q = \tilde{q}\tilde{a}$ in the long run. A natural escape from this pattern for the politician is thus to find some scapegoat to shift the blame and sustain public support. The common blame-shifting strategies include but are not limited to attributing to natural disasters, inciting hatred against a certain group, averting public attention to irrelevant matters, etc.

Firm Funding. A firm presents investors with the prospect of its product ($\tilde{q}$) to increase the funding (e). If the performance of the product does not match the prospect, investors would doubt the trustworthiness of the firm ($\tilde{a} < 1$), and figure out the real value of the product $Q = \tilde{q}\tilde{a}$ in the long run.

Goal Setting. An individual make a daily goal ($\tilde{q}$) for self-motivation. Aiming high stimulates effort (e) in the short run. However, as she fails to fulfill her goals repeatedly, the individual doubts the efficacy of her goal ($\tilde{a} < 1$) and gets less motivated by the goal setting. Ultimately, the individual acts under what she can actually accomplish $\tilde{q}\tilde{a} = Q$. Considering the emotional cost (e.g., depression and self-doubt) inflicted by unsatisfied goals, it may thus be optimal for individuals to set a realistic goal, or even a lower bound to what one can do, in order to build self-efficacy and empower oneself.

5. Extensions

5.1. *A Simplified Three-Period Model*

In the baseline model, I use misspecified learning as a modeling tool to capture the short-run versus long-run tradeoff for the principal. The basic economic mechanism remains at work in a simple three-period framework. I will present this simplified version in this section.

I consider three periods in the interaction between a principal and an agent, $t = 0, 1, 2$. At period 0, the principal influences the agent's perceived project quality $\tilde{q}$. At period 1, the agent chooses her optimal effort e_1 under her perceived project quality $\tilde{q}$ and the prior belief about her own ability π_0. At period 2, the agent observes the output at the previous period $Y(e_1, Q, A)$, updates her belief about her own ability to π_1, and then chooses the optimal effort e_2 under the perceived project quality $\tilde{q}$ and the posterior belief π_1. The stage-game payoff remains the same as in the baseline model, and the agent is still myopic whereas the principal weighs his payoff in period 1 and period 2. Thus payoffs for the agent and the principal are given as follows,

$$U_t^A(e_t; \tilde{q}, \pi_{t-1}) = E[Y_t - c(e_t)|\tilde{q}, \pi_{t-1}], \text{ for } t = 1, 2,$$
$$U_0^P(\tilde{q}; Q, \pi_0) = E[(1-\gamma)Y_1 + \gamma Y_2 | Q, \pi_0],$$

where $\gamma \in [0, 1]$ indicates the weight the principal attaches to his payoff in period 2.

As in the baseline model, the agent's effort strategy is given by $e^*(\cdot)$. Thus her effort at period 1 is given by (7) and is strictly increasing in her perceived product quality $\tilde{q}$.

At period 2, since the agent directly observes the output at period 1 without any noise, she immediately reaches a conclusion about her ability. Formally, by *Lemma 1*, there exists a unique $\tilde{a}_1 \in \mathbb{R}$ such that

$$Y(e_1, Q, A) = Y(e_1, \tilde{q}, \tilde{a}_1), \tag{15}$$

and $\tilde{a}_1$ strictly decreases in $\tilde{q}$. Since the agent observes $Y(e_1, Q, A)$, her posterior of her own ability is degenerate at $\tilde{a}_1$, i.e., $\pi_1 = \delta_{\tilde{a}_1}$. Thus her effort at period 2 is given by

$$e_2 = e^*(\tilde{q}, \tilde{a}_1).$$

In this simple framework, the tension between the short-run and long-run effects of manipulation resurfaces. Overselling the project ($\tilde{q} > Q$) boosts the agent's effort at period 1; but it lowers the agent's assessment of her own ability $\tilde{a}_1$, which potentially diminishes the agent's effort at period 2. Apparently, the immediate effect of manipulation here is exactly the same as in the baseline model. The following proposition states that the long-run effect of manipulation also coincides with *Proposition 1* in the baseline model.

PROPOSITION 2. *For any $(Q, A, \pi_0) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \Delta(\mathbb{R})$, the following statements hold true in the simplified three-period model.*

(i) If the production features a sufficient statistic, then $e_2 = e^{FB}$ for any $\tilde{q} \in [\underline{q}, \bar{q}]$.
(ii) If the production is ability-laden, then $(e_2 - e^{FB})(\tilde{q} - Q) > 0$ if $\tilde{q} \neq Q$.
(iii) If the production is quality-laden, then $(e_2 - e^{FB})(\tilde{q} - Q) < 0$ if $\tilde{q} \neq Q$.

Accordingly, the underlying logic of *Theorem 1* stands. If the production is weakly ability-laden, then the principal always profits from overselling the project, regardless of her time preferences γ; however, if the production is quality-laden, then the side effect of overselling arises, even to an extent where "under promise, over delivery" becomes a winning management strategy.

REMARK 4. *As the preceding arguments reveal, the simplified three-period model perfectly captures the immediate versus long-run tradeoff demonstrated in Section 3. The baseline model has additional merit in that it illustrates how manipulation perpetuates itself and persists over time. In the game where the agent observes outputs noiselessly, two periods of output observation suffice for the agent to detect that the principal is cheating (i.e., $\tilde{q} \neq Q$) and undo the manipulation. In other words, manipulation is alive for at most two periods. To understand long-term manipulation, the interesting case is thus to study how manipulation operates under imperfect output signals. The baseline model shows that the agent's misspecified learning, triggered by the principal, can result in a stable state in which the agent engages in closed-loop reasoning. In the stable state, the agent acts optimally under her perceived project quality and perceived ability, leading to an output distribution that perfectly matches her perceived output distribution. As such, noisy observations enable the principal to manipulate the agent in a long-term relationship; now the effect of manipulation can last permanently.*

5.2. *Perception Inception*

I have shown in what direction the principal *aims* to manipulate the agent's perceived project quality, and how this decision depends on his time preference and the nature of the project. One may wonder how such manipulation is possible in the first place, especially for an ability-laden project where an incentive to oversell the project is clearly present. In this section, I give a simple model of how a principal oversells his ability-laden project to an agent.

There are two levels of project quality, $q \in \{0, 1\}$. The principal knows the project quality whereas the agent is uncertain about the project, with the prior on $q = 1$ being $p_q = 1/2$. The production function is given by *Example* 2. The cost function is $c(e) = e^2/2$. The principal and the agent share the same prior on the agent's ability, which is uniformly distributed over $[0, 2]$, i.e., $\pi_0 = U[0, 2]$. In this case, the agent's myopically optimal effort is given by her expected project quality. Without any information from the principal, it is $e^*(p_q, \pi_0) = p_q = 0.5$.

The principal reports that the project quality is $\tilde{q} \in \{0, 1\}$. There are two types of principals, honest (H) or manipulative (M). Misreporting the project quality ($\tilde{q} \neq q$) inflicts a huge moral cost to an honest principal whereas it inflicts no moral cost to a manipulative principal. As a consequence, an honest principal always reports the true project quality whereas a manipulative principal misreports the project quality in his favor.[8] The agent is uncertain about the principal's type, and her prior on the principal being honest is $p_h \in [0, 1]$.

The agent is aware that a manipulative principal profits from overselling this project in this environment. Her posterior mean of the project quality after observing a report $\tilde{q} = 1$ is thus

$$p'_q = \frac{p_q}{p_q + (1 - p_q)(1 - p_h)} = \frac{1}{2 - p_h},$$

which is strictly increasing in p_h. If the perceived probability of meeting an honest principal $p_h = 1$, we have $p'_q = 1$. This case can be used to capture a naive agent, who always takes the principal's words at face value. If $p_h = 0$, then $p'_q = p_q = 1/2$. This case can be used to describe a skeptic, who always disregards any shared knowledge. As long as the agent perceives a slight chance that the principal is honest ($p_h > 0$), then her posterior mean rises upon observing a good report $\tilde{q} = 1$ (i.e., $p'_q > p_q = 1/2$). A manipulative principal can thus exploit this pattern to oversell a project (i.e., $q = 0$, $\tilde{q} = 1$).

[8] Readers can view the honest type versus the manipulative type of the informed principal as the committed (or behavioral) type versus payoff (or rational) type in the reputation literature (McKelvey and Palfrey, 1992). The framework can also be generalized to the case where there is a continuum of types of principal on the moral scale, each bearing different costs from cheating, and the agent is uncertain about the principal's moral type (i.e., his cost incurred by cheating).

In this respect, uncertainty about the informed principal's type invites manipulation. The agent is more vulnerable to manipulation if (i) she is more uncertain about the true project quality — marked by higher entropy in p_q; (ii) she tends to believe the good intention of other people, or the sort of relationship is generally supposed (or purposefully pretend) to be reliable — marked by a higher p_h.

6. Discussion

6.1. *Related Literature*

Misspecified Learning. The literature on misspecified learning posits that individuals update their beliefs from observations when the prior is misspecified in the sense that it assigns zero possibility to the true state. The majority of this literature takes a certain misspecified model as exogenously given and analyzes its implications. In this paper, I endogenize model misspecification in a principal-agent framework. Here, the principal exploits the agent's lack of information and affects her learning model (i.e., her perception). Analysis of the principal's manipulation strategy thus informs the specific model misspecification (i.e., misperception) that is fostered within a given environment.

The misspecified learning literature is pioneered by Berk (1966) and has received increasing attention in economics since Esponda and Pouzo (2016). Berk (1966) examines the asymptotic distribution of a parametric estimate under a possibly misspecified model. The author shows that, under some regularity conditions, the asymptotic distribution is confined to a set of values that minimize the divergence between the subjective distribution of data and the objective distribution. Esponda and Pouzo (2016) introduce this concern for misspecified learning in economic theory and propose the notion of *Berk-Nash Equilibrium* for settings where a single agent or multiple players hold potentially misspecified views (i.e., subjective models) of their decision environments.

This literature primarily focuses on three strands. The first strand provides techniques to analyze the asymptotic properties of a misspecified learning process, especially deriving conditions for belief convergence (Nyarko, 1991; Fudenberg, Romanyuk and Strack, 2017; Esponda, Pouzo and Yamamoto, 2021; Esponda and Pouzo, 2021; Heidhues, Kőszegi and Strack, 2021; Fudenberg, Lanzani and Strack, 2021, 2023; Frick, Iijima and Ishii, 2023). The second strand of the literature studies the implications of misspecified learning in varied applications such as individual learning under behavioral biases (Gervais and Odean, 2001; Heidhues, Kőszegi and Strack, 2018; Gagnon-Bartsch and Bushong, 2022; Bohren and Hauser, 2023), learning about oneself (Kőszegi, Loewenstein and Murooka, 2022; Heidhues, Kőszegi and Strack, 2023), social learning (Andreoni and Mylovanov, 2012; Bohren, 2016; Frick, Iijima and Ishii, 2020; Bohren and Hauser, 2021; He, 2022; Ba and Gindin, 2023), and political cycles (Levy, Razin and Young, 2022). The third strand investigates model selection among competing models, asking (1) which (misspecified) model persists over time (He and Libgober, 2020; Fudenberg and Lanzani, 2023; Ba, 2023), and (2) how the model selection is determined by the learning environment such as available sample size (Montiel Olea et al., 2022) and the complexity of information structure (Ba, Bohren and Imas, 2022). By comparison, I study model selection in a principal-agent framework. Assuming that individual perception (i.e., learning model) is malleable, I examine how a particular party (i.e., the principal), who is informed but has conflicting interests with the agent, can gain advantages through perception manipulation. In this context, I show that the agent's learning model crucially depends on the principal's patience and the properties of work at hand.

The most closely related paper is Heidhues, Kőszegi and Strack (2018). In their paper, an overconfident agent (i.e., $\tilde{q} > Q$) engages in the misspecified learning about a fundamental factor (i.e., a) upon output observations. Under the assumption that overconfidence is exogenously given and perceived ability and fundamental factor change the marginal return to effort in the opposition directions (i.e., $Y_{eq} > 0, Y_{ea} \leq 0$), they show that the agent's active learning is self-defeating — it results in a belief that is further away from the truth and a worse decision relative to the case of fixed action. My paper endogenizes the agent's perception and shows that the agent's overconfidence favors the principal in (self-)manipulation for the type of project that they study (i.e., $Y_{eq} > 0, Y_{ea} \leq 0$). Furthermore, I generalize the production function and allow for both factors to be complements to effort (i.e., $Y_{eq} > 0$ and $Y_{ea} > 0$). This case is nuanced due to the conflicting long-run effects of misspecification. For this case, I provide sufficient conditions on the production function to identify projects that foster unrealistically high expectations and those that foster unduly low expectations.

***Theories on Motivated Belief*.** Although this paper focuses on perception manipulation in interpersonal contexts, the analysis naturally extends to an intrapersonal setting where a former self (planner/principal) seeks to motivate subsequent selves (doer/agent) to exert effort. Such self-manipulation links to the literature of motivated beliefs (Akerlof and Dickens, 1982; Akerlof and Kranton, 2000; Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005; Kőszegi, 2006; Bénabou and Tirole, 2011; Dillenberger and Sadowski, 2012; Gottlieb, 2014; Bénabou and Tirole, 2016; Battigalli and Dufwenberg, 2022). This literature posits that the individual belief results from a tradeoff between *accuracy*, which is required by pure rationality, and *preference*, which captures the intrinsic value attached to a particular belief. It thus identifies a middle ground between the classic rational paradigm and the typical behavioral paradigm marked by built-in bounded rationality.

Furthermore, motivated beliefs can be perpetuated through self-deception (i.e., strategically manipulating one's own information). Typical self-deception strategies include wishful thinking, information avoidance, self-signaling, selective memory, and post rationalization (Carrillo and Mariotti, 2000; Bodner and Prelec, 2003; Bénabou and Tirole, 2004; Bénabou, 2013; Levy, 2014; Eyster, Li and Ridout, 2021). Aligned with this literature, this paper emphasizes the instrumental value of belief in incentivizing effort. I study how belief is managed and maintained in interpersonal and intrapersonal settings. Moreover, this paper demonstrates the long-run effect of misperception through the lens of misspecified learning.

***Evidence for Perception Manipulation*.** This paper hinges on this assumption that the agent's perception (her learning model) can be influenced by oneself or others. Empirical studies, primarily in psychology and neuroscience, reveal several self-manipulation mechanisms and show that individuals can interpret their experience strategically to motivate themselves, which breeds systematic cognitive biases (e.g., Gur and Sackeim, 1979; Pintrich, 2004; Sharot, Korn and Dolan, 2011; Eil and Rao, 2011; Schacter, 2012; Korn et al., 2014; Gin et al., 2021; Bolotnyy, Basilico and Barreira, 2022; Fan and Bolte, 2022). For example, experiments in Norem and Cantor (1986) show that pessimism can be used to cope with anxiety. Individuals may thus display defensive pessimism, and correcting pessimism blindly impairs their performances. Von Hippel and Trivers (2011) demonstrate various types of self-deception involving simultaneous awareness and ignorance within an individual. From an evolutionary approach, they posit that self-deception can be used to facilitate interpersonal deception, and this hypothesis is confirmed by experiments in Schwardmann and Van der Weele (2019). At the group level, public opinions are often influenced by social media and political propaganda (e.g., Bosmajian, 1974; Klemperer, 2002; Anand, Ashforth and Joshi, 2004; DellaVigna and Kaplan, 2007; Bazerman and Tenbrunsel, 2011). Collective misperception can lead to detrimental outcomes, such as sugarcoating discriminative practices due to entrenched prejudices, or building a booming financial bubble due to inflated confidence in the market.

In the paper, I emphasize the role of perceived agency in decision-making. The linkage between one's motivation and the (subjective) contingency of outcomes on their choices has been well explored in psychology. Experiments (Alloy and Abramson, 1979; Quattrone and Tversky, 1984) find that non-depressed individuals exhibit self-serving motivational bias: they overestimate the causal relationship between a desired outcome and their choices, and underestimate the degree of contingency when the outcome is undesired. Conversely, evidence on learned helplessness (Hiroto and Seligman, 1975; Maier and Seligman, 1976) emphasizes the harms of uncontrollability: subjects (including dogs, mice, fish, and humans) who have learned that (aversive) outcomes are noncontigent on their choices display lower incentives for initiating responses and become depressed. The possible departure of one's subjective controllability from reality opens the door for manipulation. This paper illustrates how the principal can exploit information asymmetry to manipulate the agent's perceived controllability and induce more effort from the agent.

***Bayesian Persuasion** vs **Narrative Provision**.* In the literature on Bayesian persuasion and information design (e.g., Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), the informed party chooses the signaling scheme and thus directly controls the data-generating process. The commitment power in signal choices enables the informed party to influence the uninformed party, and the induced posteriors must follow Bayesian plausibility. In this paper, I posit alternatively that the informed party cannot control the signal that the other party *observes*; instead, the informed party, acting as an expert or authority, can influence how the other party *interprets* their observed signals.

Several recent papers capture this idea of controlling interpretation as providing a *narrative*; namely, a "causal story" that maps actions into consequences (e.g., Eliaz and Spiegler, 2020; Aina, 2023). In contrast to providing an entire narrative, the principal in my paper merely manipulates the agent's perception of a single factor, leaving the agent to deduce the remaining factors to complete her subjective story. Besides, I emphasize the long-run (stable) effects of perception manipulation in an infinite-horizon framework.

6.2. *Concluding Remarks*

I propose a model of manipulation in which the principal can influence how the agent perceives the project at hand and thus affect her learning about her own ability. Overselling project quality stimulates immediate effort but undermines the agent's perceived ability, which potentially demotivates effort in the long run. I characterize how the direction of manipulation (and accordingly the agent's model misspecification) is endogenously determined by the principal's time preference and the nature of the project. Specifically, I identify three types of projects: (i) if the production features a sufficient statistic, such that the impact of project quality and the agent's ability are not separable, then the agent eventually exerts her first-best effort, regardless of the principal's manipulation strategy; (ii) if the production is ability-laden, in the sense that ability has a separate value to output independent of the agent's effort, then overselling stimulates the agent's long-run effort; (iii) if the production is quality-laden, in the sense that quality has a separate value to output, then underselling stimulates long-run effort.

Importantly, this paper underscores a central mechanism of manipulation. By downplaying external contribution that is independent of the agent's effort, the principal frames a misspecified model for the agent. Using this model, the agent internalizes contribution from external forces and thus perceives a seemingly better control over the output. Higher agency over the project implies higher responsibility for its output and a higher return to effort. As such, the agent becomes motivated to exert excessive effort that is more aligned with the principal's interests.

This paper paves the way for two promising avenues of future research. The first line of research is to investigate how the agent recognizes and breaks from manipulation. Subsequent research could consider a forward-looking agent who experiments with different models or effort levels to improve her learning model. One may also think of an agent who sets boundaries for her learning results. For instance, the agent has upper and lower bounds on her own ability and she quits if the inferred ability hits either bound. The second line of research can proceed by combining a certain communication protocol (cheap talk/signaling/Bayesian persuasion) with misspecified learning. The current paper is the first attempt, assuming that the agent fully trusts the principal and takes what the principal states about the project quality at face value. It adopts a reduced-form approach to identify the principal's profitable direction of manipulation. Future work can examine the extent to which manipulation can be sustained by a specific communication protocol where the agent is sophisticated in deciphering the stated project quality.

Appendices

Appendix A. Further Discussion

A.1. *Competing Principals*

In the baseline model, I study the case in which a single principal induces effort from the agent. In this appendix, I discuss how his manipulation strategy changes when the principal needs to compete with another principal in exploiting the agent's effort. Examples include marketing, political campaigns, and startup firms competing for venture capital.

Consider a scenario with two principals $i = 1, 2$ and one agent. Principal i has a project of quality $Q_i \in [\underline{q}_i, \bar{q}_i]$. Q_i is privately known to principal i while others share the same prior $F_i \in \Delta[\underline{q}_i, \bar{q}_i]$ about project i's quality. The principals first simultaneously influence the agent's perception of the project quality, and the agent selects one project of higher perceived quality to attend. Conditional on being selected, principal i's payoff remains (5); otherwise, his payoff is 0.

If $\underline{q}_i > \bar{q}_j$ ($i, j = 1, 2$), then it is publicly known that project i is absolutely better than project j. In this case, the agent selects project i for sure, and thus the manipulation strategy of principal i is determined as in the baseline model exempt from competition pressure.

As the overlapping of $[\underline{q}_1, \bar{q}_1]$ and $[\underline{q}_2, \bar{q}_2]$ grows, the competition between these two projects is intensified. To the extreme, the two projects are ex-ante identical, i.e., $[\underline{q}, \bar{q}] \equiv [\underline{q}_1, \bar{q}_1] = [\underline{q}_2, \bar{q}_2]$, $F \equiv F_1 = F_2$. This is when the competition between the two projects attains its peak. Now to attract the agent to work for them and avoid being left unchosen, principals have an additional incentive to oversell their projects on top of the immediate versus long-run consideration illustrated in *Section 3*.

For example, consider the case when both projects have the same quality ($Q \equiv Q_1 = Q_2$). Regardless of the production function, the competition pressure encourages overselling. To see this, if principal i claims that his project is of quality $\tilde{q}_i \in [\underline{q}, \bar{q})$, then the other principal j can outcompete him by claiming a higher quality. Therefore in equilibrium, both principals oversell their projects as far as they can, i.e., $\tilde{q}_1 = \tilde{q}_2 = \bar{q}$. Note this is true even if the principal originally prefers to be truth-telling without competition.

Herein, instead of enhancing market efficiency, competition nurtures overselling and muddles the information provision, leaving all players worse off.

A.2. *Immediate Effect vs Short-Run Effect*

In this section, I show that the tension between quality promise and ability frustration already arises for two periods, even under noisy signals. Therefore, for the concern on misinference to matter, the principal does not need to be sufficiently patient to care about the long-run stable state where the belief converges. Additionally, the main results in the paper hold generally for this short-run analysis as well.

Define the evidence-oriented effort $\hat{e}_2$ as

$$\hat{e}_2 = e^*(\tilde{q}, \tilde{a}_1),$$

where $\tilde{a}_1 \in \mathbb{R}$ stands for the stable belief on the ability when the action is fixed at the immediate effort e_1 and is given by

$$Y(e_1, Q, A) = Y(e_1, \tilde{q}, \tilde{a}_1).$$

Alternatively, $\tilde{a}_1$ can be interpreted as the degenerate posterior on ability when the agent observes the output without noise as in *Section 5.1*.

Consider a short-sighted principal, who only cares about outputs in the first two periods. The principal's payoff is given by

$$U^p = (1-\delta)E[y_1] + \delta E[y_2] = (1-\delta)Y(e_1, Q, A) + \delta E[Y(e_2, Q, A)],$$

where $\delta \in (0,1)$, $e_1 = e^*(\tilde{q}, \pi_0)$ and $E[Y(e_2, Q, A)] = w_2 Y(\hat{e}_2, Q, A) + (1-w_2)Y(e_1, Q, A)$. The weight $w_2 \in (0,1)$ increases with the precision of output observation and decreases with the precision of the agent's prior on her ability.

We already know that $e_1 = e^*(\tilde{q}, \pi_0)$ can be strictly boosted by overselling the project. Now, I provide sufficient conditions on the primitives of the production function to compare $\hat{e}_2 = e^*(\tilde{q}, \tilde{a}_1)$ and $e^{FB} = e^*(Q, A)$. This result illustrates the incentive for the short-sighted principal to mislead the agent.

Proposition A.1 (Short-Run Effect of Misspecification)**.** *For any* $(Q, A, \pi_0) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \Delta(\mathbb{R})$*, we have:*

(i) If the production features a sufficient statistic, then $\hat{e}_2 = e^{FB}$ *for any* $\tilde{q} \in [\underline{q}, \bar{q}]$.

(ii) If the production is ability-laden, then $(\hat{e}_2 - e^{FB})(\tilde{q} - Q) > 0$ *if* $\tilde{q} \neq Q$.

(iii) If the production is quality-laden, then $(\hat{e}_2 - e^{FB})(\tilde{q} - Q) < 0$ *if* $\tilde{q} \neq Q$.

Note that the evidence-oriented action $\hat{e}_2$ is exactly the same as $e_2 = e^*(\tilde{q}, \tilde{a}_1)$ in the simplified three-period model in *Section 5.1*. Therefore, *Proposition A.1* follows from *Proposition 2*, whose proof is given by *Appendix B.4*.

Again, if the project is ability-laden, then the principal always gains from overselling the project, regardless of her time preferences δ; however, if the project is quality-laden, then the side effect of overselling arises, and underselling becomes advisable if the principal is sufficiently future-oriented.

Appendix B. Omitted Proofs

B.1. *Proof of Lemma 1*

Fix any arbitrary $(Q, \tilde{q}, A, e) \in [\underline{q}, \bar{q}]^2 \times \mathbb{R} \times \mathbb{R}_{++}$.

Existence. It follows from the mean production function $Y(\cdot)$ being twice continuously differentiable that $Y(\cdot)$ is continuous. Additionally, by *Assumption 3*, $\lim_{a \to -\infty} Y(e, q, a) = -\infty$, $\lim_{a \to \infty} Y(e, q, a) = \infty$. Therefore, by the intermediate value theorem, there exists $\hat{a} \in \mathbb{R}$ that satisfies (8).

Uniqueness. Suppose, for the purpose of contradiction, that there exists an alternative $a' \neq \hat{a}$ satisfying (8). Then since $Y_a > 0$, we have

$$Y(e, Q, A) = Y(e, Q, \max\{\hat{a}, a'\}) > Y(e, Q, \min\{\hat{a}, a'\}) = Y(e, Q, A),$$

which is a contradiction. Therefore, there exists a unique $\tilde{a}$ that satisfies (8).

Variation. It follows from $Y_a > 0$ that, if $\tilde{q} = Q$, then $\hat{a} = A$.

Fix $(Q, A, e) \in [\underline{q}, \bar{q}] \times \mathbb{R} \times \mathbb{R}_{++}$, and write the induced ability $\hat{a}$ as $\hat{a}(\tilde{q})$ to emphasize its dependence on $\tilde{q}$. For any $\tilde{q} \in [\underline{q}, \bar{q}]$, we have

$$Y(e, \tilde{q}, \hat{a}(\tilde{q})) = Y(e, Q, A).$$

Differentiating both sides of the equation above with regard to $\tilde{q}$ gives

$$Y_q(e, \tilde{q}, \hat{a}(\tilde{q})) + Y_a(e, \tilde{q}, \hat{a}(\tilde{q}))\hat{a}'(\tilde{q}) = 0.$$

Then $Y_q(\cdot) > 0, Y_a(\cdot) > 0$ imply that $\hat{a}'(\tilde{q}) < 0$, that is the induced ability $\hat{a}$ strictly decreases in perceived project quality $\tilde{q}$.

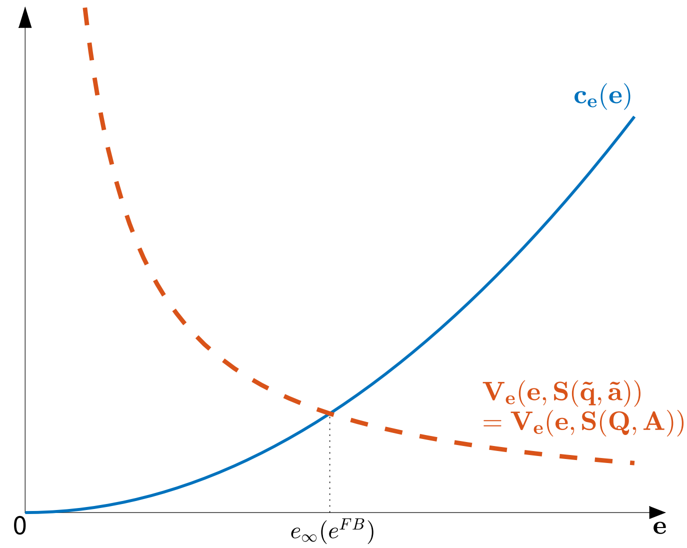


FIGURE B.1. No Long-Run Effect of Misspecification

B.2. *Proof of Lemma 2*

Lemma 2 directly follows from the definition of the stable belief $\tilde{a}$ (10) and *Lemma 1* by taking $e = e_\infty$, $\tilde{q} = Q$ and $\hat{a} = \tilde{a}$.

B.3. *Proof of Proposition 1*

I prove the long-run effect of overselling. The long-run effect of truth-selling is given by *Lemma 2*, and the long-run effect of underselling is analogous to the proof below.

Fix any arbitrary $(Q, A, \pi_0) \in [\underline{q}, \bar{q}) \times \mathbb{R} \times \Delta(\mathbb{R})$ and any $\tilde{q} > Q$.

Part (i). By *Lemma 1*, there exists a unique $\tilde{a}$ such that

$$V(e_\infty, S(\tilde{q}, \tilde{a})) = Y(e_\infty, \tilde{q}, \tilde{a}) = Y(e_\infty, Q, A) = V(e_\infty, S(Q, A)).$$

It follows from $V_S > 0$ that $S(\tilde{q}, \tilde{a}) = S(Q, A)$. Note that the stable action $e_\infty = e^*(\tilde{q}, \tilde{a})$ uniquely solves

$$V_e(e, S(\tilde{q}, \tilde{a})) = Y_e(e, \tilde{q}, \tilde{a}) = c_e(e),$$

and the first-best action $e^{FB} = e^*(Q, A)$ uniquely solves

$$V_e(e, S(Q, A)) = Y_e(e, Q, A) = c_e(e).$$

Since $S(\tilde{q}, \tilde{a}) = S(Q, A)$, we have $V_e(e, S(\tilde{q}, \tilde{a})) = V_e(e, S(Q, A))$ for any $e \geq 0$. Combining this result and $c_{ee} > 0$, we have $e_\infty = e^{FB}$. See *Figure B.1* for an illustration.

Part (ii). By *Lemma 1*, there exists a unique $\tilde{a}$ such that

$$V(e_\infty, S(\tilde{q}, \tilde{a}), \tilde{a}) = Y(e_\infty, \tilde{q}, \tilde{a}) = Y(e_\infty, Q, A) = V(e_\infty, S(Q, A), A),$$

and we have $\tilde{a} < A$. Then it follows from $V_S, V_a > 0$ that $S(\tilde{q}, \tilde{a}) > S(Q, A)$. Since $V_{eS} > 0$ and $V_{ea} \leq 0$, we obtain

$$V_e(e, S(\tilde{q}, \tilde{a}), \tilde{a}) > V_e(e, S(Q, A), A), \tag{B.1}$$

for any $e \geq 0$. Since $c_{ee} > 0$, the stable action $e_\infty = e^*(\tilde{q}, \tilde{a})$ uniquely solves

$$V_e(e, S(\tilde{q}, \tilde{a}), \tilde{a}) = Y_e(e, \tilde{q}, \tilde{a}) = c_e(e),$$

and the first-best action $e^{FB} = e^*(Q, A)$ uniquely solves

$$V_e(e, S(Q, A), A) = Y_e(e, Q, A) = c_e(e).$$

As illustrated by *Figure* B.2, we have $e_\infty > e^{FB}$.

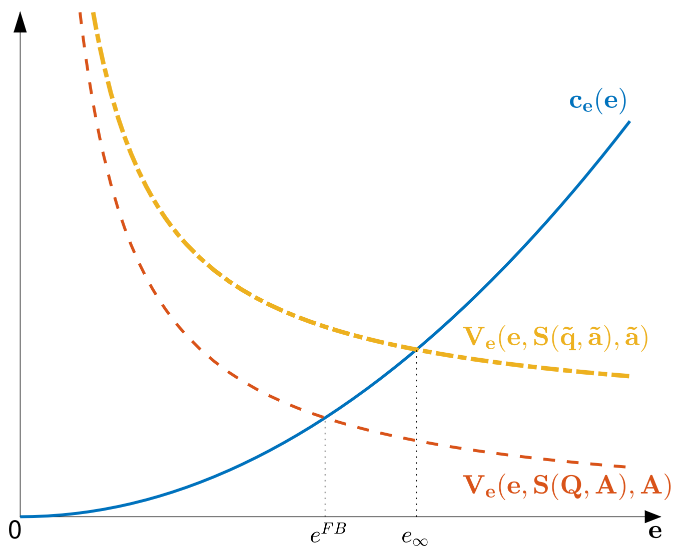


FIGURE B.2. Overselling in Part (i)

Part (iii). By *Lemma* 1, there exists a unique $\tilde{a} < A$ such that

$$V(e_\infty, S(\tilde{q}, \tilde{a}), \tilde{q}) = Y(e_\infty, \tilde{q}, \tilde{a}) = Y(e_\infty, Q, A) = V(e_\infty, S(Q, A), Q).$$

Since $V_S, V_q > 0$, we have $S(\tilde{q}, \tilde{a}) < S(Q, A)$. It follows from $V_{eS} > 0$ and $V_{eq} \leq 0$ that

$$V_e(e, S(\tilde{q}, \tilde{a}), \tilde{q}) < V_e(e, S(Q, A), Q),$$

for any $e \geq 0$. Since $c_{ee} > 0$, the stable action $e_\infty = e^*(\tilde{q}, \tilde{a})$ uniquely solves

$$V_e(e, S(\tilde{q}, \tilde{q}), \tilde{a}) = Y_e(e, \tilde{q}, \tilde{a}) = c_e(e),$$

and the first-best action $e^{FB} = e^*(Q, A)$ uniquely solves

$$V_e(e, S(Q, A), Q) = Y_e(e, Q, A) = c_e(e),$$

we have $e_\infty < e^{FB}$ (see *Figure* B.3 for an illustration).

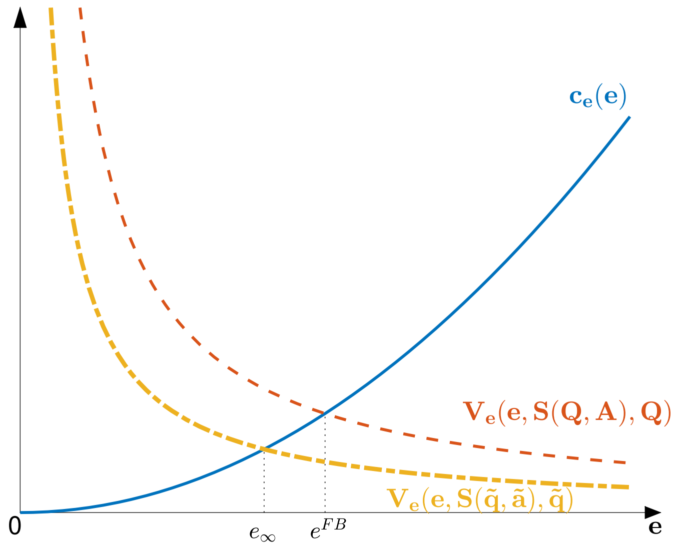


FIGURE B.3. Overselling in Part (ii)

B.4. *Proof of Proposition 2*

Fix any arbitrary $(Q, \tilde{q}, A, \pi_0) \in [\underline{q}, \bar{q}]^2 \times \mathbb{R} \times \Delta(\mathbb{R})$. If the principal is truth-telling (i.e., $\tilde{q} = Q$), then in period 2, the agent concludes correctly about her own ability A by perfectly observing the output $Y(e_1, Q, A)$. Therefore, she exerts effort under the correct incentive, and

$$e_2 = e^*(\tilde{q}, \tilde{a}_1) = e^*(Q, A) = e^{FB}.$$

What remains to prove how e_2 varies if $\tilde{q} \neq Q$.

Part (i). By the definition of $\tilde{a}_1$ (15), for any $\tilde{q}$,

$$V(e_1, S(Q, A)) = V(e_1, S(\tilde{q}, \tilde{a}_1)).$$

It follows from $V_S > 0$ that $S(Q, A) = S(\tilde{q}, \tilde{a}_1)$. Since

(a) the agent's effort at period 2, $e_2 = e^*(\tilde{q}, \tilde{a}_1)$, uniquely solves $V_e(e, S(\tilde{q}, \tilde{a}_1)) = c_e(e)$,
(b) her first-best effort $e^{FB} = e^*(Q, A)$ uniquely solves $V_e(e, S(Q, A)) = c_e(e)$,
(c) $V_e(e, S(\tilde{q}, \tilde{a})) = V_e(e, S(Q, A))$ for any $e \geq 0$,
(d) $c_e(e)$ is strictly increasing,

we can conclude that $e_2 = e^{FB}$.

Part (ii). By (15), for any $\tilde{q}$,

$$V(e_1, S(Q, A), A) = V(e_1, S(\tilde{q}, \tilde{a}_1), \tilde{a}_1),$$

If $\tilde{q} > Q$, then $\tilde{a}_1 < A$. Provided that $V_S, V_a > 0$, we have $S(Q, A) < S(\tilde{q}, \tilde{a}_1)$. Then $V_{eS} > 0$ and $V_{ea} \leq 0$ imply that $V_e(e, S(Q, A), A) < V_e(e, S(\tilde{q}, \tilde{a}_1), \tilde{a}_1)$ for any $e \geq 0$. Then

(a) e_2 uniquely solves $V_e(e, S(\tilde{q}, \tilde{a}_1), \tilde{a}_1) = c_e(e)$,

(b) e^{FB} uniquely solves $V_e(e, S(Q, A), A) = c_e(e)$,

(c) $V_e(e, S(Q, A), A) < V_e(e, S(\tilde{q}, \tilde{a}_1), \tilde{a}_1)$,

(d) $c_e(e)$ is strictly increasing,

imply that $e_2 > e^{FB}$.

If $\tilde{q} < Q$, we have $\tilde{a}_1 > A$ and $e_2 < e^{FB}$; the proof is analogous to the above.

Part (iii). If $\tilde{q} > Q$, then $\tilde{a}_1 > A$. Since $V_S, V_a > 0$, and

$$V(e_1, S(Q, A), Q) = V(e_1, S(\tilde{q}, \tilde{a}_1), \tilde{q}), \tag{B.2}$$

we have $S(Q, A) > S(\tilde{q}, \tilde{a}_1)$. Then it follows from $V_{eS} > 0$ and $V_{eq} \leq 0$ that $V_e(e, S(Q, A), Q) > V_e(e, S(\tilde{q}, \tilde{a}_1), \tilde{q})$ for any $e \geq 0$. We can conclude that $e_2 < e^{FB}$ since

(a) $e_2 = e^*(\tilde{q}, \tilde{a}_1)$ uniquely solves $V_e(e, S(\tilde{q}, \tilde{a}_1), \tilde{q}) = c_e(e)$,

(b) $e^{FB} = e^*(Q, A)$ uniquely solves $V_e(e, S(Q, A), Q) = c_e(e)$,

(c) $V_e(e, S(Q, A), Q) > V_e(e, S(\tilde{q}, \tilde{a}_1), \tilde{q})$ for any $e \geq 0$,

(d) $c_e(e)$ is strictly increasing.

If $\tilde{q} < Q$, $e_2 > e^{FB}$; the proof is analogous to the above.

References

Abramson, Kate. 2014. "Turning up the Lights on Gaslighting." *Philosophical Perspectives*, 28: 1–30.

Aina, Chiara. 2023. "Tailored Stories." Working paper.

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics*, 115(3): 715–753.

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

Alloy, Lauren B, and Lyn Y Abramson. 1979. "Judgment of Contingency in Depressed and Nondepressed Students: Sadder but Wiser?" *Journal of Experimental Psychology: General*, 108(4): 441.

Anand, Vikas, Blake E Ashforth, and Mahendra Joshi. 2004. "Business as Usual: The Acceptance and Perpetuation of Corruption in Organizations." *Academy of Management Perspectives*, 18(2): 39–53.

Andreoni, James, and Tymofiy Mylovanov. 2012. "Diverging Opinions." *American Economic Journal: Microeconomics*, 4(1): 209–232.

Ba, Cuimin. 2023. "Robust Model Misspecification and Paradigm Shifts." *arXiv preprint arXiv:2106.12727*.

Ba, Cuimin, and Alice Gindin. 2023. "A multi-agent model of misspecified learning with overconfidence." *Games and Economic Behavior*, 142: 315–338.

Ba, Cuimin, J Aislinn Bohren, and Alex Imas. 2022. "Over-and Underreaction to Information." *Available at SSRN*.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2022. "Belief-Dependent Motivations and Psychological Game Theory." *Journal of Economic Literature*, 60(3): 833–882.

Bazerman, Max H, and Ann E Tenbrunsel. 2011. *Blind Spots: Why We Fail to Do What's Right and What to Do about It.* Princeton University Press.

Bénabou, Roland. 2013. "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies*, 80(2): 429–462.

Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *The Quarterly Journal of Economics*, 117(3): 871–915.

Bénabou, Roland, and Jean Tirole. 2004. "Willpower and Personal Rules." *Journal of Political Economy*, 112(4): 848–886.

Bénabou, Roland, and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *The Quarterly Journal of Economics*, 126(2): 805–855.

Bénabou, Roland, and Jean Tirole. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs." *Journal of Economic Perspectives*, 30(3): 141–164.

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

Berk, Robert H. 1966. "Limiting Behavior of Posterior Distributions When the Model Is Incorrect." *The Annals of Mathematical Statistics*, 37(1): 51–58.

Bodner, Ronit, and Drazen Prelec. 2003. "Self-Signaling and Diagnostic Utility in Everyday Decision Making." *The Psychology of Economic Decisions*, 1(105): 26.

Bohren, J Aislinn. 2016. "Informational Herding with Model Misspecification." *Journal of Economic Theory*, 163: 222–247.

Bohren, J Aislinn, and Daniel N Hauser. 2021. "Learning with Heterogeneous Misspecified Models: Characterization and Robustness." *Econometrica*, 89(6): 3025–3077.

Bohren, J Aislinn, and Daniel N Hauser. 2023. "The Behavioral Foundations of Model Misspecification: A Decomposition." *Working Paper*.

Bolotnyy, Valentin, Matthew Basilico, and Paul Barreira. 2022. "Graduate Student Mental Health: Lessons from American Economics Departments." *Journal of Economic Literature*, 60(4): 1188–1222.

Bosmajian, Haig A. 1974. *The Language of Oppression.* University Press of America, INC.

Brunnermeier, Markus K, and Jonathan A Parker. 2005. "Optimal Expectations." *American Economic Review*, 95(4): 1092–1118.

Carrillo, Juan D, and Thomas Mariotti. 2000. "Strategic Ignorance as a Self-Disciplining Device." *The Review of Economic Studies*, 67(3): 529–544.

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *The Quarterly Journal of Economics*, 122(3): 1187–1234.

Dillenberger, David, and Philipp Sadowski. 2012. "Ashamed to Be Selfish." *Theoretical Economics*, 7(1): 99–124.

Doney, Patricia M, and Joseph P Cannon. 1997. "An Examination of the Nature of Trust in BuyerSeller Relationships." *Journal of Marketing*, 61(2): 35–51.

Eil, David, and Justin M Rao. 2011. "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics*, 3(2): 114–138.

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

Ely, Jeffrey C, and Martin Szydlowski. 2020. "Moving the Goalposts." *Journal of Political Economy*, 128(2): 468–506.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica*, 84(3): 1093–1130.

Esponda, Ignacio, and Demian Pouzo. 2021. "Equilibrium in Misspecified Markov Decision Processes." *Theoretical Economics*, 16(2): 717–757.

Esponda, Ignacio, Demian Pouzo, and Yuichi Yamamoto. 2021. "Asymptotic Behavior of Bayesian Learners with Misspecified Models." *Journal of Economic Theory*, 195: 105260.

Eyster, Erik, Shengwu Li, and Sarah Ridout. 2021. "A Theory of Ex Post Rationalization." *arXiv preprint arXiv:2107.07491*.

Fan, Qiaofeng, and Lukas Bolte. 2022. "Motivated Mislearning: The Case of Correlation Neglect." *Available at SSRN 4153191*.

Frick, Mira, Ryota Iijima, and Yuhta Ishii. 2020. "Misinterpreting Others and the Fragility of Social Learning." *Econometrica*, 88(6): 2281–2328.

Frick, Mira, Ryota Iijima, and Yuhta Ishii. 2023. "Belief Convergence under Misspecified Learning: A Martingale Approach." *The Review of Economic Studies*, 90(2): 781–814.

Fudenberg, Drew, and David K Levine. 1993. "Self-Confirming Equilibrium." *Econometrica*, 523–545.

Fudenberg, Drew, and Giacomo Lanzani. 2023. "Which Misspecifications Persist?" *Theoretical Economics*, 18(3): 1271–1315.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack. 2021. "Limit Points of Endogenous Misspecified Learning." *Econometrica*, 89(3): 1065–1098.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack. 2023. "Pathwise Concentration Bounds for Bayesian Beliefs." *Theoretical Economics (forthcoming)*.

Fudenberg, Drew, Gleb Romanyuk, and Philipp Strack. 2017. "Active Learning with a

Misspecified Prior." *Theoretical Economics*, 12(3): 1155–1189.

Gagnon-Bartsch, Tristan, and Benjamin Bushong. 2022. "Learning with Misattribution of Reference Dependence." *Journal of Economic Theory*, 203: 105473.

Gervais, Simon, and Terrance Odean. 2001. "Learning to Be Overconfident." *The Review of Financial Studies*, 14(1): 1–27.

Gin, Logan E, Nicholas J Wiesenthal, Isabella Ferreira, and Katelyn M Cooper. 2021. "PhDepression: Examining How Graduate Research and Teaching Affect Depression in Life Sciences PhD Students." *CBE—Life Sciences Education*, 20(3): ar41.

Gottlieb, Daniel. 2014. "Imperfect Memory and Choice under Risk." *Games and Economic Behavior*, 85: 127–158.

Gur, Ruben C, and Harold A Sackeim. 1979. "Self-Seception: A Concept in Search of a Phenomenon." *Journal of Personality and Social Psychology*, 37(2): 147.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack. 2018. "Unrealistic Expectations and Misguided Learning." *Econometrica*, 86(4): 1159–1214.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack. 2021. "Convergence in Models of Misspecified Learning." *Theoretical Economics*, 16(1): 73–99.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack. 2023. "Misinterpreting Yourself." *Available at SSRN 4325160*.

He, Kevin. 2022. "Mislearning from Censored Data: The Gambler's Fallacy and Other Correlational Mistakes in Optimal-Stopping Problems." *Theoretical Economics*, 17(3): 1269–1312.

He, Kevin, and Jonathan Libgober. 2020. "Evolutionarily Stable (Mis)specifications: Theory and Applications." *arXiv preprint arXiv:2012.15007*.

Hiroto, Donald S, and Martin E Seligman. 1975. "Generality of Learned Helplessness in Man." *Journal of Personality and Social Psychology*, 31(2): 311.

Holmström, Bengt. 1979. "Moral Hazard and Observability." *The Bell Journal of Economics*, 10(1): 74–91.

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

Klemperer, Victor. 2002. *The Language of the Third Reich: A Philologist's Notebook, trans.* Martin Brady, London: Continuum.

Korn, Christoph W, Tali Sharot, Hendrik Walter, Hauke R Heekeren, and Raymond J Dolan. 2014. "Depression is Related to an Absence of Optimistically Biased Belief Updating about Future Life Events." *Psychological Medicine*, 44(3): 579–592.

Kőszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association*, 4(4): 673–707.

Kőszegi, Botond, George Loewenstein, and Takeshi Murooka. 2022. "Fragile Self-Esteem." *The Review of Economic Studies*, 89(4): 2026–2060.

Levy, Gilat, Ronny Razin, and Alwyn Young. 2022. "Misspecified Politics and the Recurrence of Populism." *American Economic Review*, 112(3): 928–962.

Levy, Raphaël. 2014. "Soothing Politics." *Journal of Public Economics*, 120: 126–133.

Maier, Steven F, and Martin E Seligman. 1976. "Learned Helplessness: Theory and Evidence." *Journal of Experimental Psychology: General*, 105(1): 3.

McKelvey, Richard D, and Thomas R Palfrey. 1992. "An Experimental Study of the Centipede Game." *Econometrica*, 803–836.

Montiel Olea, José Luis, Pietro Ortoleva, Mallesh Pai, and Andrea Prat. 2022. "Competing Models." *The Quarterly Journal of Economics*, 137(4): 2419–2457.

Norem, Julie K, and Nancy Cantor. 1986. "Defensive Pessimism: Harnessing Anxiety as Motivation." *Journal of Personality and Social Psychology*, 51(6): 1208.

Nyarko, Yaw. 1991. "Learning in Mis-specified Models and the Possibility of Cycles." *Journal of Economic Theory*, 55(2): 416–427.

Pintrich, Paul R. 2004. "A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students." *Educational Psychology Review*, 16: 385–407.

Quattrone, George A, and Amos Tversky. 1984. "Causal versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion." *Journal of Personality and Social Psychology*, 46(2): 237.

Schacter, Daniel L. 2012. "Adaptive Constructive Processes and the Future of Memory." *American Psychologist*, 67(8): 603.

Schwardmann, Peter, and Joel Van der Weele. 2019. "Deception and Self-Deception." *Nature Human Behaviour*, 3(10): 1055–1061.

Sharot, Tali, Christoph W Korn, and Raymond J Dolan. 2011. "How Unrealistic Optimism is Maintained in the Face of Reality." *Nature Neuroscience*, 14(11): 1475–1479.

Stern, Robin. 2018. *The Gaslight Effect: How to Spot and Survive the Hidden Manipulation Others Use to Control Your Life.* Harmony.

Sweet, Paige L. 2019. "The Sociology of Gaslighting." *American Sociological Review*, 84(5): 851–875.

Thorp, H Holden. 2020. "Underpromise, Overdeliver." *Science*, 367(6485): 1405–1405.

Von Hippel, William, and Robert Trivers. 2011. "The Evolution and Psychology of Self-Deception." *Behavioral and Brain Sciences*, 34(1): 1–16.