Equilibrium Contracts and Boundedly Rational Expectations*

Heiner Schumacher† Heidi Christina Thysen‡

KU Leuven London School of Economics

Version: September 30, 2020

Abstract

We study a principal-agent framework in which the agent forms beliefs based on a misspecified subjective model of the principal's project. She fits this model to the objective probability distribution to predict output under alternative actions. Misspecifications in the subjective model may lead to biased beliefs. However, under mild restrictions, the agent has correct beliefs on the equilibrium path so that the optimal contract is non-exploitative. This allows for a behavioral version of the informativeness principle: The optimal contract conditions on an additional variable only if it is informative about the action according to the agent's subjective model. We further characterize when misspecifications affect the optimal contract. One implication of this characterization is that the scope for belief biases depends on the agent's job, e.g., her position in the hierarchy.

Keywords: Bayesian Networks, Principal-Agent Relationship, Bounded Rationality
JEL Classification: D03, D82, D86

---

*We gratefully acknowledge financial support by the ERC Advanced Investigator grant no. 692995. We thank Yair Antler, Felix Bierbrauer, Markus Dertwinkel-Kalt, Kfir Eliaz, Florian Englmaier, Guido Friebel, Paul Heidhues, Johannes Johnen, Heiko Karle, Michael Kosfeld, Botond Kőszegi, Gilat Levy, Debraj Ray, Ronny Razin, Karl Schlag, Klaus Schmidt, Dirk Sliwka, Balázs Szentes, and Yves Le Yaouanq, as well as seminar audiences at Aarhus University, Boston University, University of Cambridge, University of Cologne, Dalhousie University, Humboldt-University Berlin, University of Konstanz, London School of Economics, Ludwig-Maximilians University Munich, Tilburg University, the EEA-ESEM 2017 in Lisbon, the ESSET 2018 in Gerzensee, and the Psychology and Economics of Causal Reasoning Workshop 2019 in London for their valuable comments and suggestions. We especially thank Ran Spiegler for his invaluable support of the project. The usual disclaimer applies.

†KU Leuven, Department of Economics, E-mail: heiner.schumacher@kuleuven.be.

‡London School of Economics, Department of Economics, E-mail: h.c.thysen@lse.ac.uk.

1 Introduction

The canonical principal-agent model of contracting under asymmetric information assumes that the agent knows the probabilistic consequences of all available actions. Formally, these are defined by a production function $p(y \mid a)$, where y is the contractible output and a the agent's action. Given the incentives provided by the contract, the agent chooses an action that – according to this function – maximizes her expected payoff. However, in an organization, $p(y \mid a)$ is typically a complex object. It may reflect information that is unavailable to the agent or that the agent cannot process due to cognitive limitations. Herbert Simon therefore proposed that administrative behavior must be "boundedly rational" (Simon 1947, 1955).

One approach to analyze contracting with boundedly rational agents is to assume directly that beliefs $\hat{p}(y \mid a)$ about the production function are biased so that $\hat{p}(y \mid a) \neq p(y \mid a)$. An important implication of this approach is that the optimal contract may exploit the agent, in the sense that her (true) expected payoff falls below her reservation utility (e.g., Kőszegi 2014). Directly assuming biased beliefs has two disadvantages though. First, from a dynamic perspective, it is unclear how sustainable a certain belief $\hat{p}(y \mid a)$ – and hence exploitation – would be when the agent gathers data on the production function. Second, it treats beliefs as an exogenous variable. If we study how the optimal contract varies in the production function or the informational environment, the results crucially depend on how we choose beliefs.

In this paper, we examine a contracting framework in which a boundedly rational agent has beliefs about the production function that are endogenously derived from her environment. The agent estimates $p(y \mid a)$ based on data generated by the true production process, the implemented strategy α^*, and a non-parametric subjective model $\mathcal{R}$. A model $\mathcal{R}$ is a collection of variables and causal relationships between these variables. It captures what the agent knows about the production process. This model may be misspecified. For example, it may be "too simple" relative to the complexity of the organization: empirical regularities that matter for the principal's project may not appear in $\mathcal{R}$. The agent's subjective beliefs about $p(y \mid a)$ will be denoted by $p_{\mathcal{R}}(y \mid a; \alpha^*)$. An equilibrium contract implements a strategy α^* if it is optimal for the agent to follow α^* under this contract given her beliefs $p_{\mathcal{R}}(y \mid a; \alpha^*)$. We study the properties of the optimal equilibrium contract, and obtain several new results on optimal contracting and organization that we would not get (or get only under very specific assumptions) if we directly choose beliefs $\hat{p}(y \mid a)$.

To capture the agent's limited understanding of her environment, we apply Spiegler's (2016) Bayesian network approach. As an illustration, consider the following example:

> *"Marketer Example."* The agent is a marketer whose job is to increase sales y. One strategy to increase sales is to make cold-calls $a \in \{0, 1\}$, that is, calling po-

> tential customers without prior consent. Making cold-calls improves consumers' information $x_1 \in \{0, 1\}$ about the firm's product, but also reduces the firm's reputation $x_2 \in \{0, 1\}$ since some customers are annoyed by being cold-called. Expected sales increase both in consumer information x_1 and reputation x_2. However, when choosing her action, the marketer does not take the firm's reputation into account. The only mechanism on her mind is that making cold-calls improves consumer information, and that more information translates into more sales.

The Bayesian network approach roughly works as follows in the marketer example.[1] The setting describes an "extended production function" $p(x_1, x_2, y \mid a)$, i.e., a joint probability distribution over the realization of consumer information, reputation and sales for any given action. This function captures the objective model $\mathcal{R}^*$ of the project: $\mathcal{R}^*$ contains all relevant variables, {action, consumer information, reputation, sales}, and the causal relationships between these variables. The agent's subjective model $\mathcal{R}$ is a simplified version of $\mathcal{R}^*$ as it only contains the variables {action, consumer information, sales}, and their causal relationships. Her beliefs are derived by fitting $\mathcal{R}$ to the objective probability distribution, which is generated by the implemented strategy α^* and the extended production function $p(x_1, x_2, y \mid a)$. Thus, the different elements in the agent's subjective model $\mathcal{R}$ are quantified using input from the true data-generating process. Combining these elements yields the agent's subjective beliefs $p_{\mathcal{R}}(y \mid a; \alpha^*)$, which in general are not invariant to changes in α^*.

If $\mathcal{R}$ differs from $\mathcal{R}^*$, the agent's beliefs about $p(y \mid a)$ may be biased, and both the incentive compatibility and the participation constraint could in principle be affected by this bias. Our first important observation is that a weak restriction on the agent's subjective model guarantees that the participation constraint is not affected. This restriction is that $\mathcal{R}$ is "perfect", which means that the agent takes into account the link between any two variables in $\mathcal{R}$ that have a joint influence on a third variable in $\mathcal{R}$ (Spiegler 2017). She then correctly predicts the marginal equilibrium distribution over output, so that the optimal equilibrium contract does not exploit the agent. Importantly, a perfect $\mathcal{R}$ ensures in many cases that there are no informational cues in the data the agent gathers on the equilibrium path that could alert her about the misspecification in $\mathcal{R}$. Therefore, the agent's possibly incorrect beliefs and the corresponding equilibrium contract can be sustainable.

A perfect $\mathcal{R}$ does however not ensure that the incentive compatibility constraints are unaffected by the model misspecification. In the marketer example, if the principal implements making cold-calls, then, by not taking reputation into account, the agent overestimates the drop in sales after deviation to not making cold-calls, i.e., she is "control optimistic." This relaxes

[1] Missing technical details will be explained thoroughly in the next section.

the incentive compatibility constraint, so that the principal can implement cold-calls with fewer incentives than if the agent had rational expectations. Thus, the principal can strictly benefit from the misspecification in the agent's model even when exploitation is infeasible.

The property of correct expectations on the equilibrium path has further implications for the optimal equilibrium contract. An important question in contract theory is on which variables the optimal contract should condition the agent's wage. According to the informativeness principle (e.g., Holmström 1979, Chaigneau et al. 2019), the optimal contract conditions on an additional signal z only if z provides information about the agent's action that is not contained in y. Our second important observation is that we can derive an analogous statement when the agent has correct expectations on the equilibrium path about the joint distribution of y and z (with a further qualification this holds if $\mathcal{R}$ is perfect). In this case, the optimal equilibrium contract conditions on z only if the agent's action a and z are not independent conditional on y according to the agent's subjective beliefs. This result does not depend on other properties of the agent's subjective model $\mathcal{R}$, and hence would hold in any setting where the agent's beliefs about the joint distribution of y and z are correct. However, we can use results from the Bayesian network literature to state sufficient conditions on $\mathcal{R}$ so that the result's requirements are satisfied; these results also provide a tool to visually inspect for a given subjective model whether the optimal contract conditions on z.

To illustrate our behavioral version of the informativeness principle, we consider a classic application. Suppose the principal can condition the agent's wage both on her output y and on her relative peer performance z. A common shock influences both y and z so that under rational expectations the optimal contract conditions on both variables to filter out windfall gains and losses. If the agent's subjective model does not include the common shock, then z is for the agent only a noisy signal of y. The optimal equilibrium contract then only conditions on y and therefore remains incomplete. Thus, using our results we can give sufficient conditions on the agent's causal model under which the inclusion of peer-performance in the contract is inefficient. This provides a new explanation for why most executive compensation contracts do not condition on peer-performance (Bebchuk and Fried 2004).

Misspecifications in $\mathcal{R}$ do not always affect the optimal equilibrium contract. We call the agent "behaviorally rational" if she correctly anticipates the production function, or, formally, $p_{\mathcal{R}}(y \mid a; \alpha) = p(y \mid a)$ for all possible a and α (regardless of the parametrization of the extended production function). Our third important observation is that we can find a correspondence $H^*(\mathcal{R}^*)$ which indicates for a given objective model $\mathcal{R}^*$ the set of variables the agent must take into account in her simplified subjective model $\mathcal{R}$ so that she is behaviorally rational. We show that $H^*(\mathcal{R}^*)$ is often a strict subset of the variables in $\mathcal{R}^*$, and that the difference between a variable $i \in H^*(\mathcal{R}^*)$ and a variable $j \notin H^*(\mathcal{R}^*)$ can be quite nuanced. Here is a simple example: Consider a version of the marketer example where the agent's action does not influence reputation, but where consumer information affects reputation. The objective model $\mathcal{R}^*$ then has no link between action and reputation, but a link between consumer information and reputation. An agent with the subjective model from the marketer example is now behaviorally rational. She correctly anticipates the production function even though she ignores the influence of reputation on output.

The characterization of $H^*(\mathcal{R}^*)$ shows which variables matter for the agent's beliefs. An important interpretation of the objective model $\mathcal{R}^*$ is that it captures the agent's job, i.e., through which tasks, interactions, and decision-making powers she influences the final output.[2] In the canonical contracting model, these aspects are immaterial since behavior is governed by the production function $p(y \mid a)$. Similarly, they do not matter when beliefs $\hat{p}(y \mid a)$ are fixed exogenously. In our framework, we can have two extended production functions that give rise to the same "reduced-form" production function $p(y \mid a)$, but that differ in their causal model $\mathcal{R}^*$, and hence in the extent to which simplifications affect $p_{\mathcal{R}}(y \mid a; \alpha^*)$.

One application of this finding is that we can examine which organizational features potentially cause the agent to overestimate the productivity of her effort. As the marketer example shows, this happens if the agent does not take into account a partial negative effect of her effort on the output. There are several intuitive reasons why this may be the case. Consider an agent in a management position in which her effort influences the behavior of other workers (e.g., a group of telemarketers). If the agent does not understand the difficulties of their job (e.g., that telemarketing has a partial negative effect on sales through reputation), she overestimates her subordinates' – and hence her own – productivity. There are different instances where this could happen: The agent may be a technical expert who is promoted into a management position in which she oversees the actions of workers whose job she does not fully understand. Alternatively, it may be the case that subordinates do not communicate the problems they face to their managers (due to career concerns). These phenomena are usually discussed critically in the management literature, but in our framework they advance the agent's effort motivation and hence benefit the principal.

Related Literature. Our basic model is the principal-agent framework introduced by Holmström (1979) and Grossman and Hart (1983). Holmström (1979) states a version of the informativeness principle. A generalization of it can be found in, e.g., Chaigneau et al. (2019). In the canonical framework, both principal and agent know the production function $p(y \mid a)$.

There are different approaches in behavioral contract theory that relax the assumption of

[2]For example, one can interpret $\mathcal{R}^*$ as an adjusted depiction of the organizational chart of the principal's project. As the CEO the agent would influence his senior managers who in turn influence their subordinates' behavior and so forth. A simplification in $\mathcal{R}$ then captures that the agent ignores a certain part of the organization.

unbiased beliefs about $p(y \mid a)$. First, several contracting models directly assume that the agent's beliefs about the production function are biased, i.e., $\hat{p}(y \mid a) \neq p(y \mid a)$; see Fang and Moscarini (2005), Van den Steen (2005), Gervais and Goldstein (2007), Santos-Pinto (2008), De la Rosa (2011), Sautmann (2007, 2013), Spinnewijn (2013, 2015). Specifically, this approach is used to model an overconfident agent who overestimates the probability of good states and underestimates the probability of bad states. This typically allows the principal to exploit the agent by paying more after high output and much less after low output, in which case the agent's expected payoff is below her reservation utility.

Second, a rich literature builds state-space models of "unawareness" (e.g., Dekel et al. 1998, Heifetz et al. 2006, 2013) and applies them to contracting settings. Auster (2013) examines a principal-agent model with an agent who is unaware of some output levels y, which again implies that the contract is exploitative. Von Thadden and Zhao (2012, 2014) assume that the agent is unaware of her available actions a and chooses a default action unless the principal educates her; unawareness then relaxes incentive compatibility at the default action.

Third, in order to justify biased beliefs, several papers assume that the agent knows the link between action and outcomes $p(y \mid a)$, but derives anticipatory utility from optimistic beliefs. She therefore chooses beliefs $\hat{p}(y \mid a)$ that solve the trade-off between the losses from biased decision-making and the gains from anticipation; see Bénabou and Tirole (2002), Brunnermeier and Parker (2005), and Kőszegi (2006). For an organizational context, Bénabou (2013) shows how the interaction between group members can make the suppression of bad news a strategic complement, so that collective denial of adverse signals ("groupthink") occurs in equilibrium. Immordino et al. (2015) show that if the anticipatory utility is not too important, the principal may provide incentives so that it is optimal for the agent to choose correct beliefs.

Our approach to boundedly rational expectations and contracting is more conservative. The agent derives her beliefs from the true data-generating process, as in the canonical model; she just may not take into account all empirical regularities that matter for the principal's project. The misspecification in the agent's subjective model may cause her to overestimate her productivity. However, under a weak restriction, she still correctly anticipates the equilibrium distribution over output. The optimal equilibrium contract then does not exploit the agent, and we can derive a behavioral version of the informativeness principle. Moreover, our framework allows for misspecifications that do not affect the agent's beliefs about the production function. We use this structure to study which aspects of the agent's job affect the scope for control optimism.

We also contribute to the literature on Bayesian networks/directed acyclic graphs (DAGs), which have been used extensively in the artificial intelligence literature. Pearl (2009) promotes the view that DAGs represent causal relationships and provides a broad introduction to DAGs.

In economics, Spiegler (2016, 2017) uses Bayesian networks to model agents with boundedly rational expectations. DAGs provide a general method to capture a variety of different inference errors such as reverse causation and coarseness. We build on these insights and apply them to contracting. Other recent papers use causal models to capture boundedly rational decision makers in monetary policy (Spiegler 2019), political competition (Eliaz and Spiegler 2018), Bayesian persuasion (Eliaz et al. 2019), and decision theory (Schenone 2019).

The remainder of the paper is organized as follows. Section 2 describes our framework. In Section 3, we examine how a misspecification in the agent's subjective model affects the contracting problem. In Section 4, we state a behavioral version of the informativeness principle. In Section 5, we characterize when a misspecification leads to biased beliefs about the production function, and illustrate the implications of this characterization. In Section 6, we revisit two classic comparative statics of the canonical contracting framework. Section 7 concludes. Omitted proofs and further results can be found in the Online Appendix.

2 The Model

We consider a standard principal-agent problem and combine it with the Bayesian network model of boundedly rational beliefs, as introduced in Spiegler (2016).

Basic Framework. Let $A \subset \mathbb{R}$ be a finite set of actions, $Y \subset \mathbb{R}$ a finite set of outputs, and $W \subseteq \mathbb{R}^{|Y|}$ the set of possible incentive schemes. The principal proposes a contract $(w(y), p(a))$, where $w(y) \in W$ is the agent's wage conditional on the output $y \in Y$ and $p(a) \in \Delta(A)$ is the probability with which the principal wishes the agent to choose action $a \in A$. The agent can reject or accept the contract. If she rejects it, she enjoys the outside option value $\bar{U}$, while the principal earns zero. If she accepts the contract, she chooses an action $a \in A$. The agent's personal cost of choosing a is given by a function $c(a)$. The action stochastically influences the project's output. The agent's utility from wage w is given by the utility function $u(w)$, which is weakly concave and exhibits $\lim_{w \to -\infty} u(w) = -\infty$. When the output is y and the agent's action is a, the principal's payoff is $V = y - w(y)$ and the agent's payoff is $U = u(w(y)) - c(a)$.

Causal Structure. We model the causal structure through which the agent's action affects the output y. Let $N^* = \{0, ..., n\}$ be the set of relevant variables (or nodes). This set contains the agent's action and output, but may also include other variables. A generic realization of variable i is given by $x_i \in X_i$, where X_i is a finite set that contains at least two elements. Node 0 is the agent's action ($x_0 = a$, $X_0 = A$) and node n is the output ($x_n = y$, $X_n = Y$). The state is a vector $x^* = (x_0, x_1, ..., x_n)$ and the set of all states is $X^* = \times_{i \in N^*} X_i$. For every subset $M \subseteq N^*$ and $x^* \in X^*$, let $x_M = (x_k)_{k \in M}$.

Denote by $p(x_1, \ldots, x_n \mid a)$ the extended production function. For any action $a \in A$, it has full support over $X_1 \times ... \times X_n$. We represent its causal structure by an irreflexive, asymmetric and acyclic binary relation R^* over N^*, and denote it by the DAG $\mathcal{R}^* = (N^*, R^*)$, see the graph on the left of Figure 1 for an example. For two nodes $i, j \in N^*$ one may read iR^*j as "node i impacts on node j." The set of nodes that influence i is defined, with abuse of notation, as $R^*(i) = \{j \in N^* \mid jR^*i\}$. Nothing influences the agent's action, $R^*(0) = \emptyset$. The probability distribution over states, $p(x^*) \in \Delta(X^*)$, then naturally factorizes according to $\mathcal{R}^*$ via the formula

$$p(x^*) = \prod_{i \in N^*} p(x_i \mid x_{R^*(i)}). \tag{1}$$

We assume that the "objective model" $\mathcal{R}^*$ is one of the sparsest DAGs so that $p(x^*)$ factorizes according to $\mathcal{R}^*$. That is, $\mathcal{R}^*$ contains exactly those conditional independence assumptions that are satisfied by $p(x^*)$.[3]

Figure 1: An objective model $\mathcal{R}^*$ (left) and the agent's subjective model $\mathcal{R}$ (right).

Beliefs, Personal Equilibrium, and Equilibrium Contract. The agent has her own subjective model $\mathcal{R} = (N, R)$, see Figure 1 for an example. We assume that $\{0, n\} \in N \subseteq N^*$ and $R(0) = \emptyset$. The assumption that the agent includes her own action and the output in her subjective model ensures that her utility is measurable with respect to her beliefs. $N \subseteq N^*$ is assumed purely for simplicity. $R(0) = \emptyset$ implies that the agent knows that she does not receive any information about other variables prior to choosing an action, and that she has correct beliefs about the marginal distribution over her own action.

Definition 1. *We say that $\mathcal{R}$ is misspecified if $\mathcal{R} \neq \mathcal{R}^*$, and that $\mathcal{R}$ is a simplification if $N \subset N^*$ and $R = N \times N \cap R^*$.*

A simplification is a misspecification where the agent's subjective model $\mathcal{R}$ emerges from $\mathcal{R}^*$ by dropping nodes from $\mathcal{R}^*$ and the links adjacent to them. It will receive considerable attention in this paper. Denote by $x = (x_i)_{i \in N}$ the state vector for the agent's subjective model

[3] This assumption is for convenience only and will be relaxed in Section 5.

and $X = \times_{i \in N} X_i$. The agent fits her subjective model $\mathcal{R}$ to the data generated by $p(x^*)$, so her beliefs factorize according to the formula

$$p_{\mathcal{R}}(x) = \prod_{i \in N} p(x_i \mid x_{R(i)}). \tag{2}$$

Thus, all the conditional independence assumptions embedded in $\mathcal{R}$ also appear in the agent's beliefs. For example, when the agent's subjective model is $\mathcal{R}$ from Figure 1, her beliefs factorize according to $p_{\mathcal{R}}(a, x_1, y) = p(a)p(x_1 \mid a)p(y \mid x_1)$, where $p(a)$, $p(x_1 \mid a)$ and $p(y \mid x_1)$ follow from the probability distribution $p(x^*)$. Given the objective model in Figure 1, $p(y \mid x_1)$ will depend on $p(a)$ through variables 2 and 3. Hence, the agent's beliefs in general depend on $p(a)$, even when conditioning on her action. We therefore augment notation to indicate which strategy $p(a)$ is used when deriving beliefs and write $p_{\mathcal{R}}(x; p(a))$ instead of $p_{\mathcal{R}}(x)$. For any subset $M \subset N$, the agent's belief about the marginal distribution over x_M is calculated as $p_{\mathcal{R}}(x_M; p(a)) = \sum_{x_{N \setminus M} \in X_{N \setminus M}} p_{\mathcal{R}}(x_M, x_{N \setminus M}; p(a))$.

The agent follows the prescribed strategy from the contract only if it maximizes her expected utility given the wage scheme $w(y)$ and her subjective beliefs about the output conditional on her action, which we denote by $p_{\mathcal{R}}(y \mid a; p(a))$. These are computed as

$$p_{\mathcal{R}}(y \mid a; p(a)) = \frac{p_{\mathcal{R}}(a, y; p(a))}{\sum_{y \in Y} p_{\mathcal{R}}(y, a; p(a))}. \tag{3}$$

To close the model, we need to specify the agent's strategy $p(a)$ that is used to derive these beliefs. We therefore formalize the agent's strategy as a personal equilibrium.

Definition 2. *The strategy $p(a)$ is a personal equilibrium at $\mathcal{R}$ and $w(y)$ if for all actions $a \in A$ in the support of $p(a)$ we have*

$$a \in \arg\max_{a'} \sum_{y \in Y} p_{\mathcal{R}}(y \mid a'; p(a)) u(w(y)) - c(a'),$$

where $p_{\mathcal{R}}(y \mid a'; p(a)) = \lim_{k \to \infty} p_{\mathcal{R}}(y \mid a'; p^k(a))$ for all actions $a' \in A$ and a sequence $p^k(a) \to p(a)$ of fully mixed strategy profiles.

With the full support assumption, a fully mixed action profile ensures that all conditional probabilities are well-defined. The definition requires that equilibrium beliefs are the limit of a sequence of fully mixed profiles. A personal equilibrium always exists in our framework; see Online Appendix A.1. We call a contract $(w(y), p(a))$ an "equilibrium contract" if $p(a)$ is a personal equilibrium at $\mathcal{R}$ and $w(y)$. An optimal equilibrium contract is an equilibrium contract that maximizes the principal's expected payoff. For convenience, we denote beliefs by $p_{\mathcal{R}}(y \mid a; a^*)$ when a pure action a^* is implemented, and $p_{\mathcal{R}}(y \mid a; \alpha)$ with $p(a = 1) = \alpha$ when we have a binary action set $A = \{0, 1\}$.

The proposed solution concept is static. The agent's beliefs are derived from a probability distribution that could be influenced by the strategy that the equilibrium contract implements. One interpretation is that the agent is experienced and thus has data on how her action impacts on the variables in her subjective model. An alternative interpretation is that there are (or have been) many other agents in the organization who exchange data with their new colleague to which she can fit her subjective model.

3 The Optimal Equilibrium Contract

In this section, we study the properties of the optimal equilibrium contract for a given extended production function $p(x_1, ..., x_n \mid a)$ and subjective model $\mathcal{R}$. If $(w^*(y), p^*(a))$ is an optimal equilibrium contract, then $w^*(y), p^*(a)$ solve the maximization problem

$$\max_{w(y) \in W, p(a) \in \Delta(A)} \sum_{a \in A} \sum_{y \in Y} p(a) p(y \mid a)(y - w(y)) \tag{4}$$

subject to the constraints

$$p(a) \in \Delta(A) \;\; \textit{is a personal equilibrium at } \mathcal{R} \textit{ and } w(y), \tag{IC}$$

$$\sum_{a' \in A} \sum_{y \in Y} p(a')[p_{\mathcal{R}}(y \mid a'; p(a)) u(w(y)) - c(a')] \geq \bar{U}. \tag{PC}$$

When the agent's subjective model $\mathcal{R}$ equals the objective model $\mathcal{R}^*$, the problem collapses to the canonical principal-agent problem, and can be solved as suggested by Grossman and Hart (1983). We first find for each pure action $a \in A$ the wage scheme $w(y)$ that implements this action at lowest possible cost. Then we choose the action-incentive scheme combination that maximizes the principal's profit. If the agent's subjective model $\mathcal{R}$ differs from the objective model $\mathcal{R}^*$, we find the optimal equilibrium contract by applying the same procedure. However, since the agent's beliefs $p_{\mathcal{R}}(y \mid a; p(a))$ may depend on the implemented strategy $p(a)$, the first step has to be done for all pure and mixed actions $p(a) \in \Delta(A)$.

Suppose the agent is risk-averse with unlimited liability, and the principal implements a (possibly mixed) strategy $p(a)$. The Kuhn-Tucker conditions for the principal's problem are then necessary and sufficient for an optimum. Choose any action a in the support of $p(a)$. The optimal incentive scheme is then characterized by the first-order condition

$$\frac{1}{u'(w(y))} = \frac{p_{\mathcal{R}}(y; p(a))}{p(y)} \left[\mu + \sum_{a' \in A} \lambda_{a'} \frac{p_{\mathcal{R}}(y \mid a; p(a)) - p_{\mathcal{R}}(y \mid a'; p(a))}{p_{\mathcal{R}}(y; p(a))} \right] \tag{5}$$

for all $y \in Y$, where μ and $\lambda_{a'}$ are the usual Lagrange multipliers for the participation and incentive compatibility constraint, respectively. Equation (5) allows us to disentangle how a misspecification in $\mathcal{R}$ may change the contracting problem. First, the *PC* is affected when the agent holds biased beliefs about the equilibrium distribution over output; see the first term on the right of equation (5). In Subsection 3.1, we state a sufficient condition on $\mathcal{R}$ so that this belief is unbiased. Second, the *IC* may be affected. Suppose the principal implements a pure action a and $p_{\mathcal{R}}(y; a) = p(y)$. The ratio in the squared brackets then becomes $1 - \frac{p_{\mathcal{R}}(y|a';a)}{p_{\mathcal{R}}(y|a;a)}$, in which case the optimal incentive scheme depends on a likelihood ratio as in the canonical framework. Any difference between the contracts under the objective and subjective model is then driven by differences between the corresponding likelihood ratios. In Subsection 3.2, we examine how these differences may affect the optimal equilibrium contract.

3.1 Correct Expectations on the Equilibrium Path

We use a Bayesian network result from Spiegler (2017) that characterizes under what circumstances the agent's beliefs about the equilibrium output distribution are correct, so that $p_{\mathcal{R}}(y; p(a)) = p(y)$ for all $p(a) \in \Delta(A)$. To this end, we introduce a few definitions. A v-collider is a triple of nodes (i, j, k) such that iRj, kRj and there is no link between i and k (neither iRk nor kRi is in R). The set of v-colliders of a DAG is called its v-structure. A DAG is called perfect if it has an empty v-structure. A subset of nodes $M \subset N$ is a clique in $\mathcal{R} = (N, R)$ if iRj or jRi for any two nodes $i, j \in M$. For example, in the DAG $\mathcal{R}^*$ from Figure 1, the set $M = \{1, 3, 4\}$ is a clique, while the set $M' = \{2, 3, 4\}$ is not. Each node is a clique in itself, so the output node n is a clique. The following result essentially restates Proposition 2 from Spiegler (2017).

Proposition 1 (Equilibrium Beliefs)**.** *If the agent's model $\mathcal{R} = (R, N)$ is perfect, her equilibrium beliefs satisfy $p_{\mathcal{R}}(x_M; p(a)) = p(x_M)$ for all $p(a) \in \Delta(A)$ and any clique $M \subset N$.*

If the agent's subjective model $\mathcal{R}$ is perfect, then, in a personal equilibrium, the agent correctly anticipates the marginal distribution over each variable in her model, and also the joint distribution over variables in cliques. The intuition behind this result is that perfectness excludes biased estimates due to neglect of correlation. Imagine two variables i, j that influence a third variable k. Suppose that i and j are correlated, and that the agent treats them as uncorrelated. Through the application of the factorization formula (2), the agent may then obtain a biased estimate of the marginal distribution $p(x_k)$. Perfectness implies that the agent always checks for correlations between two variables i, j when, according to her subjective model, they influence a third variable k. We obtain two useful corollaries from Proposition 1.

Corollary 1. *If the agent's model $\mathcal{R} = (R, N)$ is perfect and her equilibrium action is a pure action a^*, her equilibrium beliefs satisfy $p_{\mathcal{R}}(x_M \mid a^*; a^*) = p(x_M \mid a^*)$ for every clique $M \subset N$.*

If the equilibrium contract implements a pure strategy a^*, the agent's belief about the joint distribution of any clique M conditional on her equilibrium strategy is correct. Corollary 1 is in general not true if the equilibrium contract implements a mixed strategy $p^*(a)$. While the agent still gets the marginal equilibrium distribution over each variable right, her beliefs may also exhibit $p_{\mathcal{R}}(x_i \mid a'; p^*(a)) \neq p(x_i \mid a')$ for an action a' in the support of $p^*(a)$. Thus, the agent's expected utility conditional on a' may be biased, $\mathbb{E}_{\mathcal{R}}[u(w(y)) \mid a'; p^*(a)] \neq \mathbb{E}[u(w(y)) \mid a']$.

The second direct implication of Proposition 1 is the following result.

Corollary 2. *Suppose $(w(y), p(a))$ is an equilibrium contract. If $\mathcal{R} = (R, N)$ is perfect, the PC is satisfied at this contract if and only if this is also the case under the objective model $\mathcal{R}^*$.*

If $\mathcal{R}$ is perfect, the incentive scheme has to satisfy the same participation constraint as under the objective model. Thus, an agent with a misspecified – but perfect – model cannot be exploited. Throughout the paper, we will assume that $\mathcal{R}$ is perfect. As we see next, a perfect $\mathcal{R}$ does not imply that the principal cannot benefit from the agent's misperception.

3.2 Incentive Effects

We examine how a misspecification in the agent's subjective model $\mathcal{R}$ can change the equilibrium contract when $\mathcal{R}$ is perfect. By Corollary 2, only the incentive compatibility constraint can be affected by the misspecification. We analyze a simple setting with two effort levels $a \in \{0, 1\}$, two output levels $y \in \{y_L, y_H\}$ with $y_H > y_L$, and cost $c(1) = c > c(0) = 0$. The probability of output y_H increases in the agent's effort.

Consider the marketer example from the introduction. Figure 2 shows the objective model $\mathcal{R}^*$ and the agent's subjective model $\mathcal{R}$. Node 1 is the level of consumer information. It can be low ($x_1 = 0$) or high ($x_1 = 1$). Node 2 is the firm's reputation, which can be bad ($x_2 = 0$) or good ($x_2 = 1$). The subjective model $\mathcal{R}$ captures that the agent does not take reputation into account. For the objective probability distribution, we use the parametrization $p(x_i = 1 \mid x_{R(i)}) = \beta_i + \sum_{j \in R(i)} \beta_{ji} x_j$ for $i \in \{1, 2\}$ and $p(y_H \mid x_1, x_2) = \beta_3 + \beta_{13} x_1 + \beta_{23} x_2$. Making cold-calls increases consumer information, $\beta_{01} > 0$, and decreases reputation, $\beta_{02} < 0$;

Figure 2: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in the marketer example.

consumer information x_1 and reputation x_2 have a positive influence on sales, $\beta_{13} > 0$ and $\beta_{23} > 0$. We obtain the following result.

Proposition 2 (Marketer Example)**.** *Consider the marketer example of this subsection.*

(a) The simplification in the agent's subjective model $\mathcal{R}$ relaxes the IC for $\alpha = 1$.

(b) The optimal equilibrium contract implements $\alpha \in \{0, 1\}$. If and only if effort costs c are small enough, the optimal equilibrium contract implements $\alpha = 1$ and the principal strictly benefits from the simplification in the agent's subjective model $\mathcal{R}$.

Before we prove this result, we explain the intuition behind it and its implications. First, consider statement (a). When the principal implements $\alpha = 1$, the agent overestimates the drop in expected output when she exerts low instead of high effort. According to her subjective model $\mathcal{R}$, the only effect of her action on the output occurs through consumer information x_1; she does not take into account that a deviation to low effort would also have a positive effect on expected reputation x_2, which translates into a positive effect on expected output. Formally, the IC under the objective model $\mathcal{R}^*$ is

$$[\beta_{01}\beta_{13} + \beta_{02}\beta_{23}]\,(u(w(y_H)) - u(w(y_L))) - c \geq 0. \tag{6}$$

The term in squared brackets is the effect of effort on output and contains the consumer information channel $\beta_{01}\beta_{13}$ and the reputation channel $\beta_{02}\beta_{23}$. Under the subjective model $\mathcal{R}$, this second channel is missing, so that the IC becomes

$$\beta_{01}\beta_{13}\,(u(w(y_H)) - u(w(y_L))) \geq c. \tag{7}$$

Since the effect of effort on reputation β_{02} is negative, the simplification in $\mathcal{R}$ relaxes the IC. As long as $\alpha \in (0, 1)$, the reputation effect is partly reflected in $p(y_H \mid x_1)$; the extent of this depends on α since α affects the correlation between consumer information and reputation. A higher correlation between consumer information and reputation would mitigate some of the effect of the agent's misperception.

Next, consider statement (b). The observation that the principal implements a pure strategy would be trivial in the canonical framework with rational expectations. This is not the case here as the agent's perceived effect of effort on output $p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha)$ may vary non-monotonically in α. In the present setting, the perceived effect of effort on output is maximal at $\alpha = 1$, so that there is no reason for the principal to implement a mixed strategy. At the end of this subsection, we present an example where the unique optimal equilibrium contract indeed implements a mixed strategy $\alpha \in (0, 1)$.

Importantly, if the agent chooses a pure strategy, then, by Corollary 1 and the fact that $\mathcal{R}$ is perfect, she correctly anticipates the joint distribution over all variables in $\mathcal{R}$ conditional on her equilibrium action. Thus, in the data that the agent gets under the optimal equilibrium contract, there are no informational cues which could alarm her about a misspecification in her subjective model. This is a crucial difference between the present framework and models where beliefs about outcomes are biased for equilibrium actions.

Finally, the last part of statement (b) spells out that the principal strictly benefits from the simplification in $\mathcal{R}$ when effort costs are small enough so that it is profitable to implement high effort. The principal would have no incentive to correct the agent's view on the production process (if this were possible). This is of course not true in general. For example, if the agent's action has a positive effect on reputation, $\beta_{02} > 0$, the simplification in $\mathcal{R}$ tightens the IC for $\alpha = 1$ as the agent does not take all positive effects of her action on output into account.

Proof of Proposition 2. To illustrate our approach, we present the proof of Proposition 2. We first derive $p_{\mathcal{R}}(y_H \mid a; \alpha)$ for a given mixed equilibrium strategy $\alpha \in (0, 1)$. The agent's equilibrium belief about the joint probability distribution of the variables in $\mathcal{R}$ is given by $p_{\mathcal{R}}(a, x_1, y) = p(a)p(x_1 \mid a)p(y \mid x_1)$. Since node 0 and node 1 form a clique, the agent's belief about the joint probability distribution of a and x_1 is correct. Hence, $p(x_1 \mid a)$ is independent of α and we have $p(x_1 = 1 \mid a) = \beta_1 + \beta_{01}a$. However, $p(y \mid x_1)$ depends on α since the distribution over y also depends on x_2. To get $p(y \mid x_1)$, we first derive $p(x_2 = 1 \mid x_1)$, i.e., the probability that $x_2 = 1$ given that value x_1 is observed at node 1 when the agent's equilibrium action is α. We calculate

$$p(x_2 = 1 \mid x_1 = 1) = \frac{\alpha(\beta_1 + \beta_{01})(\beta_2 + \beta_{02}) + (1 - \alpha)\beta_1\beta_2}{\beta_1 + \alpha\beta_{01}}, \tag{8}$$

$$p(x_2 = 1 \mid x_1 = 0) = \frac{\alpha(1 - \beta_1 - \beta_{01})(\beta_2 + \beta_{02}) + (1 - \alpha)(1 - \beta_1)\beta_2}{1 - \beta_1 - \alpha\beta_{01}}. \tag{9}$$

With this we can calculate the equilibrium probability that output y_H realizes after observing

$x_1 = 1$ and $x_1 = 0$, respectively:

$$p(y_H \mid x_1 = 1) = \beta_3 + \beta_{13} + \frac{\alpha(\beta_1 + \beta_{01})(\beta_2 + \beta_{02}) + (1-\alpha)\beta_1\beta_2}{\beta_1 + \alpha\beta_{01}}\beta_{23}, \tag{10}$$

$$p(y_H \mid x_1 = 0) = \beta_3 + \frac{\alpha(1-\beta_1-\beta_{01})(\beta_2 + \beta_{02}) + (1-\alpha)(1-\beta_1)\beta_2}{1-\beta_1-\alpha\beta_{01}}\beta_{23}. \tag{11}$$

From $p_{\mathcal{R}}(a, x_1, y)$ we can now calculate the agent's subjective probability of a high output after high and low effort, respectively:

$$p_{\mathcal{R}}(y_H \mid a = 1; \alpha) = (\beta_1 + \beta_{01})p(y_H \mid x_1 = 1) + (1-\beta_1-\beta_{01})p(y_H \mid x_1 = 0), \tag{12}$$

$$p_{\mathcal{R}}(y_H \mid a = 0; \alpha) = \beta_1 p(y_H \mid x_1 = 1) + (1-\beta_1)p(y_H \mid x_1 = 0). \tag{13}$$

We then use these terms to compute the *IC* for $\alpha \in (0, 1)$,

$$[p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha)]\,(u(w(y_H)) - u(w(y_L))) = 0. \tag{14}$$

By taking the limit for $\alpha \to 1$, we obtain the *IC* for $\alpha = 1$, which is the inequality in (7). Since $\beta_{02} < 0$, this completes the proof of statement (a). To prove statement (b), note first that both *IC* and *PC* must be binding at the optimal equilibrium contract. Simple calculations show that $\beta_{01}, \beta_{13}, \beta_{23} > 0$ and $\beta_{02} < 0$ imply

$$p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha) \leq \beta_{01}\beta_{13} \tag{15}$$

for all $\alpha \in (0, 1]$; that is, when the agent exerts high effort with positive probability, her perceived effect of effort on output is largest at $\alpha = 1$. The principal then cannot gain from implementing a mixed strategy. Finally, given that the optimal equilibrium contract implements either $\alpha = 0$ or $\alpha = 1$, the last part of statement (b) follows from a simple comparison of expected profits under the equilibrium contracts that implement these two actions. □

Mixed strategy example. We show by example that it is not always optimal for the principal to implement a pure strategy. Consider again the marketer example. Assume that the agent is risk-neutral, protected by limited liability so that $w(y) \geq 0$, her outside option value is zero, and $y_L = 0$. Suppose payoff parameters are such that the principal optimally implements $\alpha > 0$. Standard arguments show that $w(y_L) = 0$, and that $w(y_H)$ is chosen so that the *IC* in (14) is satisfied. The principal's expected payoff from this contract is then

$$\mathbb{E}[V] = [\alpha p(y_H \mid a = 1) + (1-\alpha)p(y_H \mid a = 0)]\left(y_H - \frac{c}{\Delta_{\mathcal{R}}(\alpha)}\right), \tag{16}$$

where $\Delta_{\mathcal{R}}(\alpha) = p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha)$ is the agent's perceived effect of effort on output. The slope of $\Delta_{\mathcal{R}}(\alpha)$ at $\alpha = 1$ is

$$\left.\frac{d\Delta_{\mathcal{R}}(\alpha)}{d\alpha}\right|_{\alpha=1} = \beta_{01}\beta_{02}\beta_{23}\left(\frac{\beta_1}{\beta_1 + \beta_{01}} - \frac{1-\beta_1}{1-\beta_1-\beta_{01}}\right). \tag{17}$$

Let the agent's action have a positive impact on both consumer information and reputation, $\beta_{01} > 0$ and $\beta_{02} > 0$. Then for $\beta_{01} \to 1-\beta_1$ the slope in (17) converges to minus infinity. Hence, if all else equal β_{01} is sufficiently close to $1-\beta_1$, then, starting from $\alpha = 1$, a small reduction in α reduces $w(y_H)$, and in terms of profits, this reduction overcompensates the smaller probability of high output. The optimal equilibrium contract then implements a mixed strategy. Thus, when the agent is induced to switch between periods of working hard and periods of shirking, her effort appears to her as particularly important for the final output.

4 The Informativeness Principle

An important question in contract theory is on which information the principal should condition the agent's wage. For a setting with risk-averse agent who has unlimited liability, the informativeness principle states that the optimal contract conditions on an additional variable z if and only if it is informative about the agent's effort, i.e., if and only if the likelihood ratio $\frac{p(y,z|a')}{p(y,z|a)}$ varies in z for some y.[4] In this section, we derive a version of the informativeness principle that allows for boundedly rational agents. To this end, we exploit the fact that an agent with biased subjective beliefs may still have correct expectations about the joint distribution of contractible variables in equilibrium. We then apply our version of the informativeness principle to provide a rationale for why in executive compensation contracts peer-performance is mostly not used so that CEOs are rewarded for windfall gains.

The original version of the informativeness principle may no longer hold when the agent's subjective model $\mathcal{R}$ is misspecified. Consider the marketer example from Subsection 3.2 and assume that the principal can also condition the agent's wage on consumer information x_1. If the agent had rational expectations, the optimal wage scheme would condition both on consumer information x_1 and sales x_3 since neither variable is a sufficient statistic of the other (to avoid confusion below, we here use x_3 instead of y).[5] However, according to the agent's

[4] Whether this result holds or not depends on the formal details of the contracting problem; see Chaigneau et al. (2019) for a recent discussion and a further extension of the informativeness principle.

[5] A further interesting trade-off can be observed here. Recall from the marketer example that when the contract only conditions on sales x_3, the agent with subjective model $\mathcal{R}$ is control optimistic, which relaxes the *IC*. In contrast, when the contract only conditions on consumer information x_1, the agent has correct expectations about her expected payoff under alternative actions, so the *IC* is unaffected by the misspecification in $\mathcal{R}$.

subjective model $\mathcal{R}$, sales x_3 are just a noisy signal of consumer information x_1. Therefore, the optimal equilibrium contract only conditions on x_1 and appears as "incomplete."

We can generalize this finding and obtain a version of the informativeness principle that allows for misspecified subjective models $\mathcal{R}$. To get this statement, we assume that the agent's subjective model is such that she correctly anticipates the joint distribution over the two contractible variables y and z. Recall from Proposition 1 that this is the case if $\mathcal{R}$ is perfect and there is a link between y and z in $\mathcal{R}$ (so that they form a clique).

Proposition 3 (Informativeness Principle). *Suppose the agent is risk-averse and has unlimited liability. Let y and z be two contractible variables that are both part of the agent's subjective model $\mathcal{R}$. If $p_{\mathcal{R}}(z, y; p(a)) = p(z, y)$ for all $p(a) \in \Delta(A)$, the following statements hold:*

(a) Suppose that $a \in \{0, 1\}$ and $c(1) > c(0)$. The equilibrium contract that implements $\alpha = 1$ at lowest cost to the principal does not condition on z if and only if for all triples a, y, z we have $p_{\mathcal{R}}(z \mid y, a; \alpha = 1) = p_{\mathcal{R}}(z \mid y; \alpha = 1)$.

(b) If for all $p(a) \in \Delta(A)$ and all triples a, y, z we have $p_{\mathcal{R}}(z \mid y, a; p(a)) = p_{\mathcal{R}}(z \mid y; p(a))$, the optimal equilibrium contract does not condition on z.

Before we prove this result, we provide an interpretation and explain its implications. First, the condition $p_{\mathcal{R}}(z \mid y, a; p(a)) = p_{\mathcal{R}}(z \mid y; p(a))$ for all $p(a) \in \Delta(A)$ and all triples a, y, z indicates that, in the agent's mind, variable z is independent of her action conditional on variable y (regardless of the implemented action). If this condition is satisfied, the agent believes that z does not contain any information about her action that is not already in y. However, this condition alone does not imply that the optimal equilibrium contract does not condition the agent's wage on z. In addition, the agent's subjective belief about the joint equilibrium distribution of y and z needs to be correct. Otherwise, the principal may want to exploit the agent's biased perception of this distribution, and condition on z even if the agent thinks that z is uninformative about her action given y. This is equivalent to betting when two individuals have different prior beliefs about future events.

Second, Proposition 3 consists of two statements. Statement (a) is the informativeness principle for the case of binary action spaces. It is very similar to the original version: The statement implies that the optimal equilibrium contract that implements $\alpha = 1$ conditions on z if and only if the likelihood ratio $\frac{p_{\mathcal{R}}(y,z|a=0;\alpha=1)}{p_{\mathcal{R}}(y,z|a=1;\alpha=1)}$ varies in z for some y. Statement (b) for general finite action spaces is weaker since the additional information embedded in z may, according to the agent's subjective beliefs, only affect non-binding ICs.[6]

[6] This is a general issue of the informativeness principle and not specific to our framework.

Third, observe that Proposition 3 does not impose any further assumptions on the agent's subjective model $\mathcal{R}$. It therefore applies to all settings in which the agent's beliefs satisfy the conditions outlined in the proposition. Importantly, we can state sufficient conditions on $\mathcal{R}$ so that the agent's beliefs satisfy the conditional independence assumption. The Bayesian network literature establishes "d-separation" as a convenient tool to check conditional independence of two sets of variables in a model $\mathcal{R}$; we describe it in Online Appendix A.2. Here we give a simple implication of d-separation: Define a path τ in $\mathcal{R}$ as a sequence of nodes so that any adjacent nodes are linked in $\mathcal{R}$; τ is a directed path if the links between any two adjacent nodes in τ point in the same direction (from the former to the latter or vice versa). Variable z is independent from action a conditional on y in $\mathcal{R}$ if all paths from a to z are directed and contain y. Note that this is the case in the marketer example above.

Fourth, our Bayesian network framework allows for a causal interpretation of the informativeness principle. The optimal equilibrium contract conditions on both y and z, if the agent's action has partially independent effects on these two variables according to $\mathcal{R}$; it does not condition on z if, according to $\mathcal{R}$, variable z is a consequence of y. In this case, the optimal contract conditions on the variable that is "causally closer" to the agent's action.

Proof of Proposition 3. We first prove statement (b). Suppose the principal wishes to implement $p(a)$. Since the agent is risk-averse with unlimited liability and her action set A is finite, we can use the arguments in Grossman and Hart (1983) to show that the Kuhn-Tucker theorem yields necessary and sufficient conditions for an optimum. The optimal incentive scheme is therefore characterized by the first-order condition

$$\frac{1}{u'(w(y,z))} = \frac{p_{\mathcal{R}}(y,z;p(a))}{p(y,z)}\left[\mu + \sum_{a'\in A}\lambda_{a'}\frac{p_{\mathcal{R}}(y,z\mid a;p(a)) - p_{\mathcal{R}}(y,z\mid a';p(a))}{p_{\mathcal{R}}(y,z;p(a))}\right]. \tag{18}$$

By assumption, we have $p_{\mathcal{R}}(y,z;p(a)) = p(y,z)$. We can rewrite $p_{\mathcal{R}}(y,z\mid a;p(a))$ as

$$p_{\mathcal{R}}(y,z\mid a;p(a)) = p_{\mathcal{R}}(y\mid a;p(a))p_{\mathcal{R}}(z\mid y,a;p(a)) = p_{\mathcal{R}}(y\mid a;p(a))p_{\mathcal{R}}(z\mid y;p(a)), \tag{19}$$

where the last equality follows from the assumption $p_{\mathcal{R}}(z\mid y,a;p(a)) = p_{\mathcal{R}}(z\mid y;p(a))$ for all triples a,y,z. Similarly, we can write $p_{\mathcal{R}}(y,z;p(a)) = p_{\mathcal{R}}(y;p(a))p_{\mathcal{R}}(z\mid y;p(a))$. Hence, we get

$$p_{\mathcal{R}}(y,z\mid a;p(a)) - p_{\mathcal{R}}(y,z\mid a';p(a)) = \frac{p_{\mathcal{R}}(y,z;p(a))}{p_{\mathcal{R}}(y;p(a))}[p_{\mathcal{R}}(y\mid a;p(a)) - p_{\mathcal{R}}(y\mid a';p(a))]. \tag{20}$$

The first-order condition in (18) therefore simplifies to

$$\frac{1}{u'(w(y,z))} = \mu + \sum_{a'\in A}\lambda_{a'}\frac{p_{\mathcal{R}}(y\mid a;p(a)) - p_{\mathcal{R}}(y\mid a';p(a))}{p_{\mathcal{R}}(y;p(a))}. \tag{21}$$

Since the right-hand side of this first-order equation is independent of z, the optimal incentive scheme does not condition on z, which completes the proof. Next, we prove statement (a). Risk-aversion and unlimited liability imply that the optimal incentive scheme that implements $a = 1$ is characterized by the first-order condition

$$\frac{1}{u'(w(y,z))} = \frac{p_{\mathcal{R}}(y,z \mid a=1; \alpha=1)}{p(y,z \mid a=1)}\left[\mu + \lambda\left(1 - \frac{p_{\mathcal{R}}(y,z \mid a=0;\alpha=1)}{p_{\mathcal{R}}(y,z \mid a=1;\alpha=1)}\right)\right], \tag{22}$$

where μ, λ are strictly positive constants. As above, we can write $p_{\mathcal{R}}(y,z \mid a = 1; \alpha = 1) = p(y,z \mid a = 1)$, so that this first-order condition simplifies to

$$\frac{1}{u'(w(y,z))} = \mu + \lambda\left(1 - \frac{p_{\mathcal{R}}(y,z \mid a=0;\alpha=1)}{p_{\mathcal{R}}(y,z \mid a=1;\alpha=1)}\right). \tag{23}$$

Statement (a) then directly follows from this equation. □

As an application, we consider a setting in which the principal can condition the agent's wage both on her output $y \in \{y_L, y_H\}$ and on her relative performance $z \in \{-1, 0, 1\}$; the latter variable captures, for example, how the stock price of the company compares to that of the company's rivals. There is a common shock $x_1 \in \{0, 1\}$, e.g., the state of the economy, that positively affects both own output y and the rivals' output $x_3 \in \{y_L, y_H\}$. Output y has a small effect on output x_3. The objective model $\mathcal{R}^*$ on the left in Figure 3 illustrates this setting.

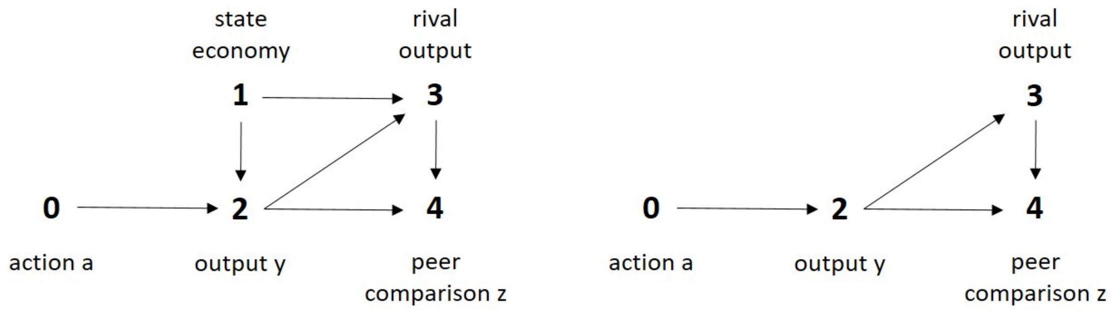


Figure 3: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in the peer-comparison example.

Under the objective model $\mathcal{R}^*$, the optimal equilibrium contract that implements high effort would, at any generic parametrization, condition the agent's wage both on output and relative performance. This can be established by visually inspecting $\mathcal{R}^*$ using d-separation.[7] The intuition is as follows: Suppose we know the agent's output y. Then information about the agent's action a provides additional information about the state of the economy x_1, and hence

[7]The "usual" way to see this is to consider a particular parametrization. Consider our linear specification with binary outcomes at all variables except z; for z we assume that $p(z = 1 \mid y > x_3) \approx 1$, $p(z = 0 \mid y = x_3) \approx 1$, and $p(z = -1 \mid y < x_3) \approx 1$. If the influence of y on x_3 is small enough, the optimal contract that implements high effort conditions on both variables, and the agent's wage increases in both y and z.

also additional information about peer performance z. Hence, a and z are not independent conditional on y in $\mathcal{R}^*$.

Now suppose that the agent does not take the common shock x_1 into account so that her subjective model is given by $\mathcal{R}$ on the right of Figure 3. Since $\mathcal{R}$ is perfect and the variables y and z are linked in $\mathcal{R}$, the agent correctly anticipates the equilibrium distribution over the two variables. Moreover, we can use the implication from d-separation above to see that a and z are independent conditional on y. The intuition is that, according to $\mathcal{R}$, output y is informative about relative performance z. However, if we already know y, we get no additional information about z from the agent's action. Proposition 3 then implies that the optimal equilibrium contract that implements $\alpha = 1$ only conditions on the agent's own output y. It is therefore incomplete and rewards the agent for windfall gains that come from good states of the economy. In the agent's mind, her relative performance is only a noisy signal of her own output. Hence conditioning her wage on relative performance would only increase the agent's exposure to risk and hence implementation costs.

Many actual compensation contracts indeed do not make use of peer-performance and reward executives for windfall gains. Bebchuk and Fried (2004) discuss this phenomenon and possible explanations. A popular explanation is that executives use their influence over the board of directors to alter their compensation, which then happens to increase in windfall gains. However, this theory cannot explain the inefficient risk allocation. In contrast, model misspecification can account for inefficient risk allocation. The manager's model is misspecified as in the application, for example, if she attributes the output to her action alone, or if she ignores the statistical implications of common shocks and therefore evaluates peer-performance as uninformative about her own action.

5 Behavioral Rationality

We learned in Section 3 that a simplification in the agent's subjective model may bias her beliefs about the production function, so that the incentive compatibility constraint is affected. However, does a simplification in $\mathcal{R}$ automatically imply that the agent's beliefs are biased? In this section, we show that the answer is negative. The agent may correctly anticipate the true production function even when her subjective model $\mathcal{R}$ omits variables from $\mathcal{R}^*$. When this statement holds for any parametrization of the extended production function that factorizes[8] according to $\mathcal{R}^*$, we say that the agent is "behaviorally rational." We state the formal definition.

[8]Importantly, we deviate in this section from our earlier assumption that $p(x^*)$ does not contain any additional conditional independence assumptions compared to $\mathcal{R}^*$.

Definition 3. *An agent with subjective model $\mathcal{R}$ is behaviorally rational if, at any probability distribution $p(x) \in \Delta(X)$ that factorizes according to $\mathcal{R}^*$, we have $p_{\mathcal{R}}(y \mid a; p(a)) = p(y \mid a)$ for all $a \in A$ and $p(a) \in \Delta(A)$.*

We characterize for a given objective model $\mathcal{R}^*$ when the agent is behaviorally rational, and when her beliefs about the production function remain unchanged if an (additional) node from N^* is dropped from her subjective model $\mathcal{R}$. We will see that two extended production functions – which involve the same set of nodes N^* and may give rise to the same $p(y \mid a)$ – can differ in the extent to which simplifications affect the agent's beliefs about $p(y \mid a)$. This extent depends on the "channels" in $\mathcal{R}^*$ through which the agent's action affects the output. Intuitively, they describe the agent's role in the organization, that is, which components or behaviors of others the agent affects directly or indirectly through her action. This allows us to identify several processes in an organization that potentially cause the agent to have biased beliefs about her productivity. We proceed as follows. In Subsection 5.1, we extend our marketer example to illustrate the influence of the agent's job on the scope for biased beliefs and control optimism. In Subsection 5.2, we characterize when the agent is behaviorally rational and generalize the main findings from Subsection 5.1.

5.1 The Agent's Job and the Scope for Control Optimism

We examine the interaction between the agent's job, model misspecification, and incentives. Let the agent first work as an ordinary marketer whose job is to increase sales. This time, making cold-calls is not part of her job. Her effort only has a (positive) effect on consumer information, for example, through informative advertising. Nevertheless, there is a group of employees engaged in telemarketing. Their effort – making cold-calls – impacts on consumer information and the firm's reputation in the usual manner. The objective model $\mathcal{R}^*$ on the left of Figure 4a represents the causal structure of this extended production function. Throughout, we use our parametrization with binary outcomes at all variables $i \in N^*$ and $p(x_i = 1 \mid x_{R(i)}) = \beta_i + \sum_{j \in R(i)} \beta_{ji} x_j$. The telemarketers either conduct cold-calls or not, $\beta_1 \in \{0, 1\}$; cold-calls have a negative effect on reputation, $\beta_{13} < 0$; consumer information has a positive effect on reputation, $\beta_{23} > 0$.[9] All formal proofs of this subsection are in Online Appendix A.3.

Imagine that the marketer neither takes into account the telemarketers' operation nor the firm's reputation so that her subjective model is given by $\mathcal{R}$ on the upper-left of Figure 4b. When choosing effort, she only considers how her action impacts sales through consumer

[9] Here we introduce the link between consumer information and reputation, and violate our full support assumption by assuming $p(x_1 = 1) \in \{0, 1\}$. The latter implies that in objective model $\mathcal{R}^*$ we could drop node 1 and factor the value $p(x_1 = 1)$ into the other conditional probabilities.

information. Does this misspecification change incentives? The answer is negative. We can show – using the results from the next subsection – that the agent's subjective beliefs about the production function are correct, so that $p_{\mathcal{R}}(y_H \mid a; \alpha) = p(y_H \mid a)$ for all $a \in \{0, 1\}$ and $\alpha \in [0, 1]$. Thus, given her role in the principal's project (as captured by $\mathcal{R}^*$), the subjective model $\mathcal{R}$ is rich enough to produce correct predictions for off-equilibrium actions. The agent may ignore important parts of the project and still act as if she were fully rational. The equilibrium contract is then the same as in the canonical model.

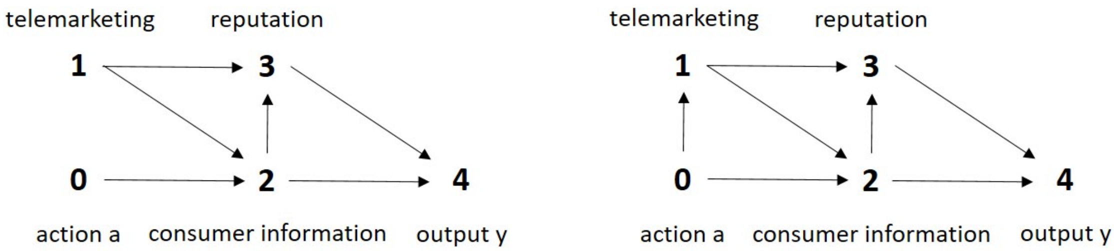


Figure 4a: Objective model $\mathcal{R}^*$ (left) when the agent works as ordinary marketer, and objective model $\mathcal{R}^{**}$ (right) when the agent works as "head of marketing."

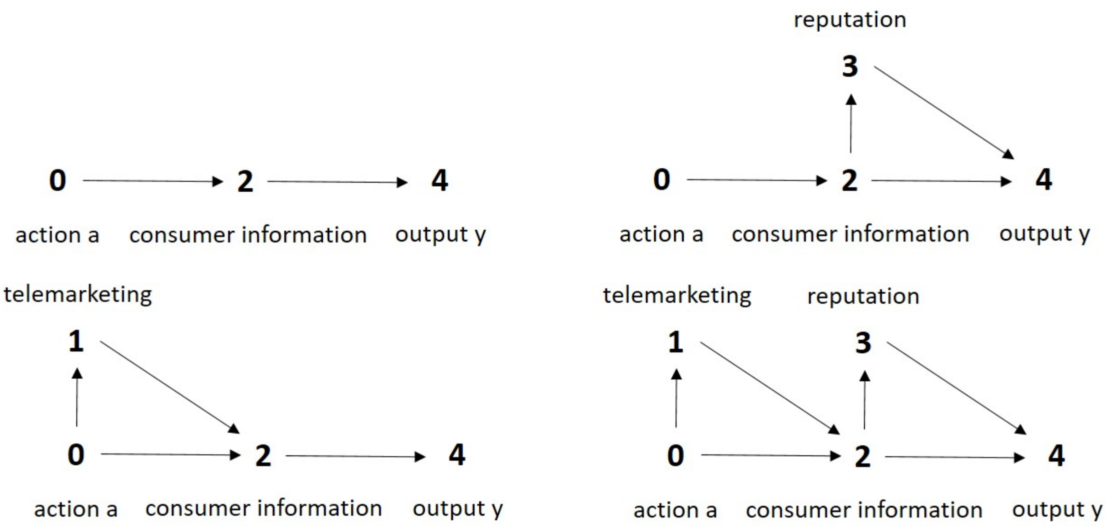


Figure 4b: Subjective models $\mathcal{R}$ (upper-left), $\mathcal{R}_1$ (upper-right), $\mathcal{R}_2$ (lower-left), and $\mathcal{R}_3$ (lower-right).

Importantly, telemarketing still matters for the principal since the probability distribution over sales depends on whether cold-calls are made or not. It is just not essential for the agent to know whether cold-calls take place. Her estimate of the production function implicitly takes into account the deterministic activity of the telemarketers, so that it will always be correct.

Is there any simplification that makes the agent overestimate the effectiveness of her effort, such that the principal benefits from it? Again, the answer is negative. If the agent does not take node 2 into account, she believes that her action has no consequences for the output. It would then be impossible to implement high effort. If only node 1 or only node 3 were omitted from her subjective model, the agent would again have correct beliefs about the production function. Thus, there is no scope for control optimism when the agent works as ordinary marketer.

Next, we alter the agent's job by promoting her to "head of marketing." Her action now influences the telemarketers' effort, for example, by motivating or inspiring the telemarketers. Instead of $p(x_1 = 1) = \beta_1$, we now have $p(x_1 = 1 \mid a) = \beta_1 + \beta_{01}a$. To keep things as close as possible to the previous case, we assume $\beta_1 = 0$ and $\beta_{01} = 1$.[10] Hence, the agent needs to act in order to get the telemarketers going (the telemarketer's activity is no longer an exogenous degenerate distribution). The objective model of the extended production function is given by $\mathcal{R}^{**}$ on the right of Figure 4a. How does a misspecification in the agent's subjective model now affect equilibrium beliefs and incentives in this environment?

Let us first assume that the agent has the same subjective model $\mathcal{R}$ as before (on the upper-left of Figure 4b). She neglects both the telemarketers' activity and the firm's reputation. This is not realistic since as "head of marketing" the agent should be aware of her subordinates' basic activities; so we will relax this assumption below. The misspecification now affects incentives. Under the objective model $\mathcal{R}^{**}$ the IC that implements $\alpha = 1$ would be

$$[(\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34}) + \beta_{01}\beta_{13}\beta_{34}](u(w(y_H)) - u(w(y_L))) \geq c. \tag{24}$$

The squared brackets contain the different channels through which effort affects output. The partial negative effect of effort on output through cold-calls and reputation is captured in the term $\beta_{01}\beta_{13}\beta_{34}$; it is negative since $\beta_{13} < 0$. Under the subjective model $\mathcal{R}$ the IC becomes

$$(\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34})(u(w(y_H)) - u(w(y_L))) \geq c. \tag{25}$$

Here the partial negative effect is missing so that the IC is relaxed. Note that through the estimate of the link between the agent's action and consumer information, the agent implicitly takes into account her positive influence on the telemarketers' effort, which in turn positively affects consumer information (see the term $\beta_{01}\beta_{12}$). Therefore, by being promoted to a job where the agent also influences telemarketing, she overestimates her productivity. The principal benefits from this since the misspecification reduces the need to provide effort incentives.

The assumption that the agent does not include the telemarketer's activity in her subjective model seems a bit odd, given that she is the head of marketing. Therefore, let her subjective model be given by $\mathcal{R}_2$ on the lower-left of Figure 4b. She now takes into account her influence on the telemarketers, and that the telemarketers increase consumer information when exerting effort. Does this inclusion correct, at least partly, the agent's beliefs? It turns out that this is

[10]Formally, we assume $\beta_1 = \varepsilon_1$ and $\beta_{01} = 1 - \varepsilon_2$ where $\varepsilon_1 < \varepsilon_2$, and consider the limit beliefs as $\varepsilon_1 \to 0$ and $\varepsilon_2 \to 0$. We show in the proofs for this subsection that our results do not depend on this assumption.

not the case. The models $\mathcal{R}$ and $\mathcal{R}_2$ produce the same beliefs about the effectiveness of effort, i.e., $p_{\mathcal{R}}(y_H \mid a; \alpha) = p_{\mathcal{R}_2}(y_H \mid a; \alpha)$ for all $a \in \{0, 1\}$ and $\alpha \in [0, 1]$. Including more variables does not necessarily make the agent more rational. This also holds for the models $\mathcal{R}_1$ and $\mathcal{R}_3$ in Figure 4b. Note that $\mathcal{R}_3$ is almost equal to the objective model $\mathcal{R}^{**}$, only the link between telemarketing and reputation is missing. Yet, all subjective models in this figure produce the same beliefs. Thus, a small misspecification in the agent's subjective model can render several important variables as inessential for estimating the production function.

Proposition 4 (Scope for Control Optimism)**.** *Consider the job examples of this subsection.*

(a) If the agent works as ordinary marketer (objective model $\mathcal{R}^*$), the misspecification in $\mathcal{R}$ has no effect on the IC and the optimal equilibrium contract is the same as in the canonical model. There is no simplification that generates control optimism.

(b) If the agent works as "head of marketing" (objective model $\mathcal{R}^{**}$), the misspecification in $\mathcal{R}$ generates control optimism and relaxes the IC; the subjective models $\mathcal{R}$, $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ generate the same beliefs about the production function.

Proposition 4 illustrates how the agent's job may matter for optimal incentives. The two jobs with objective models $\mathcal{R}^*$ and $\mathcal{R}^{**}$ may give rise to the same production function $p(y \mid a)$,[11] so that incentives would be identical under rational expectations. However, effort motivation is larger under a job with the objective model $\mathcal{R}^{**}$ when the agent's subjective model is simplified in a way that benefits the principal. The crucial difference between the jobs are the sets of channels through which the action affects the output. In the next subsection, we will formally define these channels.

The findings in Proposition 4 allow for several new interpretations. First, parts (a) and (b) combined demonstrate that an agent's degree of control optimism may be determined by the nature of her job. In the example, the agent with misspecified model $\mathcal{R}$ was behaviorally rational in her job as ordinary marketer, but overestimated the importance of her effort after being promoted to "head of marketing" where she influences the actions of others. Thus, in our framework, the agent's control optimism is not caused by certain features of her personality, but it is a consequence of her environment when her subjective model does not capture all empirical regularities of this environment.

Second, part (b) offers a new perspective on the phenomenon that managers often do not completely understand the difficulties that their rank-and-file workers face (e.g., Porter and Nohria 2018). Specifically, this can happen when an individual worked as specialist in her

[11] Specifically, when we denote parameters for the job with objective model $\mathcal{R}^*$ ($\mathcal{R}^{**}$) with "*" ("**") we only have to select parameters so that $\beta^*_{02}(\beta^*_{24} + \beta^*_{23}\beta^*_{34}) = (\beta^{**}_{02} + \beta^{**}_{01}\beta^{**}_{12})(\beta^{**}_{24} + \beta^{**}_{23}\beta^{**}_{34}) + \beta^{**}_{01}\beta^{**}_{13}\beta^{**}_{34}$.

previous position, but then was promoted to a management position where she influences the activity of individuals whose jobs she often does not fully understand. This lack of knowledge is typically regarded as a problem since it may lead to conflicts or inefficient managerial decisions. However, as our example shows, it also can have positive effects on effort motivation, in particular, when the agent does not take into account a partial negative effect of her subordinates' behavior on the final output, and she motivates this behavior through her action.

Third, what leads to control optimism in part (b) is the agent's ignorance of the partially negative consequences of her subordinates' activity for the final output. Our framework does not provide an explanation for why a certain node is in the agent's subjective model or not. However, in an organizational context, there can be good reasons why the agent only takes into account the positive aspects of her subordinates' activity. For example, subordinates may have an incentive to communicate why their effort is effective, and at the same time be reluctant to communicate the disadvantages of their activity.[12] In terms of our example, the telemarketers may know that cold-calls displease some customers. However, they may not want to make the agent aware of this, e.g., when having career concerns. Indeed, it is difficult for CEOs to obtain unbiased information about what their subordinates to. Porter et al. (2004) find that "[all] information coming to the top is filtered [...] Receiving solid information becomes even more difficult because immediately upon appointment, the CEO's relationships change. Former peers and subordinates who used to constitute an informal channel [...] go on their guard. Even those the CEO was closest to are wary of delivering bad news."

5.2 A General Result on Behavioral Rationality

To obtain a general result on behavioral rationality, we first assume that the objective model $\mathcal{R}^*$ is perfect, and that the agent's subjective model $\mathcal{R}$ is a simplification. Note that $\mathcal{R}$ will then be perfect. No v-structure emerges if we take out nodes from a perfect $\mathcal{R}^*$ and all links attached to them. The assumptions on $\mathcal{R}^*$ and $\mathcal{R}$ are not overly restrictive: Note that any probability distribution $p(x^*)$ factorizes according to some perfect DAG $\mathcal{R}^*$. The assumption on $\mathcal{R}$ is satisfied by almost all subjective models we consider in this paper. Below, we (partially) extend our behavioral rationality result to imperfect objective models. All formal proofs for this subsection are in Online Appendix A.4.

In the following, we characterize for any perfect $\mathcal{R}^*$ the subset of nodes the agent needs to have in her subjective model $\mathcal{R}$ so that she acts as if she had fully rational beliefs about the

[12]A large literature in organizational economics studies strategic information transmission in organizations (e.g., Aghion and Tirole 1997). The models in this literature are built on the common prior assumption, i.e., all parties have the same (correct) prior of what other parties may know. This is not the case in our framework. The crucial point here is that strategic communication may directly influence how the agent perceives the production process.

production function. We use the following definitions and results from the Bayesian network literature. Consider any DAG $\mathcal{R} = (N, R)$. Its skeleton $(N, \tilde{R})$ is obtained by making the DAG undirected. We have $i\tilde{R}j$ if and only if iRj or jRi.

Definition 4. *Two DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent if $p_{\mathcal{R}}(x) \equiv p_{\mathcal{G}}(x)$ for every $p(x) \in \Delta(X)$.*

Proposition 5 (Verma and Pearl 1991)**.** *Two DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent if and only if they have the same skeleton and v-structure.*

Two different models produce the same beliefs if they share the same skeleton and the same set of *v*-colliders. To illustrate, consider the two models in Figure 1. The DAGs $\mathcal{R}^*$ and $\mathcal{R}$ are not equivalent since they have different skeletons. Next, consider a DAG $\mathcal{G}$ that only differs from $\mathcal{R}$ in Figure 1 in that the link between the nodes 1 and 4 is reversed. $\mathcal{R}$ and $\mathcal{G}$ then have the same skeleton, but a different *v*-structure, so that they are not equivalent.

We need a few more definitions. A subset of nodes $M \subset N$ is called ancestral in $\mathcal{R}$ if for all nodes $i \in M$ we have $R(i) \subset M$. A path τ of length d from node i to node j is a sequence of nodes $\tau_0, \tau_1, ..., \tau_d$ so that $\tau_0 = i$, $\tau_d = j$, and $\tau_{h-1}\tilde{R}\tau_h$ for all $h \in \{1, ..., d\}$. The length of the shortest path between i and j is called the distance between these nodes and denoted by $d(i, j)$. A path of length d is active if there is no $h \in \{1, ..., d-1\}$ so that $\tau_{h-1}R\tau_h$ and $\tau_{h+1}R\tau_h$.

Define by $\mathcal{E}$ the set of DAGs in the equivalence class of $\mathcal{R}^*$ in which the action node 0 is ancestral (nothing influences the agent's action). In each of these DAGs, all active paths between the action node 0 and any node i point towards i. Thus, the assumption that node 0 is ancestral pins down the direction of many links in a perfect DAG. We call such links "fundamental links." There is a close connection between fundamental links and the set of nodes that can be removed while maintaining behavioral rationality.

Definition 5. *Consider two nodes $i, j \in N^*$. If iGj for all $\mathcal{G} = (G, N^*) \in \mathcal{E}$, then the link iGj is called fundamental link and denoted by iEj.*

An intuition for fundamental links is that they capture empirically relevant directions of causality (given agreement on the ancestral node). Specifically, they describe how the agent's action impacts on other variables. Consider $\mathcal{R}^*$ from Figure 1. Since the action node is ancestral, the links pointing from node 0 to other nodes are fundamental ($0R^*1$, $0R^*2$, and $0R^*3$). Thus, the two links pointing into the output node ($1R^*4$ and $3R^*4$) also must be fundamental. If we would turn around one of them, we would create a *v*-collider since there is no link between node 0 and node 4. The remaining links $1R^*2$, $1R^*3$, and $2R^*3$ are not fundamental. Below, we present an algorithm that identifies all fundamental links in any perfect DAG $\mathcal{R}^*$. For now, we go a step further and consider sequences of fundamental links.

Definition 6. *Let τ be an active path in $\mathcal{R}^*$. Then τ is a fundamental active path if all the links between neighboring nodes in τ are fundamental.*

Fundamental active paths are what we so far called "channels." Consider again $\mathcal{R}^*$ from Figure 1. The path $\tau = \{0, 1, 4\}$ is a fundamental active path since both links $0R^*1$ and $1R^*4$ are fundamental. In contrast, the active path $\tau' = \{0, 2, 3, 4\}$ is not fundamental since the link $2R^*3$ is not fundamental. We define the set of nodes that are part of at least one fundamental active path between the action and the output by

$$H^*(\mathcal{R}^*) := \{i \in N^* \mid i \text{ is part of a fundamental active path between 0 and } n \text{ in } \mathcal{R}^*\}.$$

It turns out that the nodes in $H^*(\mathcal{R}^*)$ are exactly those nodes the agent needs to have in her subjective model in order to be behaviorally rational, provided that her subjective model is a simplification. We can prove this by finding a DAG $\mathcal{G}$ that is equivalent to $\mathcal{R}^*$ and in which there are no links pointing from nodes in $N^* \setminus H^*(\mathcal{R}^*)$ to nodes in $H^*(\mathcal{R}^*)$. In this DAG, the nodes that are not in $H^*(\mathcal{R}^*)$ have no influence on the output, so the agent can safely ignore them. By Proposition 5, the agent correctly anticipates the production function if $H^*(\mathcal{R}^*) \subseteq N$.

Proposition 6 (Behavioral Rationality)**.** *Let $\mathcal{R}^*$ be a perfect DAG and let the agent's subjective DAG $\mathcal{R}$ be a simplification. The agent is behaviorally rational if and only if $\mathcal{R}$ contains all nodes from $H^*(\mathcal{R}^*)$.*

Proposition 6 implies that the agent does not necessarily have to take into account all variables of her (potentially) complex environment in order to be behaviorally rational. In particular, this holds independent of the parametrization of the extended production function. For example, when $p(x_1, ..., x_4 \mid a)$ factorizes according to $\mathcal{R}^*$ in Figure 1, the agent can ignore node 2 and still would behave as in the contracting model with common priors. The intuition is that when $H^*(\mathcal{R}^*) \subseteq N$, then the information captured through the variables in $H^*(\mathcal{R}^*)$ already includes the probabilistic information from variables outside $H^*(\mathcal{R}^*)$. Conversely, if the agent's subjective model does not include all variables from $H^*(\mathcal{R}^*)$, she is not behaviorally rational. In this case, we can find a parametrization of $p(x_1, ..., x_n \mid a)$ such that the incentive compatibility constraint is affected by the simplification in the agent's subjective model $\mathcal{R}$.

Next, Proposition 6 also shows that different misspecifications can have the same effect on incentives. Consider the two models $\mathcal{R}_1$ and $\mathcal{R}_2$ from the job example in Figure 4b. The set of nodes on fundamental active paths is the same for these two models, $H^*(\mathcal{R}_1) = H^*(\mathcal{R}_2) = \{0, 2, 4\}$. This implies that the agent's beliefs under these models are identical. Thus, it does not matter for the equilibrium contract whether the agent ignores node 1, node 3, or both nodes.

Therefore, the ignorance about one channel of causality may render another variable unimportant. A further interpretation is that two agents with different subjective models may have the same beliefs about the production function. We capture this result in a general statement. Consider a DAG $\mathcal{R} = (N, R)$ and a subset $\tilde{N} \subset N$. Denote by $\mathcal{R}^{[\tilde{N}]} = (\tilde{N}, \tilde{R})$ with $\tilde{R} = (\tilde{N} \times \tilde{N}) \cap R$ the DAG $\mathcal{R}$ restricted on $\tilde{N}$.

Corollary 3. *Let $\mathcal{R}_1 = (N_1, R_1)$ and $\mathcal{R}_2 = (N_2, R_2)$ be two perfect DAGs. Suppose there exists a DAG $\mathcal{R}_3$ so that $\mathcal{R}_3^{[N_1]} = \mathcal{R}_1$ and $\mathcal{R}_3^{[N_2]} = \mathcal{R}_2$. If $H^*(\mathcal{R}_1) = H^*(\mathcal{R}_2)$, then we have that $p_{\mathcal{R}_1}(y \mid a; p(a)) = p_{\mathcal{R}_2}(y \mid a; p(a))$ for all $a \in A$ and $p(a) \in \Delta(A)$.*

Identification of fundamental links. We provide an algorithm that identifies $H^*(\mathcal{R}^*)$ in perfect DAGs. Nodes that are connected by fundamental links in perfect DAGs exhibit characteristics that are easy to identify.

Proposition 7 (Fundamental Links)**.** *Let $\mathcal{R}^*$ be a perfect DAG and consider two adjacent nodes $i, j \in N^*$. The link iR^*j is fundamental if and only if at least one of the following conditions is satisfied:*

(a) we have $d(0, i) = d(0, j) - 1$;

(b) there exists a node $k \in N^*$ such that kEi and $k \notin R^*(j)$.

From this result we can derive an algorithm that finds all fundamental links in a perfect DAG $\mathcal{R}^*$. Let the topological ordering of $\mathcal{R}^*$ be such that every link is directed from an earlier to a later node. Then find for each node i the distance to the action node, $d(0, i)$. Links between nodes of differing distance are fundamental links. Next, check the links between nodes i, j that are of equal distance to the action node. Let N_d be the nodes that are at distance d to the action node. Consider the smallest element of N_d, say i, and any $j \in N_d$ with iR^*j. A link iR^*j is fundamental if and only if there exists a node k so that there is a fundamental link from k to i, but no link from k to j. Continue in this manner to evaluate all links between nodes in N_d, going sequentially from the smallest to the largest node in N_d. Do this for all distances $d > 0$.

It is not always simple to spot the nodes that are not in $H^*(\mathcal{R}^*)$. In this case, Proposition 7 is helpful. Consider, for example, the perfect DAG $\mathcal{R}^*$ in Figure 5. Condition (a) from Proposition 7 implies that all links which connect nodes of different distances to the action node are fundamental. The remaining links are $1R^*2$, $3R^*4$, $3R^*5$, $4R^*5$, $4R^*6$, and $5R^*6$. Condition (b) from Proposition 7 then implies that $4R^*6$ and $5R^*6$ are fundamental links, while the remaining links are non-fundamental. We therefore get $H^*(\mathcal{R}^*) = N^* \setminus \{3\}$.

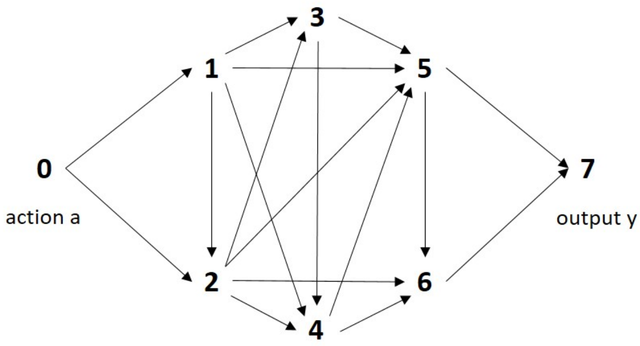


Figure 5: Example DAG $\mathcal{R}^*$.

Imperfect objective models. In several applications, the objective model $\mathcal{R}^*$ is imperfect. Nevertheless, we can apply Proposition 6 to these models to detect nodes that can be dropped from the agent's subjective model while preserving behavioral rationality. Note that one can make any imperfect DAG perfect by adding links between nodes that create v-colliders. If $p(x^*)$ is consistent with $\mathcal{R}^*$, it is consistent with any DAG that adds links to $\mathcal{R}^*$. Consider a perfect DAG $\hat{\mathcal{R}}$ that is identical to the imperfect DAG $\mathcal{R}^*$ except that it has additional links. Suppose all these additional links disappear when we take out the nodes that are not in the agent's subjective model $\mathcal{R} = (N, R)$. Then from Proposition 6 we immediately get that the agent is behaviorally rational if N contains $H^*(\hat{\mathcal{R}})$. We state this result formally.

Corollary 4. *Let $\mathcal{R}^* = (N^*, R^*)$ be the (possibly imperfect) objective DAG and $\mathcal{R} = (N, R)$ the agent's subjective DAG. The agent is behaviorally rational if there is a perfect DAG $\hat{\mathcal{R}} = (N^*, \hat{R})$ with $\hat{\mathcal{R}}^{[N]} = \mathcal{R}$ and $R^* \subseteq \hat{R}$, so that $\mathcal{R}$ contains all nodes from $H^*(\hat{\mathcal{R}})$.*

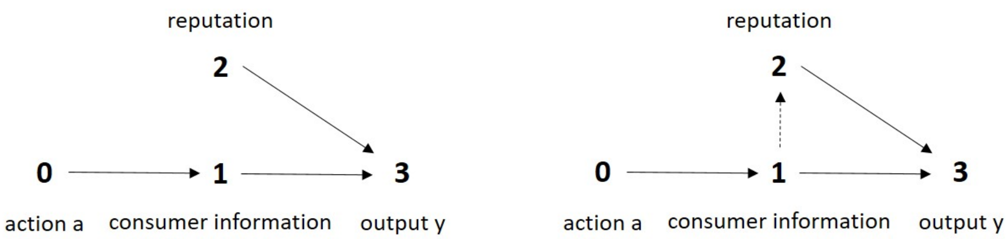


Figure 6: Imperfect model $\mathcal{R}^*$ (left) and perfect model $\hat{\mathcal{R}}$ (right).

As an illustration, consider the marketer example from Subsection 3.2 when the agent's effort has no impact on reputation, $\beta_{02} = 0$. The causal structure of this production function is then given by the imperfect DAG $\mathcal{R}^*$ on the left of Figure 6. The perfect DAG $\hat{\mathcal{R}}$ on the right is identical, except that it has an additional link $1\hat{R}2$. In this model, node 2 is not on a fundamental active path. Hence, the agent is behaviorally rational if her subjective model equals $\mathcal{R}^*$ without node 2.

6 Comparative Statics

One advantage of our approach to contracting with boundedly rational agents is that beliefs are derived endogenously from the true production process. This allows us to analyze how the optimal equilibrium contract varies in the parameters of the environment. In this section, we revisit two comparative statics that have received considerable attention in the literature: the trade-off between risk and incentives, and the relationship between team size and incentives. In both cases, the empirical evidence on these comparative statics conflicts with the predictions of the canonical model. We briefly discuss how we can explain these findings within our framework. All formal details of this section are relegated to Online Appendix A.5 and A.6.

Risk and Incentives. A risk-averse agent demands a risk premium for accepting a wage schedule with uncertain wage payments. Thus, an increase in risk drives up the costs of providing incentives. Consequently, the provision of effort incentives should decrease in the riskiness of the environment. However, empirically this relationship does not hold in general (e.g., Prendergast 2002). Field evidence on the relationship between risk an incentives for CEO compensation is mixed, and for other domains, such as franchising, a positive relationship can be observed. In contrast, a negative relationship is obtained in lab experiments where subjects know the true production function (Corgnet and Hernán-González 2019).

We can use our marketer example to show how the relationship between risk and incentives may become positive when the agent has a simplified model of the project. We consider a mean-preserving spread in $p(y \mid a)$, so that under the objective model $\mathcal{R}^*$ the provision of incentives becomes more costly when there is more risk. However, if the agent's subjective model is misspecified, there can be an additional effect of risk on incentives: The agent may perceive the riskier environment as one in which her action is more important for the output, which relaxes the incentive compatibility constraint. If this effect is sufficiently strong relative to the risk premium effect, there can be a positive relationship between risk and incentives.

Team Size and Incentives. In a team incentive problem, effort incentives are provided by tying each team member's payoff to the joint output y. The effectiveness of team incentives is constrained by the size of the team. When an agent's relative contribution to the output becomes small, it is typically no longer profitable for the principal to condition her pay on y, as the incentive effect would be outweighed by the costs of incentive provision (e.g., Kandel and Lazear 1992). An important implication of this result is that stock-options should be granted only to those employees whose actions significantly move the stock price. However, many firms grant stock options also to non-executive employees, and there is evidence that these have positive incentive effects (e.g., Hochberg and Lindsey 2010).

We can provide a belief-based explanation for this phenomenon in a setting with many agents. Each agent produces an intermediate output which positively affects the final output. A common shock affects all intermediate outputs in the same direction. If an agent ignores the intermediate outputs by other agents, she perceives a strong relationship between her intermediate output and the final output. She then overestimates the importance of her effort, which relaxes the incentive compatibility constraint. We demonstrate that output-based incentives then can remain effective even when the team becomes arbitrary large.

7 Conclusion

In this paper, we applied Spiegler's (2016) Bayesian network framework to analyze optimal contracting in a principal-agent setting where the agent forms beliefs about the production function based on a misspecified model of the principal's project. The objective causal model may be very complex, and may contain empirical regularities that the agent does not consider due to cognitive limitations or because they are never brought to her attention.

The optimal contract exhibits the following features. First, it does not exploit the agent if her subjective model takes into account the correlation between variables in her model that have a joint influence on a third variable (in which case it is "perfect"). Nevertheless, the principal benefits from a misspecification in the agent's perfect subjective model if it makes the agent control optimistic so that the incentive compatibility constraint is relaxed. Second, when the agent correctly anticipates the joint distribution of contractible variables, the optimal contract conditions on an additional variable only if it is informative about the action according to the agent's model. For example, the optimal contract may not condition on peer-performance if the agent interprets this variable as a noisy signal of her own output. Third, the optimal contract is identical to the rational benchmark if the agent is behaviorally rational. We characterize when this is the case, and apply this finding to show how the scope for control optimism may depend on the agent's job. For example, a front-line worker may not fully understand the workings of the organization around her, but still act as if she were fully rational. In contrast, a high-ranking manager, who affects the output by influencing the behavior of many subordinates, overestimates her own productivity if she does not take into account the challenges that her subordinates face in their routines.

We focused on a simple contracting framework so that we can identify precisely how misspecifications in the agent's model affect incentive contracts. Future research can extend the framework by considering team incentives, relational contracts, strategic communication and delegation. The Bayesian network approach offers a very disciplined tool to study the effects of bounded rationality on organizations, and we think that our results are useful in this respect.

References

AGHION, PHILIPPE, AND JEAN TIROLE (1997): "Formal and real authority in organizations," *Journal of Political Economy*, 105(1), 1–29.

AUSTER, SARAH (2013): "Asymmetric awareness and moral hazard," *Games and Economic Behavior*, 82, 503–521.

BEBCHUK, LUCIAN ARYE, AND JESSE FRIED (2004): *Pay without performance: The unfulfilled promise of executive compensation*, Harvard University Press, Cambridge.

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," *Quarterly Journal of Economics*, 117(23), 871–915.

BRUNNERMEIER, MARKUS, AND JONATHAN PARKER (2005): "Optimal expectations," *American Economic Review*, 95(14), 1092–1118.

CHAIGNEAU, PIERRE, ALEX EDMANS, AND DANIEL GOTTLIEB (2019): "The informativeness principle without the first-order approach," *Games and Economic Behavior*, 113, 743–755.

CORGNET, BRICE, AND ROBERTO HERNÁN-GONZÁLEZ (2019): "Revisiting the trade-off between risk and incentives: The shocking effect of random shocks?," *Management Science*, 65(3), 1096–1114.

DE LA ROSA, ENRIQUE (2011): "Overconfidence and moral hazard," *Games and Economic Behavior*, 73(2), 429–451.

DEKEL, EDDIE, BARTON LIPMAN, AND ALDO RUSTICHINI (1998): "Standard state-space models preclude unawareness," *Econometrica*, 66(1), 159–173.

ELIAZ, KFIR, AND RANI SPIEGLER (2018): "A model of competing narratives," CEPR Discussion Paper No. DP13319.

ELIAZ, KFIR, RANI SPIEGLER, AND HEIDI THYSEN (2019): "Strategic interpretations," CEPR Discussion Paper No. 13441.

FANG, HANMING, AND GIUSEPPE MOSCARINI (2005): "Morale hazard," *Journal of Monetary Economics*, 52(4), 749–777.

GERVAIS, SIMON, AND ITAY GOLDSTEIN (2007): "The positive effects of biased self-perceptions in firms," *Review of Finance*, 11(3), 453–496.

GROSSMAN, SANFORD, AND OLIVER HART (1983): "An analysis of the principal-agent problem," *Econometrica*, 51(1), 7–45.

HEIFETZ, AVIAD, MARTIN MEIER, AND BURKHARD SCHIPPER (2006): "Interactive unawareness," *Journal of Economic Theory*, 130(1), 78–94.

HEIFETZ, AVIAD, MARTIN MEIER, AND BURKHARD SCHIPPER (2013): "Unawareness, beliefs, and speculative trade," *Games and Economic Behavior*, 77(1), 100–121.

HOCHBERG, YAEL, AND LAURA LINDSEY (2010): "Incentives, targeting, and firm performance: An analysis of non-executive stock options," *Review of Financial Studies*, 23(11), 4148–4186.

HOLMSTRÖM, BENGT (1979): "Moral hazard and observability," *Bell Journal of Economics*, 10(1), 74–91.

IMMORDINO, GIOVANNI, ANNA MARIA MENICHINI, MARIA GRAZIA ROMANO (2015): "Contracts with wishful thinkers," *Journal of Economics and Management Strategy*, 24(4), 863–886.

KANDEL, EUGENE, AND EDWARD LAZEAR (1992): "Peer pressure and partnerships," *Journal of Political Economy*, 100(4), 801–817.

KŐSZEGI, BOTOND (2006): "Ego utility, overconfidence, and task choice," *Journal of the European Economic Association*, 4(4), 673– 707.

KŐSZEGI, BOTOND (2014): "Behavioral contract theory," *Journal of Economic Literature*, 52(4), 1075–1118.

PEARL, JUDEA (2009): *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

PORTER, MICHAEL, JAY LORSCH, AND NITIN NOHRIA (2004): "Seven surprises for new CEOs," *Harvard Business Review*, 82(10), 62–72.

PORTER, MICHAEL, AND NITIN NOHRIA (2018): "How CEOs manage time," *Harvard Business Review*, 96(4), 42–51.

PRENDERGAST, CANICE (2002): "The tenuous trade-off between risk and incentives," *Journal of Political Economy*, 110(5), 1071–1102.

SANTOS-PINTO, LUÍS (2008): "Positive self-image and incentives in organisations," *Economic Journal*, 118(531), 1315–1332.

SAUTMANN, ANJA (2007): "Self-confidence in a principal-agent relationship," unpublished manuscript, Brown University.

SAUTMANN, ANJA (2013): "Contracts for agents with biased beliefs: Some theory and an experiment," *American Economic Journal: Microeconomics*, 5(3), 124–156.

SCHENONE, PABLO (2019): "Causality: A decision theoretic framework," Working Paper, California Institute of Technology.

SIMON, HERBERT (1947): *Administrative Behavior*, Macmillan, London.

SIMON, HERBERT (1955): "A behavioral model of rational choice," *Quarterly Journal of Economics*, 69(1), 99–118.

SPIEGLER, RAN (2016): "Bayesian networks and boundedly rational expectations," *Quarterly Journal of Economics*, 131(3), 1243–1290.

SPIEGLER, RAN (2017): "Data monkeys: A procedural model of extrapolation from partial statistics," *Review of Economic Studies*, 84(4), 1818–1841.

SPIEGLER, RAN (2019): "Can agents with causal misperceptions be systematically fooled?," *Journal of the European Economic Association*, forthcoming.

SPINNEWIJN, JOHANNES (2013): "Insurance and perceptions: How to screen optimists and pessimists," *Economic Journal*, 123(569), 606–633.

SPINNEWIJN, JOHANNES (2015): "Unemployed but optimistic: Optimal insurance design with biased beliefs," *Journal of the European Economic Association*, 13(1), 130–167.

VON THADDEN, ERNST-LUDWIG, AND XIAOJIAN ZHAO (2012): "Incentives for unaware agents," *Review of Economic Studies*, 79(3), 1151–1174.

VON THADDEN, ERNST-LUDWIG, AND XIAOJIAN ZHAO (2014): "Multitask agency with unawareness," *Theory and Decision*, 77(2), 197–222.

VAN DEN STEEN, ERIC (2005): "Organizational beliefs and managerial vision," *Journal of Law, Economics, and Organization*, 21(1), 256–283.

VERMA, THOMAS, AND JUDEA PEARL (1991): "Equivalence and synthesis of causal models," *Uncertainty in Artificial Intelligence*, 6, 255–268.

A Online Appendix

A.1 Existence of a Personal Equilibrium

We show that a personal equilibrium exists at any admissible $\mathcal{R}$ and $w(y) \in W$. Note that $\Delta(A)$ is non-empty, compact, and convex. Define the best-response correspondence $BR : \Delta(A) \to \Delta(A)$ by

$$BR(p(a)) = \arg \max_{\tilde{p}(a') \in \Delta(A)} \sum_{a' \in A} \sum_{y \in Y} \tilde{p}(a')[p_{\mathcal{R}}(y \mid a'; p(a))u(w(y)) - c(a')]. \tag{A.1}$$

For every $p(a) \in \Delta(A)$ we have that $BR(p(a))$ is non-empty and convex. The latter statement follows since any convex combination of pure actions that are optimal for the agent is an element of $BR(p(a))$. Definition 1 and the factorization formula in (2) imply that the agent's beliefs $p_{\mathcal{R}}(y \mid a'; p(a))$ are continuous in $p(a)$. Therefore, we also must have that $\sum_{a' \in A} \sum_{y \in Y} \tilde{p}(a')[p_{\mathcal{R}}(y \mid a'; p(a))u(w(y)) - c(a')]$ is continuous in $p(a)$. Hence, $BR(p(a))$ is upper hemi-continuous. The existence of a personal equilibrium then follows from Kakutani's theorem.

A.2 A Brief Introduction to d-separation

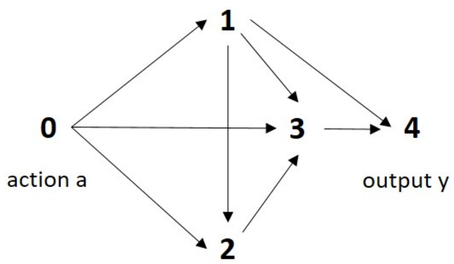
We briefly introduce the concept of d-separation, a result from the Bayesian network literature that allows us to check, for any given model $\mathcal{R}$, whether two variables (or two sets of variables) are independent when conditioning on a third variable (or set of variables). For simple models $\mathcal{R}$ it can be used as visual inspection tool; for complex models, there exists an algorithm for checking d-separation (Geiger et al. 1990). We use the definition of a (directed) path τ from Section 4. A node j is a descendant of node i if there exists a directed path from i to j. For convenience, we use the notation $i \to j$ instead of iRj in this section. The following definitions and result are adopted from Pearl (2009).

Definition 7. *A path τ is blocked in $\mathcal{R} = (R, N)$ by a set of variables $M \subset N$ if and only if one of the following condition holds:*

(a) τ contains variables i, m, j with $m \in M$ so that $i \to m \to j$ or $i \leftarrow m \to j$, or

(b) τ contains variables i, m, j so that $i \to m \leftarrow j$, $m \notin M$, and no descendant of m is in M.

To illustrate, consider the DAG $\mathcal{R}^*$ from Figure 1 in the paper, reproduced here on the left of Figure 7. The path $\tau = 0 \to 2 \leftarrow 1 \to 3 \to 4$ between the nodes 0 and 4 is blocked by node 1 and node 3, but not by node 2. To see this, note that conditions (a) and (b) are both satisfied



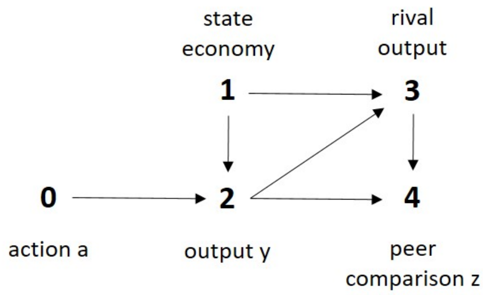


Figure 7: Objective model $\mathcal{R}^*$ from Figure 1 (left) and objective model $\mathcal{R}^*$ from Figure 3 (right).

if we define $M = \{1\}$, or $M = \{3\}$; however, none of the conditions is satisfied if we define $M = \{2\}$.

Definition 8. *Let $\mathcal{R} = (R, N)$ be a DAG and M', M'', M disjoint subsets of N. M' and M'' are d-separated by M in $\mathcal{R}$, if M blocks every path between any node in M' and any node in M''.*

Consider the DAG $\mathcal{R}^*$ from Figure 3 in the paper, reproduced here on the right of Figure 7. We check whether the nodes 0 and 4 are d-separated in $\mathcal{R}^*$ by $M = \{2\}$. For this, we have to consider three paths, $\tau = 0 \to 2 \to 4$, $\tau' = 0 \to 2 \leftarrow 1 \to 3 \to 4$, and $\tau'' = 0 \to 2 \to 3 \to 4$. By condition (a) in Definition 7, the paths τ and τ'' are blocked by $M = \{2\}$. In contrast, the path τ' is not blocked by $M = \{2\}$. Hence, the nodes 0 and 4 are not d-separated in $\mathcal{R}^*$ by $M = \{2\}$. However, they are d-separated in $\mathcal{R}^*$ by $M = \{1, 2\}$, $M = \{2, 3\}$, or $M = \{1, 2, 3\}$. Suppose, for example, that $M = \{1, 2\}$. Now not only the paths τ and τ'' are blocked according to condition (a) in Definition 1, but also path τ' (we see this from the segment $2 \leftarrow 1 \to 3$). The implication of d-separation is given in the following result.

Proposition 8 (Implications of d-separation)**.** *If the variables 0 and n are d-separated by variable i in $\mathcal{R}$, then $p_{\mathcal{R}}(x_n \mid x_0, x_i; p(a)) = p_{\mathcal{R}}(x_n \mid x_i; p(a))$ for all $p(a) \in \Delta(A)$ and all triples x_0, x_i, x_n. If the variables 0 and n are not d-separated by variable i in $\mathcal{R}$, then x_0 and x_n are dependent conditional on x_i for at least one distribution compatible with $\mathcal{R}$.*

A.3 Omitted Proofs from Subsection 5.1

We first derive the IC under the objective model $\mathcal{R}^*$. The probabilities of high output after high and low effort, respectively, are given by

$$
\begin{aligned}
p(y_H \mid a = 1) &= \beta_4 + [\beta_2 + \beta_{02} + (\beta_1 + \beta_{01})\beta_{12}]\beta_{24} \\
&\quad + [\beta_3 + (\beta_1 + \beta_{01})\beta_{13} + (\beta_2 + \beta_{02} + (\beta_1 + \beta_{01})\beta_{12})\beta_{23}]\beta_{34}, \qquad (A.2) \\
p(y_H \mid a = 0) &= \beta_4 + [\beta_2 + \beta_1\beta_{12}]\beta_{24} + [\beta_3 + \beta_1\beta_{13} + (\beta_2 + \beta_1\beta_{12})\beta_{23}]\beta_{34}, \qquad (A.3)
\end{aligned}
$$

so that the effect of effort on the probability of high output equals

$$p(y_H \mid a = 1) - p(y_H \mid a = 0) = (\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34}) + \beta_{01}\beta_{13}\beta_{34}. \tag{A.4}$$

Next, we drive the *IC* under the subjective model $\mathcal{R}$ when the equilibrium action is $\alpha \in [0, 1]$. We calculate

$$p(x_1 = 1 \mid x_2 = 1) = \frac{\alpha(\beta_1 + \beta_{01})(\beta_2 + \beta_{02} + \beta_{12}) + (1-\alpha)\beta_1(\beta_2 + \beta_{12})}{\beta_2 + \beta_1\beta_{12} + \alpha(\beta_{02} + \beta_{01}\beta_{12})}, \tag{A.5}$$

$$p(x_1 = 1 \mid x_2 = 0) = \frac{\alpha(\beta_1 + \beta_{01})(1 - \beta_2 - \beta_{02} - \beta_{12}) + (1-\alpha)\beta_1(1 - \beta_2 - \beta_{12})}{1 - \beta_2 - \beta_1\beta_{12} - \alpha(\beta_{02} + \beta_{01}\beta_{12})}, \tag{A.6}$$

and

$$p(x_3 = 1 \mid x_2 = 1) = \beta_3 + p(x_1 = 1 \mid x_2 = 1)\beta_{13} + \beta_{23}, \tag{A.7}$$

$$p(x_3 = 1 \mid x_2 = 0) = \beta_3 + p(x_1 = 1 \mid x_2 = 0)\beta_{13}. \tag{A.8}$$

The agent's belief about the probability of high output after $x_2 = 1$ and $x_2 = 0$, respectively, is therefore given by

$$p(y_H \mid x_2 = 1) = \beta_4 + \beta_{24} + [\beta_3 + p(x_1 = 1 \mid x_2 = 1)\beta_{13} + \beta_{23}]\beta_{34}, \tag{A.9}$$

$$p(y_H \mid x_2 = 0) = \beta_4 + [\beta_3 + p(x_1 = 1 \mid x_2 = 0)\beta_{13}]\beta_{34}. \tag{A.10}$$

The agent correctly anticipates $p(x_2 \mid a)$. Hence, her belief about the effect of effort on the probability of high output under $\mathcal{R}$ equals

$$\begin{aligned} p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha) &= (\beta_{02} + \beta_{01}\beta_{12})(\beta_{24} + \beta_{23}\beta_{34}) + (\beta_{02} + \beta_{01}\beta_{12})\beta_{13}\beta_{34} \\ &\quad \times [p(x_1 = 1 \mid x_2 = 1) - p(x_1 = 1 \mid x_2 = 0)]. \end{aligned} \tag{A.11}$$

Recall that $\beta_{13} < 0$. By comparing (A.4) and (A.11) we get that at $\alpha = 1$ the misspecification in $\mathcal{R}$ relaxes the *IC* if and only if

$$\beta_{01} > \frac{\beta_{12}(\beta_1 + \beta_{01})(1 - \beta_1 - \beta_{01})(\beta_{02} + \beta_{01}\beta_{12})}{(1 - \beta_2 - \beta_{02} - \beta_{12}(\beta_1 + \beta_{01}))(\beta_2 + \beta_{02} + \beta_{12}(\beta_1 + \beta_{01}))}, \tag{A.12}$$

which implies the statement in the main text.

Proof of Proposition 4. We prove the statements in (a). Since $\beta_1 \in \{0, 1\}$, we can rewrite the probability model without variable 1. The corresponding objective model $\tilde{\mathcal{R}}^*$ equals $\mathcal{R}^*$ in Figure 4a without node 1. We now apply Propositions 6 and 7. In model $\tilde{\mathcal{R}}^*$, node 3 is not on a fundamental active path. Hence, the agent with subjective model $\mathcal{R}$ is behaviorally rational, which yields the results. We prove the statements in (b). The first statement is shown in the text. The second statement follows from Corollary 3. Note that, in all models of Figure 4b, the set of nodes on fundamental active paths is identical. □

A.4 Omitted Proofs from Subsection 5.2

We first prove Proposition 7 and then Proposition 6. To this end, we prove several intermediate results. We first note that in a perfect DAG $\mathcal{R}^*$ the link iR^*j is fundamental if the nodes i and j differ in their distance to the action node 0.

Lemma 1. *Let $i, j \in N^*$ be adjacent nodes in $\mathcal{R}^*$. If $d(0,i) = d(0,j) - 1$, then iEj.*

Proof. First, suppose $d(0,i) = 0$ so that $i = 0$. Since node 0 is ancestral, we must have iGj in every DAG $\mathcal{G} \in \mathcal{E}$. Next, suppose $d(0,i) = d > 0$. Since $\mathcal{R}^*$ is perfect and node 0 is ancestral, there exists an active path of length d from node 0 to node i. Denote by k the direct ancestor of i on this path. There cannot exist a link between k and j, otherwise we would have $d(0,i) = d(0,k)$, a contradiction. Thus, we must have iGk in every DAG $\mathcal{G} \in \mathcal{E}$, otherwise we would have a v-collider at node i. □

Lemma 2. *Let $i, j \in N^*$ and iR^*j. If there exists a node $k \in N^*$ such that kEi and $k \notin R^*(j)$, then iEj.*

Proof. If there is a fundamental link from node k to node i, then iR^*j implies that we cannot have jR^*k. Otherwise, we would have a directed cycle. Node j and node k are therefore not adjacent. Hence, if jGi in some DAG $\mathcal{G} \in \mathcal{E}$, there would be a v-collider at i, a contradiction. □

The "if"-statement of Proposition 7 follows directly from Lemma 1 and Lemma 2. For the "only if"-statement we need two more results. The first one provides a condition under which a link is not fundamental.

Lemma 3. *Let $i, j \in N^* \setminus \{0\}$ and iR^*j. If $R^*(i) \subset R^*(j)$, then the link between i and j is not fundamental.*

Proof. Consider the DAG $\mathcal{G} = (G, N^*)$ that is identical to $\mathcal{R}^*$ except that it reverses the link between i and j. The assumption $R^*(i) \subset R^*(j)$ rules out that there are v-colliders in $\mathcal{G}$. Assume that there is a cycle in $\mathcal{G}$. Since $\mathcal{R}^*$ is acyclic, the cycle must contain jGi. Further, there must exists a node k and a link kGj which is part of the cycle. Since $\mathcal{R}^*$ is perfect, we must have $k\tilde{R}^*i$. Assume first that we have kR^*i. Then jGi implies that kGi is not part of the cycle. Thus, there must exist an active path τ of some length d so that $\tau_0 = i$ and $\tau_d = k$. But then there is a cycle consisting of the link kGi and τ. This cycle also exists in $\mathcal{R}^*$, a contradiction. Next, assume that we have iR^*k. Since $i \neq 0$ and $R^*(i) \subset R^*(j)$, there exists a node l with lR^*i and lR^*j. Since $\mathcal{R}^*$ is perfect, we also must have $l\tilde{R}^*k$. The same applies to all $l' \in R^*(i)$. Hence, starting from $\mathcal{R}^*$, we can reverse the links between i and j as well as between i and k and obtain a DAG $\mathcal{G}' \in \mathcal{E}$. □

The second result needed for the proof of the "only if"-statement of Proposition 7 demonstrates that for each node i in a perfect DAG $\mathcal{R}^*$ there exists a DAG $\mathcal{G} \in \mathcal{E}$ in which there is no non-fundamental link that points to i.

Lemma 4. *For all nodes $i \in N^*$ there exists a DAG $\mathcal{G} \in \mathcal{E}$ in which all non-fundamental links adjacent to node i point away from i.*

Proof. Let N_d be the set of nodes that have distance $d > 0$ to the action node 0. Denote by $N_d^{[\kappa]}$, $\kappa = 1, 2, ...$, the maximal subset of nodes that (i) are at distance $d > 0$ from the action node 0, and (ii) are connected through non-fundamental links (i.e., for any two nodes $i, j \in N_d^{[\kappa]}$ there exists a path between i and j consisting of non-fundamental links). **Step 1.** We show that all nodes in a given set $N_d^{[\kappa]}$ have the same parents outside of $N_d^{[\kappa]}$. Consider two nodes $i, j \in N_d^{[\kappa]}$ that are connected through the non-fundamental link iR^*j. By definition, we have kEi for each $k \in R^*(i) \setminus N_d^{[\kappa]}$ for each $i \in N_d^{[\kappa]}$. Since $\mathcal{R}^*$ is perfect, this implies that $R^*(j) \setminus N_d^{[\kappa]} \subset R^*(i) \setminus N_d^{[\kappa]}$. Since iR^*j is non-fundamental, we also must have $R^*(i) \setminus N_d^{[\kappa]} \subset R^*(j) \setminus N_d^{[\kappa]}$ so that $R^*(i) \setminus N_d^{[\kappa]} = R^*(j) \setminus N_d^{[\kappa]}$. The result follows from the fact that, by assumption, all nodes in $N_d^{[\kappa]}$ are connected through non-fundamental links. **Step 2.** Consider two links $i \in N_d^{[\kappa]}$ and $i' \in N_d^{[\kappa']}$ with $\kappa \neq \kappa'$ that are adjacent. Assume w.l.o.g. that iR^*i'. By definition, iR^*i' is a fundamental link. Step 1 then implies that iEj' for all $j' \in N_d^{[\kappa']}$. Thus, there cannot exist nodes $j \in N_d^{[\kappa]}$ and $j' \in N_d^{[\kappa']}$ so that $j'R^*j$. Otherwise, we would have $j'Ej$ and $j'Ei$ for all $i \in N_d^{[\kappa]}$, a contradiction. Thus, there cannot exist nodes $i, j \in N_d^{[\kappa]}$ and $i', j' \in N_d^{[\kappa']}$ such that iR^*i' and $j'R^*j$. **Step 3.** Note that, since $\mathcal{R}^*$ is perfect, by Lemma 1 all links between N_d and N_{d+1} point away from the nodes in N_d. **Step 4.** We now can prove Lemma 4. Take any node $i \in N^*$ and assume w.l.o.g. that $i \in N_d^{[\kappa]}$. Consider the DAG $\mathcal{G}^{[\kappa]} = (N_d^{[\kappa]}, G^{[\kappa]})$ where $G^{[\kappa]}$ is identical to R^* restricted on $N_d^{[\kappa]}$. Since $\mathcal{R}^*$ is perfect, $\mathcal{G}^{[\kappa]}$ also must be perfect. Corollary 1 from Spiegler (2019) implies that there exists a DAG $\mathcal{Q}^{[\kappa]}$ in which node i is ancestral and that is equivalent to $\mathcal{G}^{[\kappa]}$. Choose such a $\mathcal{Q}^{[\kappa]}$ and replace $\mathcal{G}^{[\kappa]}$ in the original DAG $\mathcal{R}^*$ by $\mathcal{Q}^{[\kappa]}$. Call the resulting DAG $\mathcal{Q}^*$. Step 1 implies that there are no v-colliders in $\mathcal{Q}^*$, and Step 2 and 3 imply that there are no cycles in $\mathcal{Q}^*$, which proves the result. □

Proof of Proposition 7. The "if"-statement follows from Lemma 1 and Lemma 2. We prove the "only if"-statement. Consider any two adjacent nodes $i, j \in N^*$ with iR^*j and $d(0, i) =$

$d(0, j)$. Suppose that for any node $k \in R^*(i)$ with a fundamental link kR^*i we also have $k \in R^*(j)$. By Lemma 4, we can find a DAG $\mathcal{G} \in \mathcal{E}$ in which all non-fundamental links are turned away from node i. In this DAG, we have $G(i) \subset G(j)$. From Lemma 3 it then follows that the link iR^*j is not fundamental. This completes the proof. □

Before we can prove Proposition 6, we need two more results. We will use the following definitions. Recall that a path τ of length d is directed if for any $h \in \{1, ..., d\}$ we have $\tau_{h-1}R\tau_h$ on this path. For any DAG, the topological ordering is a sequence of nodes such that every link is directed from an earlier to a later node in the sequence.

Lemma 5. *Let $M \subset N^* \setminus H^*(\mathcal{R}^*)$ be a set of nodes connected through non-fundamental links. Suppose there are two nodes $i, j \in H^*(\mathcal{R}^*)$ with non-fundamental links to nodes in M. Then i and j are adjacent.*

Proof. As in the proof of Lemma 4, let N_d be the set of nodes that have distance $d > 0$ to the action node 0. Let $E(i)$ be the set of nodes k with kEi. By Lemma 1, there is a $d > 0$ so that $i, j \in N_d$ and $M \subset N_d$. By Lemma 2, we must have $E(i) = E(j)$ since these nodes are connected through non-fundamental links. Choose any node $k \in N_{d-1}$ with $k \in H^*(\mathcal{R}^*)$ and kR^*i. By Lemma 2, we also have kR^*j. We can now choose two fundamental active paths $\tau^{[i]}, \tau^{[j]}$ from node 0 to node n so that (i) $k \in \tau^{[i]}$ and $k \in \tau^{[j]}$, (ii) $i \in \tau^{[i]}$ and $j \in \tau^{[j]}$, (iii) all nodes on $\tau^{[i]}$ and $\tau^{[j]}$ before k are identical, and (iv) there is not any node on $\tau^{[i]}$ ($\tau^{[j]}$) between k and i (k and j). Since $i, j \in H^*(\mathcal{R}^*)$ this is possible. Now define by $m_1^{[i]}$ ($m_1^{[j]}$) the last node on $\tau^{[i]}$ ($\tau^{[j]}$) before node n; by $m_2^{[i]}$ ($m_2^{[j]}$) the penultimate node on $\tau^{[i]}$ ($\tau^{[j]}$) before node n, and so forth. Since $\mathcal{R}^*$ is perfect, $m_1^{[i]}$ and $m_1^{[j]}$ must be adjacent. Since $m_1^{[i]}$ and $m_1^{[j]}$ are adjacent and $\mathcal{R}^*$ is perfect, $m_2^{[i]}$ and $m_2^{[j]}$ must be adjacent, and so forth. If nodes i and j are both the t'th node from n in $\tau^{[i]}$ ($\tau^{[j]}$), we are done. Assume that this is not the case, and that w.l.o.g. node i is the t'th node from n while node j is the t''th node from n, with $t' > t$. Then i is adjacent to $m_t^{[j]}$, and also to all nodes on $\tau^{[j]}$ between $m_t^{[j]}$ and j (including j) through non-fundamental links, otherwise there would be a contradiction to $E(i) = E(j)$. □

The next result is crucial for the proof of Proposition 6. It shows that all nodes that are not on a fundamental active path between action and output can be made "unimportant", in the sense that we can find a DAG in $\mathcal{E}$ in which any link between a node in $H^*(\mathcal{R}^*)$ and a node in $N^* \setminus H^*(\mathcal{R}^*)$ points towards the node in $N^* \setminus H^*(\mathcal{R}^*)$.

Lemma 6. *There exists a DAG $\mathcal{G}^* \in \mathcal{E}$ such that in $\mathcal{G}^*$ all links with one end in $H^*(\mathcal{R}^*)$ and the other in $N^* \setminus H^*(\mathcal{R}^*)$ point from $H^*(\mathcal{R}^*)$ to $N^* \setminus H^*(\mathcal{R}^*)$.*

Proof. The proof proceeds by steps. **Step 1.** Consider any maximal set $M \subset N^* \setminus H^*(\mathcal{R}^*)$ of nodes connected through non-fundamental links and let $M^+ \subset H^*(\mathcal{R}^*)$ be the set of nodes that have non-fundamental links to nodes in M. By Lemma 1, there is a $d > 0$ so that $M, M^+ \subset N_d$. Denote by M^{++} the set of nodes in $N_d \cap H^*(\mathcal{R}^*)$ with fundamental links into M. Since the nodes in M are connected through non-fundamental links, there is a fundamental link from any node $i \in M^{++}$ to any node in M. Thus, any node in M^{++} must also be adjacent to any node in M^+, so $M^+ \cup M^{++}$ is a clique. **Step 2.** Consider the DAG $\bar{\mathcal{G}} = (N, \bar{G})$, where $N = M \cup M^+ \cup M^{++}$ and $\bar{G}$ is identical to R^* restricted on N. By construction, this DAG is perfect. Hence, Corollary 1 from Spiegler (2019) implies that there exists a DAG $\bar{\mathcal{G}}^+$ in which the clique $M^+ \cup M^{++}$ is ancestral and that is equivalent to $\bar{\mathcal{G}}$. We choose such a $\bar{\mathcal{G}}^+$ with the property that the ordering of the nodes in $M^+ \cup M^{++}$ is the same as in $\bar{\mathcal{G}}$ (this is possible since $M^+ \cup M^{++}$ is a clique, and all links between nodes $M^+ \cup M^{++}$ and nodes in M point towards the latter one). Consider now the DAG $\mathcal{G}$ that is identical to $\mathcal{R}^*$ except that $\bar{\mathcal{G}}$ is replaced by $\bar{\mathcal{G}}^+$. We show that there are no cycles or v-colliders in $\mathcal{G}$ so that it is equivalent to $\mathcal{R}^*$. Consider any node $i \in N_{d-1} \cup N_d$ that is outside $M \cup M^+ \cup M^{++}$ and that has a fundamental link into a node in M. Since the nodes in M are connected through non-fundamental links, node i has a fundamental link into every node in M (otherwise, i would belong to M, a contradiction). This rules out v-colliders. Any link between a node in N_d and a node in N_{d+1} points into the latter one. Hence, by construction, there cannot be cycles or v-colliders in $\mathcal{G}$. We obtain $\mathcal{G}^*$ by performing the same changes for any maximal set $M \subset N^* \setminus H^*(\mathcal{R}^*)$ of nodes connected by non-fundamental links in $\mathcal{R}^*$. □

Proof of Proposition 6. First, we show the "if"-statement. Assume that the agent's subjective model $\mathcal{R}$ contains all the nodes in $H^*(\mathcal{R}^*)$. Consider the DAG $\mathcal{G}^* \in \mathcal{E}$ in which all links with one end in $H^*(\mathcal{R}^*)$ and the other in $N^* \setminus H^*(\mathcal{R}^*)$ point from $H^*(\mathcal{R}^*)$ to $N^* \setminus H^*(\mathcal{R}^*)$. By Lemma 6, this DAG exists. From Proposition 5 it follows that $p_{\mathcal{G}^*}(x_{H^*(\mathcal{R}^*)}) = p(x_{H^*(\mathcal{R}^*)})$ for all distributions $p(x) \in \Delta(X)$. Consider the subgraph $\mathcal{G} = (G, N)$ where G equals G^* restricted on N. Since none of the nodes in $N \setminus H^*(\mathcal{R}^*)$ impacts on any node in $H^*(\mathcal{R}^*)$, we have $p_{\mathcal{G}}(x_{H^*(\mathcal{R}^*)}) = p_{\mathcal{G}^*}(x_{H^*(\mathcal{R}^*)})$ for all $p(x) \in \Delta(X)$. By construction, the DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent so that we have $p_{\mathcal{R}}(x_{H^*(\mathcal{R}^*)}) = p_{\mathcal{G}}(x_{H^*(\mathcal{R}^*)}) = p_{\mathcal{G}^*}(x_{H^*(\mathcal{R}^*)}) = p(x_{H^*(\mathcal{R}^*)})$ for all distributions $p(x) \in \Delta(X)$, which proves the "if"-statement. Next, we show the "only if"-statement. Assume that there is one node $i \in H^*(\mathcal{R}^*)$ that is not in the agent's subjective model. This node is on a fundamental active path τ between the action node 0 and the output node n. We then can find a probability distribution $p(x) \in \Delta(X)$ so that $p_{\mathcal{R}}(x_n \mid x_0) \neq p(x_n \mid x_0)$. Let k be the k'th node in τ. Consider a probability distribution with the following properties: $p(x_j \mid x_{R^*(j)}) = p(x_j)$ for all nodes $j \notin \tau$ that are between the nodes 0 and n, and $p(x_k \mid x_{R^*(k)}) = p(x_k \mid x_{k-1})$. Clearly, such a distribution can have the desired property. □

Proof of Corollary 3. Denote $H^*(\mathcal{R}_1) = H^*(\mathcal{R}_2) = H$. By Proposition 6, there exists a DAG $\mathcal{R}_1^{[1]}$ that is equivalent to $\mathcal{R}_1$ and in which all links between any node $i \in H$ and any node $j \in N_1 \setminus H$ is turned away from i. Thus, we have

$$p_{\mathcal{R}_1}(x_H) = \sum_{x_{N_1 \setminus H} \in X_{N_1 \setminus H}} p_{\mathcal{R}_1}(x_{N_1}) = \sum_{x_{N_1 \setminus H} \in X_{N_1 \setminus H}} p_{\mathcal{R}_1^{[1]}}(x_{N_1}) = p_{\mathcal{R}_1^{[1]}}(x_H). \tag{A.13}$$

Note that for all $i \in H$ we have that $R_1^{[1]}(i) \subset H$. Consider the restriction of $R_1^{[1]}$ on H, $R_1^{[H]}$. We then have

$$p_{\mathcal{R}_1^{[1]}}(x_H) = \prod_{i \in H} p(x_i \mid x_{R_1^{[1]}(i)}) = \prod_{i \in H} p(x_i \mid x_{R_1^{[H]}(i)}) = p_{\mathcal{R}_1^{[H]}}(x_H). \tag{A.14}$$

Define $\mathcal{R}_2^{[1]}$ and $\mathcal{R}_2^{[H]}$ just like $\mathcal{R}_1^{[1]}$ and $\mathcal{R}_1^{[H]}$. By assumption, the link $iR_1^{[H]}j$ is in $R_1^{[H]}$ if and only if we have $iR_2^{[H]}j$ or $jR_2^{[H]}i$. Thus, $\mathcal{R}_1^{[H]}$ and $\mathcal{R}_2^{[H]}$ have the same skeleton. Since $\mathcal{R}_1$ and $\mathcal{R}_2$ are perfect, so are $\mathcal{R}_1^{[H]}$ and $\mathcal{R}_2^{[H]}$. Hence $\mathcal{R}_1^{[H]}$ and $\mathcal{R}_2^{[H]}$ are equivalent, so that

$$p_{\mathcal{R}_1^{[H]}}(x_H) = p_{\mathcal{R}_2^{[H]}}(x_H). \tag{A.15}$$

From the equations (A.13) to (A.15), we get $p_{\mathcal{R}_1}(x_H) = p_{\mathcal{R}_2}(x_H)$, which implies the result. □

A.5 Risk and Incentives

To study the relationship between risk and incentives, the literature typically uses a setting with continuous actions, normally distributed output and exponential utility (so that the optimal contract is linear). To properly apply our framework, we consider a setting with discrete actions and outputs that captures the negative relationship between risk and incentives.

Let there be a binary action $a \in \{0, 1\}$ and three equidistant output levels, y_L, y_M, y_H with $y_H > y_M > y_L > 0$. The level of risk is indexed by a parameter $\xi \in [0, \bar{\xi}]$. The production function is $p(y_L \mid a) = \beta_L(\xi) - \beta a$, $p(y_M \mid a) = \beta_M(\xi)$, and $p(y_H \mid a) = \beta_H(\xi) + \beta a$, where $\beta_L(\xi) = \beta_H(\xi)$ for all ξ. An increase in risk ξ shifts probability mass from the medium output y_M to the extreme outputs y_L and y_H, i.e., $\beta'_L(\xi) = \beta'_H(\xi) = \varepsilon$ for some $\varepsilon > 0$ and $\beta'_M(\xi) = -2\varepsilon$. The agent has a piecewise linear utility function $u(w) = w$ for $w \geq 0$, and $u(w) = \lambda w$ with $\lambda > 1$ for $w < 0$; her reservation utility is $\bar{U} = 0$.

We now fit the marketer example from Subsection 3.2 to the present setting. The objective causal model is given by $\mathcal{R}^*$ on the left of Figure 2, while the agent's subjective model is given by $\mathcal{R}$ on the right of this figure. We use our usual parametrization, except for the output. The probability of low, middle, and high output conditional on x_1 and x_2 is given by

$$p(y_H \mid x_1, x_2) = \beta_3^H(\xi) + \beta_{13}(\xi)x_1 + \beta_{23}(\xi)x_2, \tag{A.16}$$

$$p(y_M \mid x_1, x_2) = \beta_3^M(\xi), \tag{A.17}$$

$$p(y_L \mid x_1, x_2) = \beta_3^L(\xi) - \beta_{13}(\xi)x_1 - \beta_{23}(\xi)x_2. \tag{A.18}$$

We assume that the level of risk ξ changes the importance of consumer information and reputation for the final output. The larger the risk, the more important are these two factors to obtain a high rather than a small output. We capture this by assuming

$$\beta_{13}(\xi) = \bar{\beta}_{13}\left(1 + \frac{\xi}{\beta_{01}\bar{\beta}_{13}}\right) \text{ and } \beta_{23}(\xi) = \bar{\beta}_{23}\left(1 + \frac{\xi}{\mid \beta_{02} \mid \bar{\beta}_{23}}\right) \tag{A.19}$$

for two values $\bar{\beta}_{13}, \bar{\beta}_{23} > 0$ with $\beta_{01}\bar{\beta}_{13} + \beta_{02}\bar{\beta}_{23} = \beta$. We choose the functions $\beta_3^H(\xi), \beta_3^M(\xi)$ and $\beta_3^L(\xi)$ so that the objective probability model generates the production function from above.[13]

Proposition 9 (Risk and Incentives)**.** *Consider the marketer example of this subsection.*

(a) Suppose the agent's subjective model equals $\mathcal{R}^*$. The expected wage payment needed to implement $\alpha = 1$ then increases in risk ξ, and there exists an interval $[c_L, c_H]$ so that if $c \in (c_L, c_H)$, then for some $\xi^* \in (0, \bar{\xi})$ the optimal equilibrium contract implements $\alpha = 1$ if $\xi < \xi^*$ and $\alpha = 0$ if $\xi > \xi^*$.

(b) Suppose the agent's subjective model equals $\mathcal{R}$. The expected wage payment needed to implement $\alpha = 1$ then decreases in risk ξ if the slope $\beta_L'(\xi) = \beta_H'(\xi) = \varepsilon$ is small enough. In this case, there is an interval $[c_L, c_H]$ so that if $c \in (c_L, c_H)$, then for some $\xi^* \in (0, \bar{\xi})$ the optimal equilibrium contract implements $\alpha = 0$ if $\xi < \xi^*$ and $\alpha = 1$ if $\xi > \xi^*$.

Below we provide the proof of Proposition 9. We explain why part (a) holds. When the agent has rational expectations, the *IC* that ensures high effort equals

$$\beta(u(w_H) - u(w_L)) \geq c, \tag{A.20}$$

and the optimal wage schedule that implements high effort is given by

$$w(y_L) = -\frac{1}{2\lambda\beta}c, \; w(y_M) = 0, \text{ and } w(y_H) = \frac{1}{2\beta}c. \tag{A.21}$$

[13] Specifically, we derive $\beta_3^H(\xi)$ and $\beta_3^L(\xi)$ from $\beta_H(\xi) = \beta_3^H(\xi) + \beta_1\beta_{13}(\xi) + \beta_2\beta_{23}(\xi)$ and $\beta_L(\xi) = \beta_3^L(\xi) - \beta_1\beta_{13}(\xi) - \beta_2\beta_{23}(\xi)$. Since $\beta_H(\xi) = \beta_L(\xi)$ for all ξ, we have $\beta_3^M(\xi) = 1 - 2[\beta_3^H(\xi) + \beta_1\beta_{13}(\xi) + \beta_2\beta_{23}(\xi)]$.

Note that a change in risk ξ affects neither the optimal wage schedule, nor the incentive compatibility constraint in (A.20). In terms of effort incentives, the effect of risk on the importance of consumer information and reputation cancel each other out. However, an increase in risk exposes the agent to more variation in her wage, so that she requires a higher risk-premium. Hence, when the principal implements high effort, his expected payment to the agent under the optimal contract increases in risk. Therefore, there exists an interval of cost levels $[c_L, c_H]$, so that if $c \in (c_L, c_H)$, the optimal equilibrium contract implements high effort if and only if the level of risk is sufficiently small. We thus obtain a negative relationship between risk and incentives.

Next, consider part (b). If the agent does not take the reputation channel into account, an increase in risk appears to her as an increase in the productivity of her effort, as the association between consumer information and sales becomes stronger. The IC that ensures high effort now equals

$$\beta_{01}\beta_{13}(\xi)(u(w_H) - u(w_L)) \geq c. \tag{A.22}$$

Recall that $\beta_{13}(\xi)$ increases in ξ. Hence, an increase in risk ξ relaxes this IC. The optimal wage schedule that implements $\alpha = 1$ is now given by

$$w(y_L) = -\frac{\beta_H(\xi) + \beta - \beta_{01}\beta_{13}(\xi)}{\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c, \ w(y_M) = 0, \text{ and } w(y_H) = \frac{\beta_L(\xi) - \beta + \beta_{01}\beta_{13}(\xi)}{(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c. \tag{A.23}$$

A change in risk now has two countervailing effects on the expected payment when the principal implements high effort. It again increases the risk premium that the agent requires, but it also relaxes the incentive compatibility constraint. Which effect dominates depends on the probability model and the utility function. If the slope $\beta_L'(\xi) = \beta_H'(\xi) = \varepsilon$ is small enough, an increase in risk reduces the expected payment to the agent at all risk levels $\xi \in [0, \bar{\xi}]$. We then obtain a positive relationship between risk and incentives: For an interval of cost levels $[c_L, c_H]$, if $c \in (c_L, c_H)$, the optimal equilibrium contract implements high effort if the level of risk is sufficiently large, and otherwise low effort through a fixed wage.

Proof of Proposition 9. We first prove statement (a). For this, we derive the optimal contract under the objective model $\mathcal{R}^*$ that implements high effort. For convenience, we abbreviate $w_H = w(y_H)$, $w_M = w(y_M)$, and $w_L = w(y_L)$. Standard arguments show that both IC and PC must be binding at the optimal contract, and that $w_L < 0$ and $w_H > 0$ at the optimum. Assume for the moment that $w_M \geq 0$ under the optimal contract. The IC is then

$$\beta(w_H - \lambda w_L) = c, \tag{A.24}$$

and the *PC* equals

$$(\beta_H(\xi) + \beta)w_H + \beta_M(\xi)w_M + (\beta_L(\xi) - \beta)\lambda w_L = 0. \tag{A.25}$$

From the *IC* we get

$$w_H = \frac{c}{\beta} + \lambda w_L, \tag{A.26}$$

We plug this into the *PC*, solve for w_M, and get

$$w_M = -\frac{\beta_H(\xi)}{\beta_M(\xi)\beta}c - \frac{\beta_L(\xi) + \beta_H(\xi)}{\beta_M(\xi)}\lambda w_L. \tag{A.27}$$

The expected wage payment of the principal when he implements $\alpha = 1$ equals

$$\mathbb{E}[w \mid \alpha = 1] = (\beta_H(\xi) + \beta)w_H + \beta_M(\xi)w_M + (\beta_L(\xi) - \beta)w_L. \tag{A.28}$$

Using the results from above, we can write the expected wage payment as

$$\mathbb{E}[w \mid \alpha = 1] = c - (\beta_L(\xi) - \beta)(\lambda - 1)w_L. \tag{A.29}$$

The optimal wage w_L minimizes this term subject to the constraint that w_M in (A.27) remains weakly positive. The solution implies that $w_M = 0$, and $w(y_L) = -\frac{1}{2\lambda\beta}c$ as well as $w(y_H) = \frac{1}{2\beta}c$. We obtain the same result when we go through the same steps while assuming $w_M \leq 0$. With this we can compose the expected wage payment $\mathbb{E}[w \mid \alpha = 1]$ and obtain

$$\frac{\partial\mathbb{E}[w \mid \alpha = 1]}{\partial\xi} = \frac{\varepsilon}{2\beta}c - \frac{\varepsilon}{2\lambda\beta}c > 0. \tag{A.30}$$

Hence, the expected wage payment to implement $\alpha = 1$ strictly increases in risk. The expected wage payment to implement $\alpha = 0$ is zero for all risk levels. This yields us statement (a).

Next, we prove statement (b). We first derive the agent's beliefs about the production function at $\alpha = 1$. As in the proof of Proposition 2, we find $p(x_2 = 1 \mid x_1 = 1)$ and $p(x_2 = 1 \mid x_1 = 0)$. At $\alpha = 1$, we have $p(x_2 = 1 \mid x_1 = 1) = p(x_2 = 1 \mid x_1 = 0) = \beta_2 + \beta_{02}$, and thus

$$\begin{aligned}
p(y_H \mid x_1 = 1) &= \beta_3^H(\xi) + \beta_{13}(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.31)}\\
p(y_H \mid x_1 = 0) &= \beta_3^H(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.32)}\\
p(y_M \mid x_1 = 1) &= \beta_3^M(\xi), && \text{(A.33)}\\
p(y_M \mid x_1 = 0) &= \beta_3^M(\xi), && \text{(A.34)}\\
p(y_L \mid x_1 = 1) &= \beta_3^L(\xi) - \beta_{13}(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.35)}\\
p(y_L \mid x_1 = 0) &= \beta_3^L(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi). && \text{(A.36)}
\end{aligned}$$

From this, we can derive the agent's beliefs about the production function at $\alpha = 1$ as

$$
\begin{aligned}
p_{\mathcal{R}}(y_H \mid a = 1; \alpha = 1) &= \beta_3^H(\xi) + (\beta_1 + \beta_{01})\beta_{13}(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.37)}\\
p_{\mathcal{R}}(y_M \mid a = 1; \alpha = 1) &= \beta_3^M(\xi), && \text{(A.38)}\\
p_{\mathcal{R}}(y_L \mid a = 1; \alpha = 1) &= \beta_3^L(\xi) - (\beta_1 + \beta_{01})\beta_{13}(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.39)}
\end{aligned}
$$

and

$$
\begin{aligned}
p_{\mathcal{R}}(y_H \mid a = 0; \alpha = 1) &= \beta_3^H(\xi) + \beta_1\beta_{13}(\xi) + (\beta_2 + \beta_{02})\beta_{23}(\xi), && \text{(A.40)}\\
p_{\mathcal{R}}(y_M \mid a = 0; \alpha = 1) &= \beta_3^M(\xi), && \text{(A.41)}\\
p_{\mathcal{R}}(y_L \mid a = 0; \alpha = 1) &= \beta_3^L(\xi) - \beta_1\beta_{13}(\xi) - (\beta_2 + \beta_{02})\beta_{23}(\xi). && \text{(A.42)}
\end{aligned}
$$

At $\alpha = 1$, the *IC* is therefore given by

$$\beta_{01}\beta_{13}(\xi)(u(w_H) - u(w_L)) \geq c. \tag{A.43}$$

The rest of the proof proceeds as in the proof of statement (a). We derive the equilibrium contract that implements $\alpha = 1$ at lowest cost to the principal when the agent's subjective model is given by $\mathcal{R}$. Assume that we have $w_M \geq 0$ at this contract. From the *IC*, we get

$$w_H = \frac{c}{\beta_{01}\beta_{13}(\xi)} + \lambda w_L, \tag{A.44}$$

and from the *PC* we get that

$$w_M = -\frac{\beta_H(\xi) + \beta - \beta_{01}\beta_{13}(\xi)}{\beta_M(\xi)\beta_{01}\beta_{13}(\xi)} - \frac{\beta_L(\xi) + \beta_H(\xi)}{\beta_M(\xi)}\lambda w_L. \tag{A.45}$$

With this, we can calculate the expected wage payment under the optimal equilibrium contract that implements $\alpha = 1$ as

$$\mathbb{E}[w \mid a = 1; \mathcal{R}] = c - (\beta_L(\xi) - \beta)(\lambda - 1)w_L. \tag{A.46}$$

The optimal wage w_L minimizes this term subject to the constraint that w_M in (A.45) remains weakly positive. The solution implies that $w_M = 0$ as well as

$$w_L = -\frac{\beta_H(\xi) + \beta - \beta_{01}\beta_{13}(\xi)}{\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c \text{ and } w_H = \frac{\beta_L(\xi) - \beta + \beta_{01}\beta_{13}(\xi)}{(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}c. \tag{A.47}$$

We obtain the same result when we go through the same steps while assuming $w_M \leq 0$. We then can compose the expected wage payment at the optimal equilibrium contract that implements

$\alpha = 1$ as

$$\mathbb{E}[w \mid a = 1; \mathcal{R}] = \frac{(\lambda - 1)(\beta_H(\xi) + \beta)(\beta_L(\xi) - \beta) + (\lambda + 1)\beta_{01}\beta_{13}(\xi)}{\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)}. \tag{A.48}$$

We differentiate this expression with respect to risk ξ and find

$$\lim_{\varepsilon \to 0} \frac{\partial \mathbb{E}[w \mid a = 1; \mathcal{R}]}{\partial \xi} = -\frac{\lambda(\lambda - 1)(\beta_H(\xi) + \beta_L(\xi))(\beta_H(\xi) + \beta)(\beta_L(\xi) - \beta)}{[\lambda(\beta_H(\xi) + \beta_L(\xi))\beta_{01}\beta_{13}(\xi)]^2} < 0. \tag{A.49}$$

Hence, if ε is sufficiently small, the expected wage payment needed to implement $\alpha = 1$ decreases in risk ξ. The rest of the proof of statement (b) proceeds in the same way as for statement (a). □

A.6 Team Size and Incentives

Team incentives and optimal team size. We consider a simple team setting in which the principal chooses both incentives and the size of the team. Let there be m identical agents who can choose between high and low effort $a \in \{0, 1\}$. We suppress notation for individual agents. For convenience, we assume that agents are risk-neutral and protected by limited liability, so that $w(y) \geq \bar{w} > 0$ for all $y \in Y$. The project output is either large ($y = y_H$) or small ($y = y_L$). The team size m scales these payoffs. We have $y_H = m^\theta \bar{y}_H$ and $y_L = m^\theta \bar{y}_L$, for some $\theta \in (0, 1)$, and normalize $\bar{y}_L = 0$. If the share k of agents exerts high effort, the probability of high output is $kB + D$, where B, D are positive constants with $B + D < 1$. Thus, as the team becomes large, the relative influence of a single agent on the output becomes small. The cost of high effort for the individual agent is c and the cost of low effort is 0.

The principal chooses both team size m and agents' incentives $w(y)$. If he wishes to implement high effort from m agents, the optimal wage scheme is a bonus scheme with $w(y_H) = \bar{w} + \frac{mc}{B}$ and $w(y_L) = \bar{w}$. The principal's profit is then given by

$$(B + D)\left(m^\theta \bar{y}_H - \frac{m^2 c}{B}\right) - m\bar{w}. \tag{A.50}$$

Observe that $w(y_H)$ converges to infinity for $m \to \infty$. As the team size increases, it becomes prohibitively costly to provide effort incentives, as an individual agent's influence on the output becomes small. If the principal wishes to implement low effort from m agents, the optimal wage scheme is a fixed-wage $w(y_H) = w(y_L) = \bar{w}$ and the corresponding profit is $Dm^\theta \bar{y}_H - m\bar{w}$. Denote by $m^{[a]}(c)$ the optimal team size if the principal implements action $a \in \{0, 1\}$ and effort costs are given by c. This value is uniquely defined. We then get the following result: There is a $c^* > 0$ such that the principal optimally implements high effort with team size $m^{[1]}(c)$ if $c \leq c^*$, and low effort with team size $m^{[0]}(c)$ if $c > c^*$.

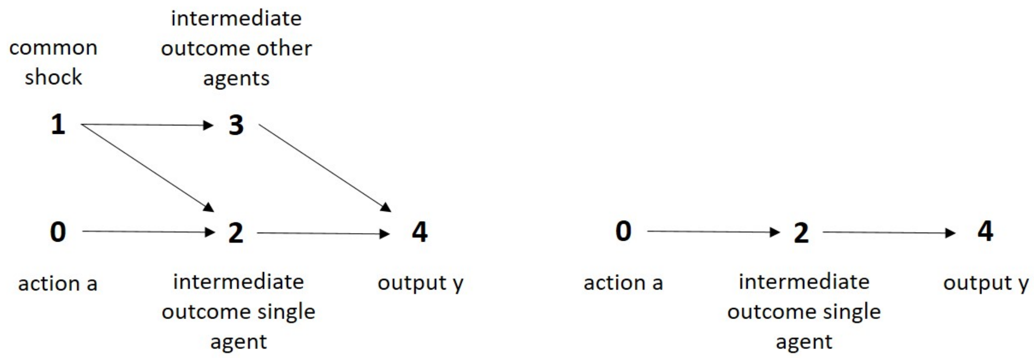


Figure 8: The objective model $\mathcal{R}^*$ (right) and the agent's subjective model $\mathcal{R}$ (left) in the team size example.

Team incentives and optimal team size with misspecified model. We now consider an extended production function that is consistent with the production function indicated above, and that allows us to study how team size and incentives change when agents do not take their colleagues' effort into account. Consider the objective model $\mathcal{R}^*$ on the left of Figure 8. Node 0 is the effort of a single agent. We use the fact that all agents are symmetric, and assume that each of the other $m-1$ agents exerts high effort with probability α^o. Through her effort, the single agent affects an intermediate outcome $x_2 \in \{0,1\}$. Denote by $x_3 \in \{0,1\}^{m-1}$ the $m-1$-dimensional vector of intermediate outcomes of all other agents. The probability of high output increases linearly in the number of high intermediate outcomes, $p(y_H \mid x_2, x_3) = \beta_{24}x_2 + \beta_{24} \parallel x_3 \parallel$, where $\parallel . \parallel$ is the sum of entries in a vector. There is a common shock $x_1 \in \{0,1\}$ that occurs with probability $p(x_1 = 1) = \beta_1$. It positively affects each agent's intermediate outcome, $p(x_2 = 1 \mid a, x_1) = \beta_{02}a + \beta_{12}x_1$, where $\beta_{02} + \beta_{12} < 1$ and $\beta_1\beta_{12} \geq \frac{1}{2}$; for any other agent, the probability of a high intermediate outcome is $\beta_{02}\alpha^o$ if $x_1 = 0$ and $\beta_{02}\alpha^o + \beta_{12}$ if $x_1 = 1$. We define $B \equiv \beta_{02}\beta_{24}$ with $\beta_{24} = \frac{\bar{\beta}_{24}}{m}$ for some $\bar{\beta}_{24}$, and $D \equiv \beta_1\beta_{12}\bar{\beta}_{24}$. The production function is then the same as above; optimal team size and incentives would remain unchanged if the agents' subjective model would be given by $\mathcal{R}^*$. We assume now that agents ignore the contributions of others. Let an agent's subjective model be given by $\mathcal{R}$ on the right of Figure 8. We then obtain the following result.

Proposition 10 (Team Size and Incentives)**.** *Consider the team size example of this section.*

(a) Suppose the agent's subjective model equals $\mathcal{R}^*$. Then there is a unique value $c^* > 0$ such that the principal optimally implements high effort with team size $m^{[1]}(c)$ if $c \leq c^*$, and low effort with team size $m^{[0]}(c)$ if $c > c^*$.

(b) Suppose the agent's subjective model equals $\mathcal{R}$. Then there is a unique value $c^{**} > c^*$ such that the principal optimally implements high effort with team size $m^{[1]}_{\mathcal{R}}(c) > m^{[1]}(c)$ if $c \leq c^{**}$, and low effort with team size $m^{[0]}(c)$ if $c > c^{**}$.

Thus, if the agent's subjective model is misspecified and effort costs are small enough, the principal chooses a team size that is "too large" for tying the agents' pay to the output. It then appears as if incentives are provided to too many employees. However, the simplification in the agents' subjective model causes them to overestimate the importance of their effort for the final output, so that granting these incentives remains profitable for the principal.

Before we prove Proposition 10, we explain its intuition. Suppose that the probabilities of high effort α and α^o are given. According to model $\mathcal{R}$, the agent's belief about how her intermediate outcome affects the final output then equals

$$p(y_H \mid x_2 = 1) - p(y_H \mid x_2 = 0) = \beta_{24}[1 + \xi(\alpha, \beta_1, \beta_{02}, \beta_{12})(m-1)], \tag{A.51}$$

where $\xi(\alpha, \beta_1, \beta_{02}, \beta_{12}) = \frac{\beta_1(1-\beta_1)\beta_{12}^2}{(\beta_{12}+\alpha\beta_{02})(1-\beta_{12}-\alpha\beta_{02})} \in (0,1)$. In contrast, under the objective model, the value in (A.51) would be equal to β_{24} and therefore vanish as the team size m becomes large. Thus, under the subjective model, the agent overestimates the importance of her intermediate outcome for the output. The reason is that a high intermediate outcome indicates a positive common shock, which also increases the chance of high intermediate outcomes for all other agents. Under model $\mathcal{R}$, the agent falsely attributes the corresponding increase in the probability of a high output y_H to the significance of her intermediate outcome x_2. We show below that her perception of the significance of her intermediate outcome decreases in team size, but converges against a positive constant for $m \to \infty$. Thus, the agent maintains a certain belief in the importance of her effort even when her true impact on the final output vanishes.

When $\mathcal{R}$ is the subjective model of all agents, the principal's profit from implementing high effort at team size m with the optimal incentive scheme is given by

$$(B + D)\left(m^\theta \bar{y}_H - \frac{m^2 c}{B} \frac{1}{1 + \xi(1, \beta_1, \beta_{02}, \beta_{12})(m-1)}\right) - m\bar{w}. \tag{A.52}$$

From this we can derive the optimal team size $m_{\mathcal{R}}^{[1]}(c)$ at cost c. The profit from implementing low effort from m agents remains the same as under the objective model. Proposition 10 then follows from a comparison of the profit levels in (A.50) and (A.52).

Proof of Proposition 10. We provide the remaining formal details needed to prove Proposition 10. For this, we fit the agent's subjective model $\mathcal{R}$ to the probability distribution, taking α and α^o as given. First, we calculate the probabilities that there is a common shock, given that

$x_2 = 1$ and $x_2 = 0$, respectively. We get

$$p(x_1 = 1 \mid x_2 = 1) = \frac{\beta_1\beta_{12} + \alpha\beta_1\beta_{02}}{\beta_1\beta_{12} + \alpha\beta_{02}}, \tag{A.53}$$

$$p(x_1 = 1 \mid x_2 = 0) = \frac{\beta_1(1 - \beta_{12} - \alpha\beta_{02})}{1 - \beta_1\beta_{12} - \alpha\beta_{02}}. \tag{A.54}$$

The probability of a high output y_H after a high intermediate outcome $x_2 = 1$ is then given by

$$\begin{aligned} p(y_H \mid x_2 = 1) = {} & \frac{\beta_1\beta_{12} + \alpha\beta_1\beta_{02}}{\beta_1\beta_{12} + \alpha\beta_{02}} \left[1 + \sum_{k=0}^{m-1} \binom{m-1}{k} (\beta_{12} + \alpha\beta_{02})^k (1 - \beta_{12} - \alpha\beta_{02})^{m-1-k} k\right] \beta_{24} \\ & + \left(1 - \frac{\beta_1\beta_{12} + \alpha\beta_1\beta_{02}}{\beta_1\beta_{12} + \alpha\beta_{02}}\right) \left[1 + \sum_{k=0}^{m-1} \binom{m-1}{k} (\alpha\beta_{02})^k (1 - \alpha\beta_{02})^{m-1-k} k\right] \beta_{24}. \end{aligned} \tag{A.55}$$

Using $\binom{m}{k} p^k (1-p)^{m-k} k = mp$ we get

$$p(y_H \mid x_2 = 1) = \beta_{24}(1 + \alpha^o\beta_{02}(m-1)) + \frac{\beta_1\beta_{12} + \alpha\beta_1\beta_{02}}{\beta_1\beta_{12} + \alpha\beta_{02}} \beta_{24}\beta_{12}(m-1). \tag{A.56}$$

Similarly, we get

$$p(y_H \mid x_2 = 0) = \beta_{24}\alpha^o\beta_{02}(m-1) + \frac{\beta_1(1 - \beta_{12} - \alpha\beta_{02})}{1 - \beta_1\beta_{12} - \alpha\beta_{02}} \beta_{24}\beta_{12}(m-1). \tag{A.57}$$

From equations (A.56) and (A.57) we can then derive $p(y_H \mid x_2 = 1) - p(y_H \mid x_2 = 0)$ and the incentive compatibility constraint. From this IC we can derive that if the principal wishes to implement high effort from m agents, then the optimal incentive scheme is

$$w(y_H) = \bar{w} + \frac{cm}{B[1 + \xi(1, \beta_1, \beta_{02}, \beta_{12})(m-1)]} \quad \text{and} \quad w(y_L) = \bar{w}. \tag{A.58}$$

From this the principal's profit in equation (A.52) follows. Note that $\beta_1\beta_{12} \geq \frac{1}{2}$ implies that $\xi(\alpha, \beta_1, \beta_{02}, \beta_{12})$ is maximal at $\alpha = 1$. Thus, the principal cannot gain by implementing a mixed strategy profile.

Finally, we show that $\xi(\alpha, \beta_1, \beta_{02}, \beta_{12}) < 1$ for all admissible values $\alpha, \beta_1, \beta_{02}, \beta_{12}$. This inequality is identical to $\beta_1\beta_{12}(1 - \beta_{12}) + \alpha\beta_{02}(1 - 2\beta_1\beta_{12} - \alpha\beta_{02}) > 0$. Since $1 > \beta_{02} + \beta_{12}$ and $\alpha \leq 1$, this inequality is implied by $\beta_1\beta_{12}(1 - \beta_{12}) - \beta_1\beta_{12}\beta_{02} = \beta_1\beta_{12}(1 - \beta_{02} - \beta_{12}) > 0$, which implies the statement above. □

References

GEIGER, DAN, THOMAS VERMA, AND JUDEA PEARL (1990): "Identifying independence in Bayesian networks," *Networks*, 20(5), 507–534.

PEARL, JUDEA (2009): *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

SPIEGLER, RAN (2019): "Can agents with causal misperceptions be systematically fooled?," *Journal of the European Economic Association*, forthcoming.