Equilibrium Contracts and Boundedly Rational Expectations*

Heiner Schumacher† Heidi Christina Thysen‡

KU Leuven London School of Economics

Version: February 6, 2019

Abstract

We study a principal-agent framework in which the agent has a misspecified model of the principal's project. She fits this model to the objective probability distribution in order to predict outcomes under alternative actions. Under mild restrictions, the agent has correct beliefs on the equilibrium path, but may incorrectly extrapolate how off-equilibrium actions map into outcomes. The agent's misspecification may relax incentive compatibility or even eliminate the incentive problem. We examine optimal workplace design when the principal has discretion over the agent's subjective model, and obtain new results on technology choice, transparency, narcissistic leadership, and relative performance evaluation.

Keywords: Bayesian Networks, Principal-Agent Relationship, Moral Hazard
JEL Classification: D03, D82, D86

*We gratefully acknowledge financial support by the ERC Advanced Investigator grant no. 692995. We thank Yair Antler, Felix Bierbrauer, Kfir Eliaz, Florian Englmaier, Guido Friebel, Heiko Karle, Ronny Razin, Karl Schlag, Klaus Schmidt, Dirk Sliwka, and Yves Le Yaouanq as well as seminar audiences at Aarhus University, Boston University, University of Cologne, Dalhousie University, London School of Economics, Ludwig-Maximilians University Munich, Tilburg University, the EEA-ESEM 2017 in Lisbon, and the ESSET 2018 Gerzensee for their valuable comments and suggestions. We especially thank Ran Spiegler for his invaluable support of the project. The usual disclaimer applies.

†Corresponding Author. KU Leuven, Department of Economics, Naamsestraat 69, 3000 Leuven, Belgium, ++32 163 74 579, E-mail: heiner.schumacher@kuleuven.be.

‡Department of Economics, London School of Economics, h.c.thysen@lse.ac.uk.

1 Introduction

In an organization, an agent may not understand all details of her complex environment. Consider the following example of a "misspecified model" that an agent may have.

> **"Marketer Example."** The agent is a marketer whose job is to increase sales. One strategy to increase sales is to make cold-calls (calling potential customers without prior consent). This increases the set of customers who know about the firm's product, but also reduces the firm's reputation since some customers are annoyed by being cold-called.[1] Both a larger customer set and a better reputation increase expected sales. However, when choosing her strategy, the marketer does not take into account the negative effect of cold-calls on the firm's reputation. The only mechanism that is on her mind is that making cold-calls enlarges the set of potential customers.

We analyze a principal-agent model in which the agent exhibits such misspecifications. We show that, in an organizational context, misspecifications like this can be fairly robust for two reasons. First, the principal may strictly benefit from the agent's misspecification. In this case, he has no incentive to correct her model, and a strict incentive to hire agents with misspecified models. Second, the beliefs over outcomes generated by the agent's misspecified model can be consistent with the actual equilibrium outcomes. That is, the data collected on the equilibrium path will not create suspicion or invalidate the agent's model. We analyze the optimal contract when the agent's misspecification exhibits these two features, and we study how the principal optimally designs the workplace when he has some discretion over the agent's model. Several new implications for optimal organization emerge from this analysis.

There are by now many well-documented empirical cases where experienced agents choose inferior actions, despite strong incentives for improvement and learning. For example, before the invention of germ theory in the second half of the 19th century, doctors saw no value in washing hands before treating patients, thereby killing many individuals through the transmission of diseases. This did not change even after being confronted with clear statistical evidence for the effectiveness of hygienic measures (Nuland 2004). Bloom et al. (2013) show how managers and owners of large companies do not recognize how the cleanliness of the factory floor or the careful documentation of inventory matter for productivity. Experienced farmers may ignore important input dimension and thus produce off the Pareto frontier (Hanna et al. 2014). Tech companies may greatly overestimate the effectiveness of their online marketing efforts when they neglect the relationship between search clicks and purchase intent (Blake et al.

[1] Alternatively, some savvy consumers may infer from such marketing efforts that the quality of the product must be low; see, for example, Miklós-Thal and Zhang (2013).

2015).[2] In all these cases, decision makers are not inexperienced or overconfident, but they do not pay attention to important aspects of their operation. We demonstrate in this paper that such ignorance – unlike in these cases – can increase organizational performance, so that the principal is inclined to keep it.

To capture the agent's misunderstanding of her environment, we apply Spiegler's (2016, 2017) Bayesian network approach. The principal's project can be described by a number of variables and an objective joint probability distribution that describes the probabilistic relationships between these variables. The agent has a subjective model of how and which variables are related to each other. She fits this model to the objective probability distribution, and uses the calibrated subjective model to predict outcomes under alternative actions. A crucial advantage of the Bayesian network approach to boundedly rational beliefs is that it is non-parametric. The agent's beliefs are directly derived from the objective probability distribution. A parametric assumption (as in the case of overconfidence) is not required. This allows us to study the link between biased beliefs and organizational structure, and (to some degree) to endogenize the agent's misspecification.

To illustrate, consider the marketer example from above. The labels of relevant variables are {*cold calls*, *customer base*, *reputation*, *sales*}, where *cold calls* is the agent's action and *sales* is the contractible variable. In the agent's mind, only the variables {*cold calls*, *customer base*, *sales*} matter. She fully understands how *cold calls* affect the *customer base*. She also understands how, given her equilibrium action (which we define precisely below), *sales* depend on the *customer base*. We will see that on the equilibrium path the agent therefore knows the true distribution over *sales*. However, she does not understand that the relationship between *customer base* and *sales* changes when she deviates to an off-equilibrium action. Suppose that the contract implements "make *cold calls*" as equilibrium action. Since *reputation* is not in the agent's model, she ignores the partial positive effect of a deviation on sales through an increase in reputation, which relaxes the incentive constraint. In short, the principal benefits from the agent's misspecification, and the agent cannot detect her belief bias from the data she collects on the equilibrium path.

We exploit a restriction on the agent's subjective model (introduced in Spiegler 2017) which ensures that the agent correctly predicts the marginal equilibrium distribution over outcomes. Loosely speaking, this restriction is that the agent takes into account the correlation between any two variables in her model that have a joint influence on a third variable of her model

[2]There are many further examples outside organizational economics where experience does not help decision makers. Juveniles may overlook the effectiveness of a polite request in situations of conflict (Heller et al. 2017). Commuters may continuously take suboptimal routes because they do not take into account how quickly a vehicle moves between single stops (Larcom et al. 2019). Social behavior is impacted if individuals believe in the validity of certain social norms even when a majority rejects them privately (Bursztyn et al. 2018).

(so that there is no "neglect of correlation"). Misspecifications that respect this restriction can change the incentive problems in several ways that benefit the principal. A misspecification in the agent's model of the mapping from action to output can relax the incentive constraint, as in the marketer example; a misspecification in her model of the mapping from action to her non-pecuniary cost or benefits can even eliminate the incentive problem altogether.

We derive a general result under what circumstances a misspecification in the agent's model can change her beliefs in a way that is relevant for the optimal contract. This result has two implications. First, the agent may miss out important aspects of the contracting problem (such as the contributions of other workers in a team incentives setting) and still behave as if she were fully rational. Omitting variables from a model therefore does not *per se* create incentive effects, and paying attention to more details of the operation does not necessarily improve decision-making, even when they are crucial for the description of the contracting problem. Second, when the agent does not take into account a variable of the project, other variables may become inconsequential for incentives. Thus, different misspecifications – which may reflect different psychological predispositions – can have identical incentive effects. We illustrate these insights in several applications.

The main implication of our basic model is that the agent's misunderstanding of her environment can increase organizational performance. We address the following question: Which agent misspecifications and organizational features are robust in successful organizations? We consider a number of classic topics in organizational economics, and give the principal – who designs the agent's workplace – some discretion over her subjective model. He can hide or highlight certain empirical regularities when choosing the organizational structure. Note that organization affects the objective probability distribution over all relevant variables. From this distribution the agent's beliefs are derived. The design of the organization therefore not only determines incentives, but also has an indirect effect on effort motivation through the agent's subjective model and beliefs. We find a number of organizational features that increase effort motivation through misspecifications that entail correct expectations on the equilibrium path.

(i) The principal has a preference for production technologies with "extreme" features when he can make their advantages salient to the agent, while hiding their disadvantages. This can induce the choice of inefficient technologies and technological inertia.

(ii) Transparency increases effort motivation when the agent's effort inspires her subordinates (Winter 2010). If the agent does not perfectly understand her subordinates' job, she may perceive her contribution as more "pivotal" than it really is. Transparency thus can have advantages beyond exploiting complementarities in the production function.

(iii) An egocentric agent focuses on her own information and solicits too little tacit knowl-

edge from others. The equilibrium outcome then may imply narcissistic leadership – the agent overestimates her own contribution to the output and shows little empathy for others. The principal may have a preference for hiring such an agent.

(iv) The principal may want to use "camouflaged" relative performance evaluation (RPE) to reduce the agent's exposure to common shocks. To maintain cooperation among employees, he does not make RPE explicit, but uses "subjective" performance evaluation that takes the performance of others' into account.

(v) The principal offers an incomplete contract if the inclusion of more variables updates the agent's subjective model so that the implementation of effort becomes more costly.

We discuss how these results advance previous work in organizational economics and what implications they have for empirical work.

Related Literature. The first comprehensive analysis of organizations with boundedly rational decision makers is Herbert Simon's (1947) *Administrative Behavior*. Herbert Simon proposes that rationality requires the decision maker to know the consequences of all possible options. In an organizational context, this is typically impossible. Thus, administrative behavior must be "boundedly rational." Subsequent work in managerial economics, such as the influential works by March and Simon (1958) and Cyert and March (1963), builds on this premise, usually without formalizing it in a mathematical model. In contrast, the standard models in contract theory assume rational expectations and common priors. In our framework, the agent extrapolates from partial data sets to predict outcomes under alternative actions. We therefore use a framework that formalizes to some extent Herbert Simon's bounded rationality and that can be applied to many models in contract theory and organizational economics.

This paper offers a new approach to the literature on contracting between parties with non-common priors. Several models examine how an agent's bias affects optimal incentives and organization, by assuming overconfidence (Rotemberg and Saloner 2000, Fang and Moscarini 2005, Van den Steen 2005, Gervais and Goldstein 2007, De la Rosa 2011, Gervais et al. 2011, Spinnewijn 2015, Bhaskar and Thomas 2019), endogenously biased beliefs (Bénabou 2013), or unawareness (Filiz-Ozbay 2012, Auster 2013, Von Thadden and Zhao 2012, 2014, Auster and Pavoni 2018). In these models, the agent is "biased on the equilibrium path." She either does not correctly predict the outcomes which result from her equilibrium action (as in the case of overconfidence and biased beliefs) or she is unaware of certain actions or outcomes (as in the case of unawareness). In our framework, the agent is aware of all actions and potential outcomes. She makes correct predictions on the equilibrium path, but the misspecification in her model causes her to incorrectly extrapolate how off-equilibrium actions map into outcomes.

Thus, our approach is applicable to a setting where the agent has substantial experience and data on the consequences of her actions, but misses out important regularities of her environment.

Finally, we also contribute to the literature on Bayesian networks and directed acyclic graphs (DAG), which have been used extensively in the artificial intelligence literature. In biomedical research, DAGs are used to eliminate confounding biases in the estimated treatment effect. Moreover, they are used as visual inspection tool when choosing explanatory variables, see, for example, Shrier and Pratt (2008) and Farzaneh-Far et. al. (2010). In these papers, DAGs are interpreted as a representation of causal relationships. This viewpoint is also promoted by Pearl (2009) who provides a broad introduction to DAGs.[3] In economics, Spiegler (2016, 2017) uses Bayesian networks to model agents with boundedly rational expectations. He shows that DAGs can be used to capture a variety of different inference errors such as reverse causation, coarseness and mis-attribution biases. We build on these insights and apply them to contracting and optimal organization. A number of recent papers apply the Bayesian network models to monetary policy (Spiegler 2018), political economy (Eliaz and Spiegler 2018), and Bayesian persuasion (Eliaz et al. 2018).

The remainder of the paper is organized as follows. Section 2 describes our basic model. In Section 3, we characterize the optimal equilibrium contract and derive several general results. Section 4 advances the model by allowing the principal to "choose" the agent's misspecification. In Section 5, we apply our model to study optimal workplace design. Section 6 concludes. All proofs, mathematical details, and further results can be found in the Online Appendix.

2 The Model

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

Basic Framework. The principal proposes a contract $(\mathcal{R}, w(y), a)$, where $w(y) \in W$ is the agent's wage conditional on the output $y \in Y$, $a \in A$ the action that the principal wishes the agent to choose, and $\mathcal{R}$ the agent's subjective model (which we explain in detail below).[4] Let W be the set of possible incentive schemes, $A \subset \mathbb{R}$ a finite set of actions, $Y \subset \mathbb{R}$ a finite set of outputs, and $C \subset \mathbb{R}$ a finite set of possible (non-monetary) effort costs. 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 $a \in A$. Mixed action profiles

[3] For other general introductions to DAGs see, for example, Cowell et al. (1999) or Koski and Noble (2009).

[4] Below, we allow the principal to choose the agent's subjective model. Thus, $\mathcal{R}$ appears in the contract.

are denoted by $p(a) \in \Delta(A)$. The agent's action stochastically influences the project's output and her costs. Her 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 cost is c, the principal's payoff is $V = y - w(y)$ and the agent's payoff is $U = u(w(y)) - c$.

Causal Structure. We model the causal structure though which the agent's action affects the output y and her costs c. Let $N^* = \{0, ..., n, n+1\}$ be the set of relevant project variables (or nodes). They comprise the agent's action, output, and costs, but may also include other variables like customer base or reputation (as in the marketer example). 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 ($x_0 = a$, $X_0 = A$), node n is the output ($x_n = y$, $X_n = Y$), and node $n+1$ is the agent's personal cost ($x_{n+1} = c$, $X_{n+1} = C$). We use these labels interchangeably. The state is a vector $x^* = (x_0, x_1, ..., x_{n+1})$ and the set of all states is $X^* = \times_{i\in N^*} X_i$. Let x_S be the vector of variables in $S \subset N^*$.

The production- and cost-function is given by $p(x_1, \ldots, x_{n+1} \mid a)$. Together with the agent's action $p(a)$ it generates the objective joint distribution over all variables $p(x^*)$. Output and costs are independent conditional on the agent's action.[5] We can now write down the causal structure of the principal's project by an irreflexive, asymmetric and acyclic binary relation R^* over N^*. We denote it by the DAG $\mathcal{R}^* = (N^*, R^*)$, see the graph on the left in 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 then naturally factorizes according to $\mathcal{R}^*$ via the formula

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

The "objective model" $\mathcal{R}^*$ is one of the sparsest DAGs so that $p(x^*)$ factorizes according to $\mathcal{R}^*$.

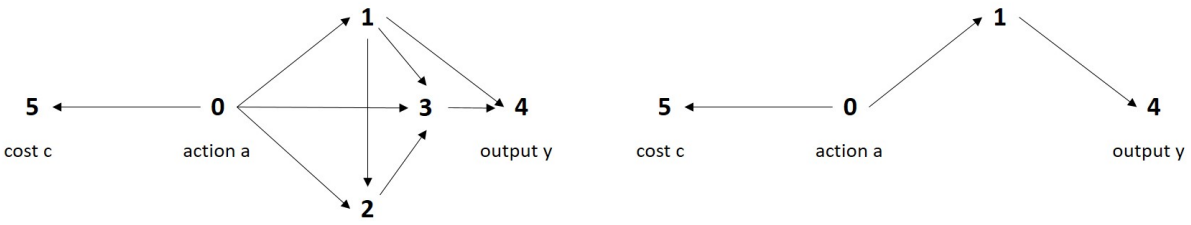

Figure 1: An objective model $\mathcal{R}^*$ (left) and the agent's subjective model $\mathcal{R}$ (right).

[5]This means that the set of variables that influence the output is disjoint from the set of variables that influence effort costs. We will also assume this for the agent's subjective model (defined below).

Beliefs, Personal Equilibrium, and Equilibrium Contract. The agent may have her own subjective model $\mathcal{R} = (N, R)$, see, for example, the graph on the right in Figure 1. We assume that N is a subset of N^* and contains at least the action a, output y, and cost c. Denote by $x = (x_i)_{i \in N}$ the corresponding state vector and $X = \times_{i \in N} X_i$. The agent fits her causal model to the data generated by p, so her beliefs factorize according to the formula

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

She chooses the prescribed action from the contract only if it maximizes her expected utility given the wage scheme $w(y)$ and her subjective belief $p_{\mathcal{R}}$. Since her action potentially influences her beliefs, we formalize the agent's action choice as a personal equilibrium. For this, define by $p_{\mathcal{R}}(y, c \mid a'; p(a))$ the agent's belief about the distribution over outcomes after choosing action a' when her subjective model is $\mathcal{R}$ and her personal equilibrium is $p(a)$.

Definition 1. *The action $p(a)$ 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*

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

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

A fully mixed action profile ensures that all conditional probabilities are well-defined, in particular, those at variables in $\mathcal{R}$ that are directly influenced by a (such as the variables $\{1, 5\}$ in $\mathcal{R}$ of Figure 1). The definition requires that equilibrium beliefs are the limit of a sequence of fully mixed profiles. In the Online Appendix, we show that a personal equilibrium always exists in our framework. We call a contract $(\mathcal{R}, w(y), p(a))$ 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.

The proposed solution concept is static. The agent's beliefs are derived from a probability distribution that could be influenced by the action 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 that she can fit to her subjective model.

3 The Optimal Equilibrium Contract

In this section, we study the properties of the equilibrium contract that is optimal for the principal, given the agent's subjective model $\mathcal{R}$. If $(\mathcal{R}, w^*(y), p^*(a))$ is an optimal equilibrium contract, then $w^*(y), p^*(a)$ solve the maximization problem

$$\max_{w(y)\in W, p(a)\in\Delta(A)} \sum_{a\in A}\sum_{y\in Y} p(a)p(y \mid a)(y - w(y)) \tag{3}$$

subject to the constraints

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

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

If the agent's subjective model $\mathcal{R}$ equals the objective model $\mathcal{R}^*$, the problem collapses to a canonical principal-agent problem and can be solved as suggested by Grossman and Hart (1983). We find for each pure action $a \in A$ the wage scheme $w(y)$ that implements this action at lowest possible cost, and then 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 $\mathcal{R}^*$, one finds the optimal equilibrium contract by applying the same procedure, with one change. Since the agent's beliefs about the distribution over outcomes possibly depend on the implemented action $p(a)$, the first step has to be done for all pure and mixed actions $p(a) \in \Delta(A)$.

We characterize properties of the optimal equilibrium contract in four stages. In Subsection 3.1, we state sufficient conditions on $\mathcal{R}$ and $p^*(a)$ so that the agent has correct expectations on the equilibrium path. These conditions imply that the participation constraint *PC* is not affected by the agent's misspecification. In Subsection 3.2, we show how a misspecification can change the *IC* and how it might eliminate the incentive problem altogether. In Subsection 3.3, we discuss when the principal implements a pure or mixed action $p^*(a)$ and highlight that implementing a pure action may not be optimal. In Subsection 3.4, we characterize for a certain type of $\mathcal{R}^*/\mathcal{R}$-combinations when an agent with misspecified model acts "as if" her model were equal to $\mathcal{R}^*$.[6]

[6]Moreover, in the Online Appendix, we study under what circumstances the contract is also optimal from the agent's (potentially biased) perspective. We thereby introduce the refinement of "justifiability" from the unawareness literature (Filiz-Ozbay 2012, Heifetz et al. 2013) to our framework.

3.1 Correct Expectations on the Equilibrium Path

We examine under what circumstances the agent's beliefs over outcomes are identical to the equilibrium distribution over outcomes. We build on a Bayesian network result from Spiegler (2017) that turns out to be very useful in our setting and from which we can derive several implications. To state this result, we have to introduce a few graph-theoretical concepts. 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. Next, a subset of nodes $S \subset N$ is a clique in $\mathcal{R} = (N, R)$ if iRj or jRi for any two nodes $i, j \in S$. For example, in the DAG $\mathcal{R}^*$ from Figure 1, the set $S = \{1, 3, 4\}$ is a clique, while the set $S' = \{2, 3, 4\}$ is not. Each node is a clique in itself. The following result is a direct implication of Proposition 2 from Spiegler (2017).

Proposition 1 (Equilibrium Beliefs)**.** *If the agent's model $\mathcal{R} = (R, N)$ is perfect, her equilibrium beliefs satisfy $p_{\mathcal{R}}(x_S) = p(x_S)$ for every clique $S \subset N$.*

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 $p(x_k)$.[7] 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 $p_{\mathcal{R}}(x_S \mid a^*; a^*) = p(x_S \mid a^*)$ for every clique $S \subset N$.*

If the equilibrium contract implements a pure action a^*, the agent's beliefs over the joint distribution of any clique in $\mathcal{R}$ conditional on her equilibrium action are correct. Since output y and costs c are independent (both in $\mathcal{R}$ and $\mathcal{R}^*$), this implies that the agent also correctly anticipates the equilibrium distribution over the output, costs, and her payoff.

Corollary 1 is in general not true if the equilibrium contract implements a mixed action $p^*(a)$. While the agent still gets the marginal equilibrium distribution over each variable right, we may also have $p_{\mathcal{R}}(x_i \mid a'; p^*(a)) \neq p(x_i \mid a')$ for an action a' in the support of $p^*(a)$. Thus,

[7] We provide an example in the Online Appendix.

the agent's expected utility conditional on a' may also be biased, $\mathbb{E}_{\mathcal{R}}[u(w(y)) - c \mid a'; p^*(a)] \neq \mathbb{E}[u(w(y)) - c \mid a']$. The second direct implication of Proposition 1 is the following result.

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

If $\mathcal{R}$ is perfect, the principal's wage 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. As we will see in the next subsection, this does not imply that the principal cannot benefit from the agent's misperception.

3.2 Incentive Effects

Next, we study how a misspecification in the agent's subjective model $\mathcal{R}$ can change the contracting problem in (3) when $\mathcal{R}$ is perfect. By Corollary 2, only the incentive compatibility constraint *IC* could then be affected by the misspecification. We examine a simple setting with two levels of effort $a \in \{0, 1\}$, two output levels $y \in \{y_L, y_H\}$, and two cost levels $c \in \{c_L, c_H\}$. Both the probability of high output y_H and the probability of high costs c_H increase in the agent's effort. We consider two "types" of misspecification: First, a misspecification in the agent's model of the production function, so that the agent does not fully understand how her action a maps into the output y; and second, a misspecification in the agent's model of the cost function, which implies that the agent does not fully grasp how a affects her personal costs c.

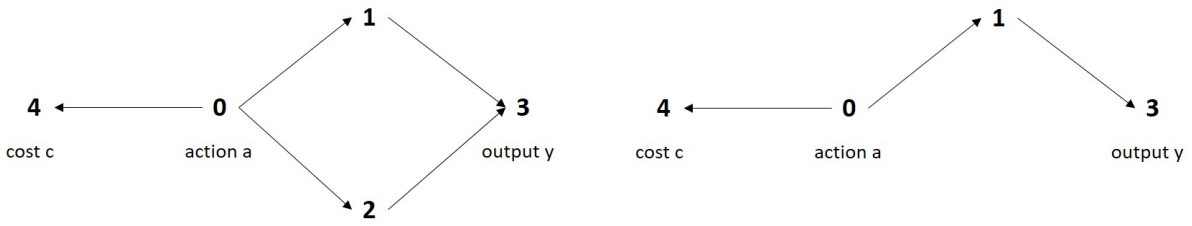

Figure 2a: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in production function example.

Misspecification in the production function model. Consider the objective model $\mathcal{R}^*$ and the agent's subjective model $\mathcal{R}$ from Figure 2a. They represent, for example, the marketer scenario from the introduction in which the agent does not take into account her action's influence on the firm's reputation. Node 1 is the set of customers who are informed about the firm's product. It can be small ($x_1 = 0$) or large ($x_1 = 1$). Node 2 is the firm's reputation, which can be bad ($x_2 = 0$) or good ($x_2 = 1$). For the objective probability distribution, we abbreviate $p(x_i = 1 \mid a) = \beta_a^i$ and $p(y_H \mid x_1, x_2) = \gamma_{x_1,x_2}$. Suppose the principal wishes to implement the

action $p(a = 1) = \alpha > 0$. The incentive compatibility constraint *IC* is then

$$[p_{\mathcal{R}}(y_H \mid a = 1; \alpha) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha)]\,(u(w(y_H)) - u(w(y_L))) \geq \mathbb{E}(c \mid a = 1) - \mathbb{E}(c \mid a = 0), \quad (4)$$

which must hold with strict equality if $\alpha \in (0, 1)$. This is a standard *IC*, except for the term in the quadratic brackets. This term is the agent's perceived effect of effort on output. Note that it depends on the implemented action α. We calculate the $p_{\mathcal{R}}(y_H \mid a; \alpha)$-terms by fitting the agent's model to the objective probability distribution, taking α as given. In the Online Appendix, we present this procedure in detail. The agent's perceived effect of effort on output equals

$$(\beta_1^1 - \beta_0^1)\left[(\gamma_{1,0} - \gamma_{0,0}) + \frac{\alpha\beta_1^1\beta_1^2 + (1-\alpha)\beta_0^1\beta_0^2}{\alpha\beta_1^1 + (1-\alpha)\beta_0^1}(\gamma_{1,1} - \gamma_{1,0}) - \frac{\alpha(1-\beta_1^1)\beta_1^2 + (1-\alpha)(1-\beta_0^1)\beta_0^2}{\alpha(1-\beta_1^1) + (1-\alpha)(1-\beta_0^1)}(\gamma_{0,1} - \gamma_{0,0})\right]. \quad (5)$$

The term in the large quadratic brackets represents $p_{\mathcal{R}}(y_H \mid x_1 = 1; \alpha) - p_{\mathcal{R}}(y_H \mid x_1 = 0; \alpha)$, which indicates how much the probability of high sales y_H changes if the size of the customer base is large rather than small. In the agent's mind, this value is a constant, independent of her action. However, through the reputation channel, this value changes in the agent's equilibrium action α. The direction of change depends on the objective probability distribution. Note that if the value in the large quadratic brackets increases (decreases) in α, an increase in α implies that the incentive compatibility constraint *IC* is relaxed (tightened). Thus, in general, it may be possible that the optimal equilibrium contract implements a mixed action.

We develop some intuition how the agent's misspecification may affect the *IC*. Suppose that the principal wants to implement making cold calls with certainty, $\alpha = 1$. For convenience, we assume that the impact of reputation on sales is positive and independent of the customer set, $\gamma_{1,1} - \gamma_{1,0} = \gamma_{0,1} - \gamma_{0,0} > 0$. Under the objective model, the effect of effort on output is then

$$p(y_H \mid a = 1) - p(y_H \mid a = 0) = (\beta_1^1 - \beta_0^1)(\gamma_{1,0} - \gamma_{0,0}) + (\beta_1^2 - \beta_0^2)(\gamma_{0,1} - \gamma_{0,0}), \quad (6)$$

while under the agent's subjective model this value equals

$$p_{\mathcal{R}}(y_H \mid a = 1; \alpha = 1) - p_{\mathcal{R}}(y_H \mid a = 0; \alpha = 1) = (\beta_1^1 - \beta_0^1)(\gamma_{1,0} - \gamma_{0,0}). \quad (7)$$

Suppose that making cold-calls increases the customer base, $\beta_1^1 > \beta_0^1$. When some people are annoyed by being cold-called, cold-calls lower the firm's reputation, $\beta_1^2 < \beta_0^2$. In this case, the agent overestimates the drop in expected sales if she deviates to low effort. The *IC* is then relaxed by the misspecification. Alternatively, a talented agent may even be able to increase the firm's reputation by making cold-calls, $\beta_1^2 > \beta_0^2$. If the agent does not take this effect into

account, the incentive constraint is tightened relative to the objective model. Finally, suppose that making cold-calls (for some strange reason) decreases the customer base, $\beta_1^1 < \beta_0^1$, but increases the firm's reputation, $\beta_1^2 > \beta_0^2$. The agent with misspecified model then perceives her effort as unproductive, so that it is impossible to implement high effort.

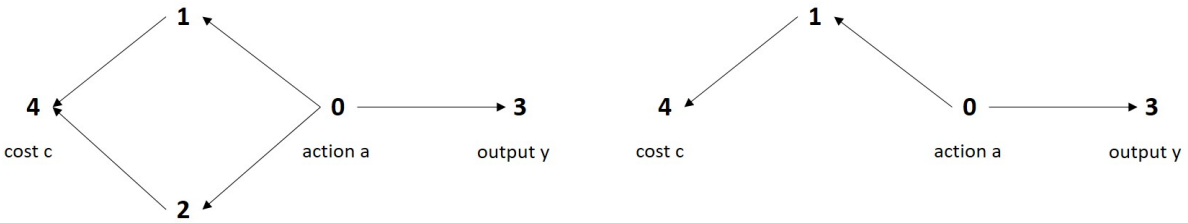

Figure 2b: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in cost function example.

Misspecification in the cost function model. Next, consider the objective model $\mathcal{R}^*$ and the agent's subjective model $\mathcal{R}$ from Figure 2b. They represent, for example, the following scenario: The agent is an accomplished technical specialist who is promoted to a management position. Her psychological well-being now depends on the atmosphere in her department. However, her people skills are poor. She does not understand that her efforts impact on her subordinates' mood and thus on her own quality of life at the workplace. Besides the interaction with colleagues, also the manager's compliance to her own work norms is important for her personal costs. Let node 1 be the degree of stress the manager experiences if she does not comply to her work norms (or does not achieve her goals), which can be low ($x_1 = 0$) or high ($x_1 = 1$). Node 2 is her relationship to her colleagues, which can be good ($x_2 = 0$) or bad ($x_2 = 1$). Both the level of stress and the quality of relationship influence the agent's cost in the expected manner. When the principal wishes to implement action $p(a = 1) = \alpha > 0$, the incentive compatibility constraint IC is

$$\mathbb{E}(u(w(y)) \mid a = 1) - \mathbb{E}(u(w(y)) \mid a = 0) \geq [p_{\mathcal{R}}(c_H \mid a = 1; \alpha) - p_{\mathcal{R}}(c_H \mid a = 0; \alpha)](c_H - c_L), \quad (8)$$

which must hold with strict equality if $\alpha \in (0, 1)$. The term in the quadratic brackets on the right-hand side now represents the agent's perceived increase in the probability of high costs if she exerts high rather than low effort. If this value decreases (increases), the incentive compatibility constraint is relaxed (tightened). Since we chose the objective and subjective model symmetric to the production function example, this term is identical to (5) when we use the abbreviations $p(x_i = 1 \mid a) = \beta_a^i$ and $p(c_H \mid x_1, x_2) = \gamma_{x_1,x_2}$.

We again build intuition for how misspecifications affect effort incentives. Suppose that the principal implements high effort with certainty, $\alpha = 1$. The objective effect of effort on

personal costs is then

$$p(c_H \mid a = 1) - p(c_H \mid a = 0) = (\beta_1^1 - \beta_0^1)(\gamma_{1,0} - \gamma_{0,0}) + (\beta_1^2 - \beta_0^2)(\gamma_{0,1} - \gamma_{0,0}), \tag{9}$$

while under the subjective model it is

$$p_{\mathcal{R}}(c_H \mid a = 1; \alpha = 1) - p_{\mathcal{R}}(c_H \mid a = 0; \alpha = 1) = (\beta_1^1 - \beta_0^1)(\gamma_{1,0} - \gamma_{0,0}). \tag{10}$$

Suppose that the manager has strong work norms and suffers from non-compliance, so that $\beta_1^1 < \beta_0^1$. In this case, the misspecification makes the manager believe that high effort is the low-cost action. High effort then can be implemented with a fixed wage so that the first-best allocation is realized. Therefore, a misspecification in the agent's subjective model can eliminate the incentive problem altogether.

Assume now that her personal norms emphasize the work-life balance, which means that $\beta_1^1 > \beta_0^1$. If working less hard improves the relationship to her colleagues (since there is more time for socializing), $\beta_1^2 > \beta_0^2$, not taking into account this effect relaxes the incentive compatibility constraint. If working less hard worsens the relationship to her colleagues (since there are strong norms in the organization that everybody works hard), $\beta_1^2 < \beta_0^2$, ignoring this effect tightens the incentive compatibility constraint.

3.3 The Agent's Action under the Optimal Equilibrium Contract

We now know how a misspecification can change the incentive compatibility constraint. Next, we ask whether the optimal equilibrium contract implements a pure action. Recall from Corollary 1 that in a pure action personal equilibrium, the agent correctly anticipates the joint distribution of any clique $S \in N$ conditional on her equilibrium action. For the subjective models from Subsection 3.2, this implies that the agent correctly anticipates the joint distribution of all variables in her model.

Unfortunately, there does not seem to be a simple set of conditions that are necessary and sufficient for an optimal equilibrium contract to implement a pure action. For a specific application, one has to show this by using ad-hoc methods.[8] We can make two general statements. First, consider a setting in which expected output increases in the agent's effort a and let a_H be the highest level of effort. Assume that personal cost c are always weakly positive, and that the agent's subjective model is so that $\mathbb{E}_{\mathcal{R}}[y \mid a; a_H]$ increases in a. Then, if ceteris paribus personal cost levels c are small enough, the optimal equilibrium contract implements a_H with certainty. Second, let a^* be a Pareto-efficient effort level. If $\mathcal{R}$ is perfect and the misspecification in $\mathcal{R}$

[8] In the Online Appendix, we do this for the production function example from Subsection 3.2.

implies that $a^* \in \arg\min_{a \in A} \mathbb{E}_{\mathcal{R}}[c \mid a; a^*]$ (as in the cost function example in Subsection 3.2), then the optimal equilibrium contract implements a^* with certainty.

However, it is not always optimal for the principal to implement a pure action, as the following example shows. Consider the production function from the previous subsection. Assume that the agent is risk-neutral, protected by limited liability so that $w(y) \geq 0$, her outside option value is zero, and $y_L = 0$. We show that the optimal equilibrium contract may implement a mixed action $\alpha \in (0, 1)$. 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 (4) is satisfied with equality. The principal's expected payoff from this contract is then

$$\mathbb{E}[V] = (\alpha p(y_H \mid a = 1) + (1 - \alpha) p(y_H \mid a = 0)) \left[y_H - \frac{\mathbb{E}[c \mid a = 1] - \mathbb{E}[c \mid a = 0]}{\Delta(\alpha)} \right], \tag{11}$$

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

$$\left. \frac{d\Delta(\alpha)}{d\alpha} \right|_{\alpha=1} = (\beta_1^1 - \beta_0^1)(\beta_1^2 - \beta_0^2) \left[\frac{\beta_0^1}{\beta_1^1} (\gamma_{1,1} - \gamma_{1,0}) - \frac{1 - \beta_0^1}{1 - \beta_1^1} (\gamma_{0,1} - \gamma_{0,0}) \right]. \tag{12}$$

Let the agent's action have a positive impact on both components, $\beta_1^1 > \beta_0^1$ and $\beta_1^2 > \beta_0^2$. Then for $\beta_1^1 \to 1$ the slope converges to minus infinity. Thus, if all else equal β_1^1 is sufficiently close to 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 $\alpha \in (0, 1)$.[9]

3.4 Behavioral Rationality

A misspecification in the agent's subjective model does not necessarily change the contracting problem in (3). We derive a result that characterizes which nodes must be in the agent's subjective model $\mathcal{R}$ to guarantee that she acts as if she were rational. This result has a number of important implications that we will illustrate in subsequent applications.

In the following, we assume that the objective model $\mathcal{R}^*$ is perfect and that the agent's subjective model $\mathcal{R} = (N, R)$ originates from $\mathcal{R}^*$ by omitting nodes and the links attached to these nodes. Formally, this means that R equals R^* restricted on N: We have iRj for $i, j \in N$ if and only if iR^*j. Note that $\mathcal{R}$ will then be perfect (no v-structure emerges if we take out nodes from $\mathcal{R}^*$ and all links attached to them). The assumptions on $\mathcal{R}^*$ and $\mathcal{R}$ are not restrictive. Any

[9]For example, if $y_H = 1$, $\mathbb{E}[c \mid a = 1] = 0.1$, $\mathbb{E}[c \mid a = 0] = 0$, $\beta_1^1 = \beta_1^2 = 0.95$, $\beta_0^1 = \beta_0^2 = 0.40$, $\gamma_{1,1} = 0.9$, $\gamma_{1,0} = 0.6$, $\gamma_{0,1} = 0.4$, and $\gamma_{0,0} = 0.1$, the optimal equilibrium contract implements $\alpha \simeq 0.94$.

probability distribution p factorizes according to some perfect DAG $\mathcal{R}^*$, and the assumption on $\mathcal{R}$ is satisfied by all subjective models we consider in this paper.

For any perfect $\mathcal{R}^*$ we characterize the subset of nodes $H^* \subseteq N^*$ the agent needs to have in her subjective model $\mathcal{R} = (N, R)$ so that she acts rationally, regardless of her incentives $w(y)$. If $H^* \subseteq N^*$, the agent acts "as if" her subjective model were given by $\mathcal{R}^*$ so that the problem in (3) collapses to the canonical principal-agent problem. We formally define this case.

Definition 2. *An agent with subjective DAG $\mathcal{R}$ is behaviorally rational if, for any probability distribution p and any incentive scheme $w(y)$, a personal equilibrium at $\mathcal{R}$ and $w(y)$ is also a personal equilibrium at the objective DAG $\mathcal{R}^*$ and $w(y)$.*

To derive the characterization of H^*, we will make use of 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 3. *Two DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent if $p_{\mathcal{R}}(x) \equiv p_{\mathcal{G}}(x)$ for every $p \in \Delta(X)$.*

Proposition 2 (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.*

To illustrate, consider the models in Figure 2a. The DAGs $\mathcal{R}^*$ and $\mathcal{R}$ are not equivalent since they have different skeletons. Consider a DAG $\mathcal{G}$ that only differs from $\mathcal{R}^*$ in that the link between the nodes 1 and 3 is reversed. The two DAGs then have the same skeleton, but a different v-structure. Then there exist probability distributions p so that the agent's subjective beliefs $p_{\mathcal{G}}$ differ from p if her subjective model is given by $\mathcal{G}$.

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 of an ancestral node 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 4. *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 ($0R^*1$, $0R^*2$, $0R^*3$, and $0R^*5$). Thus, the two links pointing into the outcome node 4 ($1R^*4$ and $3R^*4$) also must be fundamental. If we would turn around one or both of them, we would create a *v*-collider since there is no link between the action node 0 and the output 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 5. *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.*

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 outcome nodes by

$$H^* := \{i \in N^* \mid i \text{ is part of a fundamental active path between 0 and } n \text{ or } n+1\}.$$

It turns out that the nodes in H^* are exactly those nodes the agent needs to have in her subjective DAG in order to be behaviorally rational. 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 $N^* \setminus H^*$ to nodes in H^*. In this DAG, the nodes that are not in H^* have no influence on output or costs, so the agent can safely ignore them. By Proposition 2, the agent knows the true statistical relationship between actions and outcomes if $H^* \subseteq N^*$.

Proposition 3 (Behavioral Rationality)**.** *Let $\mathcal{R}^*$ be a perfect DAG. The agent is behaviorally rational if and only if her subjective DAG $\mathcal{R}$ contains all nodes from H^*.*

This result has a number of implications. First, Proposition 3 shows that the agent may not take into account all variables of her environment and still behave fully rational. In Section 5, we illustrate that $N^* \setminus H^*$ may contain important variables that are crucial for the description of the contracting problem. Conversely, the incentive compatibility constraint may be affected if the agent's subjective model misses out at least one node from H^*.

Second, Proposition 3 implies that different misspecifications – which may reflect different psychological predispositions or misperceptions – can have the same effect on incentives.

Consider two different nodes $i, j \in H^*$ in a perfect DAG $\mathcal{R}^*$. Let $\mathcal{R}_{-i}$ be the model that is identical to $\mathcal{R}^*$ except that node i is missing in $\mathcal{R}_{-i}$; define $\mathcal{R}_{-j}$ accordingly. We can now examine which nodes are on fundamental active paths in the misspecified models $\mathcal{R}_{-i}$ and $\mathcal{R}_{-j}$. Suppose that node j is not on a fundamental active path of $\mathcal{R}_{-i}$. It then does not matter for the agent's incentives whether she ignores node i or the nodes i, j. If also node i is not on a fundamental active path of $\mathcal{R}_{-j}$, it does not matter for incentives, whether the agent ignores node i, node j, or both nodes. We will discuss such a case in Subsection 5.2. Therefore, the ignorance about one channel of causality may render another channel unimportant.

Finally, a practical implication of Proposition 3 is that it can simplify the analysis of the agent's subjective beliefs (in combination with Proposition 4 below). Deriving equilibrium beliefs from the agent's subjective model $\mathcal{R}$ and her equilibrium action $p(a)$ can be cumbersome. The result indicates whether omitting a node in the subjective DAG affects beliefs or not. Nodes in $N^* \setminus H^*$ can be ignored in the calculations.

We can partially extend Proposition 3 to imperfect DAGs. Any imperfect DAG can be made perfect by adding links. This does not restrict, but extends the set of probability distributions that factorize according to the DAG. In this way, Proposition 3 can also be applied to the imperfect DAGs in Subsection 3.2 (see the Online Appendix). It remains to provide an algorithm that identifies H^* in perfect DAGs. Nodes that are connected by fundamental links in perfect DAGs fortunately exhibit characteristics that are easy to identify.

Proposition 4 (Characterization of 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. First, find for each node i the distance to the action node, $d(0, i)$. Links between nodes of differing distance are fundamental links. Second, 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 $d > 0$.

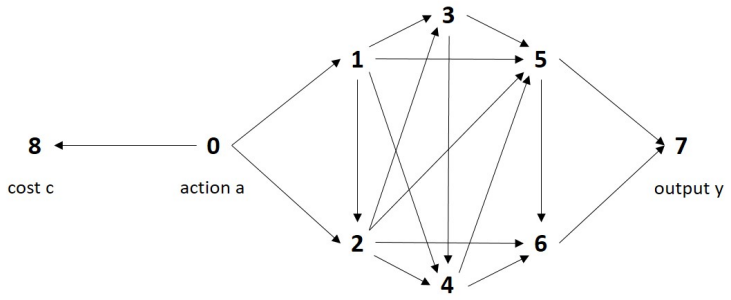

Figure 3: Example DAG $\mathcal{R}^*$.

We apply this algorithm to the perfect DAG $\mathcal{R}^*$ in Figure 3. Condition (a) from Proposition 4 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 4 then implies that $4R^*6$ and $5R^*6$ are fundamental links, while the remaining links are non-fundamental. The set of nodes on fundamental active paths is therefore given by $H^* = N^* \setminus \{3\}$. By Proposition 3, if the agent's subjective model includes this set of nodes, but not node 3, then she still acts in equilibrium as if she knew the complete project, regardless of the probability distribution p and the incentives scheme $w(y)$.

4 Endogenous Misspecification

So far, we assumed that the agent's subjective model $\mathcal{R}$ is exogenously given. However, in many cases the principal – who designs the agent's workplace – may be able to influence her subjective view on the project. This can happen by hiding certain empirical regularities, while highlighting others when describing the job to the agent. For example, in our marketer example, the principal may emphasize to the agent that it is really important to contact as many potential customers as possible in order to increase sales, while he does not mention the downsides of such a strategy.

Formally, the organization of the project defines the probability distribution p and the corresponding objective model $\mathcal{R}^*$. The principal then can decide which components of the objective model the agent takes into account, potentially subject to certain informational constraints. He chooses the subjective model $\mathcal{R}$ for the contract $(\mathcal{R}, w(y), p(a))$ out of a set Γ. This set captures all constraints, i.e., the components of the project that the agent always takes into account (e.g., that there are sales y), or, alternatively, components that the agent does not understand and thus are never in her subjective model. Γ contains at least one element. If $\Gamma = \{\mathcal{R}^*\}$, the agent is fully rational and the problem again collapses to the canonical principal-agent model. If for some misspecified model $\hat{\mathcal{R}} = (\hat{N}, \hat{R})$ we have $\Gamma = \{\hat{\mathcal{R}}\}$, the principal cannot educate

the agent about the objective model and we are in the setting of Section 3. If $\Gamma = \{\hat{\mathcal{R}}, \mathcal{R}^*\}$, the principal can choose whether he wants to keep the agent's model misspecified at $\hat{\mathcal{R}}$ or to educate her about the objective model $\mathcal{R}^*$. Recall from our marketer example in Subsection 3.2 that there may be cases for both options: If making cold-calls decreases the firm's reputation, the principal prefers to keep the agent's model misspecified when implementing high effort; if it increases the firm's reputation, the principal wants to inform the agent about this additional effect of her effort on sales.

We now find the optimal equilibrium contract by identifying the optimal incentive scheme $w(y)$ and action $p(a)$ combination for every given $\mathcal{R} \in \Gamma$, and then picking the model-action-incentive scheme combination that maximizes the principal's expected payoff (so that Grossman and Hart's 1983 two-step procedure becomes a three-step procedure). For convenience, we assume that the principal can select any subjective model $\mathcal{R} \in \Gamma$ without incurring further costs. In reality, it may require extensive effort to update the agent's model. To convince her of certain empirical regularities, the principal may present a persuasive narrative[10], demonstrate an effect through successful test trials (Bloom et al. 2013), easy-to-understand summary statistics (Hanna et al. 2014), or forced experimentation (Larcom et al. 2019).

5 Organization and Model Misspecification

In this section, we examine which misspecifications are robust in an organization, and how this affects optimal workplace design. We demonstrate in a number of concrete examples that organization and model misspecification mutually influence each other, and that in many cases the principal strictly benefits from keeping the agent's subjective view of the organization misspecified. Several new implications for optimal organization will emerge from this analysis. In Subsection 5.1, the principal can choose between different production technologies. In Subsection 5.2, the principal selects between a transparent and a non-transparent structure of the workplace. In Subsection 5.3, the principal can hire a reflective or an egocentric agent who becomes a "narcissistic leader." In Subsection 5.4, we study how the principal optimally implements relative performance evaluation. In Subsection 5.5, the principal chooses whether to offer a complete or an incomplete contract. When necessary, we slightly advance the basic model from Section 2, but the insights from Section 3 will remain valid, and we will make use of them in many instances.

[10]For example, the medical doctors of the 19th century were finally persuaded about the importance of hospital hygiene through Louis Pasteur's germ theory.

5.1 Technology Choice

A crucial design feature of the agent's workplace is the production technology. It determines how the agent's effort changes the distribution over outputs. In this subsection, we examine the principal's preferences over production technologies. We show that if the principal can keep the agent's model of the production function misspecified, he prefers "extreme" technologies, in the sense that they entail both large up- and downsides. This may cause technological inertia: The principal may not want to change the production technology even if a new technology is available that is objectively superior to the old one.

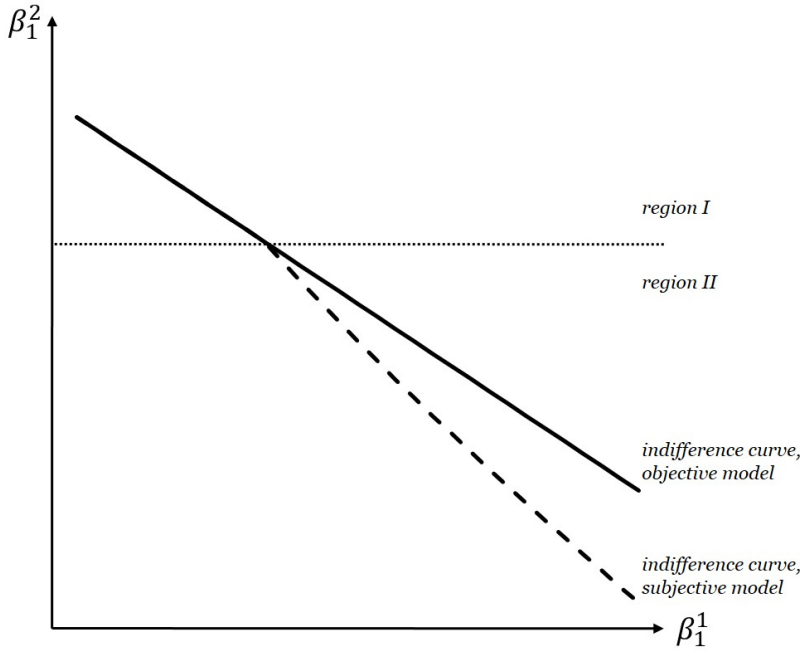

Figure 4: Principal's preferences over production technologies

We continue the production function example from Subsection 3.2. Assume that the agent is risk-neutral, protected by limited liability so that $w(y) \geq 0$, and her outside option value is zero. We normalize $y_L = 0$ and treat the parameters β_1^1 (the impact of the marketer's effort on the customer set) and β_1^2 (the impact of the marketer's effort on reputation) as technology parameters of the marketing strategy. The other parameters we keep fixed. Suppose first that the agent is rational. Standard arguments show that the optimal wage scheme that implements high effort specifies $w(y_L) = 0$, and $w(y_H)$ so that the *IC* is satisfied with equality. The agent therefore earns an information rent, and the principal's expected payoff from this scheme is

$$\mathbb{E}[V \mid \mathcal{R} = \mathcal{R}^*] = p(y_H \mid a = 1)\left[y_H - \frac{\mathbb{E}[c \mid a = 1] - \mathbb{E}[c \mid a = 0]}{p(y_H \mid a = 1) - p(y_H \mid a = 0)}\right]. \tag{13}$$

The principal's expected payoff strictly increases in $p(y_H \mid a = 1)$. The slope of his indifference curves in the $\beta_1^1 - \beta_1^2$-space – how he trades-off changes in the technology with respect to its effect on customer set and reputation – is $\frac{d\beta_1^2}{d\beta_1^1} = -\frac{\gamma_{1,0}-\gamma_{0,0}}{\gamma_{0,1}-\gamma_{0,0}}$, see the solid line in Figure 4 ("indifference curve, objective model"). The principal's preferences over technologies are identical those of a social planner whose objective is to maximize the value of the project.

Next, suppose that the principal can choose the agent's subjective model $\mathcal{R}$ from the set $\Gamma = \{\hat{\mathcal{R}}, \mathcal{R}^*\}$, where $\hat{\mathcal{R}}$ is the misspecified model on the right of Figure 2a. That is, he can inform the agent about the reputation component or keep her model misspecified. This causes a kink in the principal's preferences over technologies. Recall from Section 3.2 that at (β_1^1, β_1^2)-combinations with $\beta_1^2 > \beta_0^2$ the misspecification tightens the IC. In this case, the principal prefers $\mathcal{R} = \mathcal{R}^*$, so that her preferences locally are the same as in the rational framework. However, at (β_1^1, β_1^2)-combinations with $\beta_1^2 < \beta_0^2$ the principal prefers $\mathcal{R} = \hat{\mathcal{R}}$. His expected payoff when he implements high effort with probability $\alpha = 1$ is then

$$\mathbb{E}[V \mid \mathcal{R} = \hat{\mathcal{R}}] = p(y_H \mid a = 1)\left[y_H - \frac{\mathbb{E}[c \mid a = 1] - \mathbb{E}[c \mid a = 0]}{p_{\hat{\mathcal{R}}}(y_H \mid a = 1; \alpha = 1) - p_{\hat{\mathcal{R}}}(y_H \mid a = 0; \alpha = 1)}\right], \quad (14)$$

and the slope of the principal's indifference curves in the $\beta_1^1 - \beta_1^2$-space is

$$\frac{d\beta_1^2}{d\beta_1^1} = -\frac{\gamma_{1,0} - \gamma_{0,0}}{\gamma_{0,1} - \gamma_{0,0}}\left[1 + \frac{p(y_H \mid a = 1)\frac{\mathbb{E}[c|a=1]-\mathbb{E}[c|a=0]}{[(\beta_1^1-\beta_0^1)(\gamma_{1,0}-\gamma_{0,0})]^2}}{y_H - \frac{\mathbb{E}[c|a=1]-\mathbb{E}[c|a=0]}{(\beta_1^1-\beta_0^1)(\gamma_{1,0}-\gamma_{0,0})}}\right]. \quad (15)$$

The dashed line in Figure 4 shows the principal's indifference curves in the $\beta_1^1 - \beta_1^2$-space when he can choose the agent's subjective model. It is steeper when the agent's model is misspecified (region II) than when it is correctly specified (region I). The kink occurs when β_1^2 falls below β_0^2. Intuitively, it occurs because, in region II, improvements in the customer set domain reduce the information rent that the principal has to pay to the agent, while improvements in the reputation domain do not. Hence, the principal's preferences are distorted relative to the rational benchmark, and no longer identical to that of the social planner.

These distorted preferences have two implications. First, the principal chooses "extreme" production technologies. Suppose that he can select a technology (β_1^1, β_1^2) out of a set Ξ. For convenience, assume that all available (β_1^1, β_1^2)-combinations in Ξ roughly generate the same effect of effort on output $p(y_H \mid a = 1)$. The principal then strictly prefers the technology with the highest β_1^1 and smallest β_1^2, provided that this β_1^2 is below β_0^2. Intuitively, this technology offers the most effective "narrative" to the agent of why her effort is crucial for the output. The principal then highlights its advantage (its positive impact on the customer base), and hides its disadvantage (its negative effect on reputation) to exploit an incentive effect.

The second implication are technological inertia. Suppose that additional to the set Ξ there is a new production technology $(\bar{\beta}_1^1, \bar{\beta}_1^2)$ that is superior to any element in Ξ, i.e., it offers a higher value of $p(y_H \mid a = 1)$. The principal may still prefer the old, extreme technology to $(\bar{\beta}_1^1, \bar{\beta}_1^2)$ if the advantage of the new technology is not large enough and it is more balanced in the sense that $\bar{\beta}_1^2 > \beta_1^2$ and $\bar{\beta}_1^1 < \beta_1^1$. Thus, the principal may stick to an inefficient technology.

Indeed, it has been documented that many successful firms do not adjust to technological changes. March and Simon (1958) and Cyert and March (1963) argue that bounded rationality may cause firms to maintain their current structure as long as performance is not abnormally poor performance. Hannan and Freeman (1984) propose that selection tend to favor stable organizations. Another explanation for inertia is that changes are likely to create winners and losers within the organization (Milgrom 1988). Our explanation does not require conflicts in the organization and is consistent with a rational principal as well as an agent who has correct expectations on the equilibrium path.

5.2 Transparency

An important aspect of workplace design is how easily workers observe each others' efforts, i.e., how transparent the workplace is. Peer effects can be crucial for the success of an organization (e.g., Mas and Moretti 2009). Winter (2010) examines the interaction between incentives and transparency among peers.[11] He demonstrates that if the production function exhibits a complementarity, transparency reduces the need to provide incentives. We build on this insight to show that transparency can also affect an agent's beliefs and behavioral rationality. Therefore, transparency can have advantages beyond those suggested by Winter (2010).

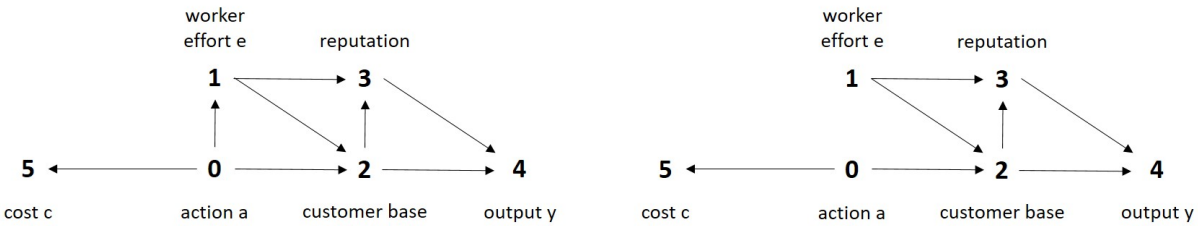

Figure 5a: Objective model $\mathcal{R}_T^*$ of the transparent organization (left) and objective model $\mathcal{R}_{NT}^*$ of the non-transparent organization (right).

Let there be two employees, the agent and the worker. Their job is to increase sales y through marketing. The agent's action a (online-marketing) increases the size of the customer base; the worker's effort e is to make cold-calls, which also has a positive effect on the cus-

[11] Interestingly, his framework also uses graphs to model the interaction structure.

tomer base, but a negative effect on the firm's reputation. Both the size of the customer base and reputation have a positive effect on sales. Some consumers become aware of the product only if approached through multiple channels. Thus, the employees' efforts have a complementary effect on the customer base. The principal can choose between a transparent and a non-transparent organization. In the transparent organization, the worker observes the agent's action a before choosing her effort e; in the non-transparent organization, agent and worker choose their efforts independently. We study how a misspecification in the agent's model changes the relative advantages of the two organizational forms.

Non-transparent Organization. If the two employees choose their efforts independently, the objective model of the organization can be represented by model $\mathcal{R}^*_{NT}$ on the right of Figure 5a ("NT" for "non-transparent"). Suppose the principal implements the worker effort $\bar{e}$. We then can simplify the objective probability model so that it can be represented by a perfect DAG. Write $\tilde{p}(x_2 \mid a) = p(x_2 \mid a, \bar{e})$ and $\tilde{p}(x_3 \mid x_2) = p(x_3 \mid \bar{e}, x_2)$. The true probability distribution over the variables $x^* = (a, x_2, x_3, y, c)$ is then given by

$$p(x^*) = p(a)\tilde{p}(x_2 \mid a)\tilde{p}(x_3 \mid x_2)p(y \mid x_2, x_3)p(c \mid a). \tag{16}$$

The model $\mathcal{R}_1$ at the upper-left of Figure 5b represents this factorization formula. In this model, the reputation node 3 is not on a fundamental active path. Thus, the agent's behavior is independent of whether she takes the reputation component into account or not (as in model $\mathcal{R}_4$ of Figure 5b). The agent is behaviorally rational, regardless of whether her subjective model is $\mathcal{R}_1$ or $\mathcal{R}_4$. Intuitively, if the agent does not affect the worker's action, it does not matter whether she fully understands how her peer's job is related to the final output.

Transparent Organization. If the worker's effort e depends on the agent's action a, the objective model of the organization can be represented by model $\mathcal{R}^*_T$ on the right of Figure 5a ("T" for "transparent"). All notes from $\mathcal{R}^*_T$ are on fundamental active paths, so the agent is no longer behaviorally rational if her subjective model omits one of them. Transparency thus has an effect on the agent's behavioral rationality.

Interestingly, different misspecifications have the same effect on incentives. Consider the subjective models in Figure 5b. Model $\mathcal{R}_1$ represents the thinking of a person who ignores others' contribution to the final output. Model $\mathcal{R}_2$ displays the reasoning of an agent who acknowledges both the contribution of the worker and the importance of reputation, but does not understand the worker's job sufficiently well to grasp the impact of effort e on reputation. Model $\mathcal{R}_3$ is the subjective model of an agent who takes into account the contribution of the worker, but ignores the reputation component. Finally, model $\mathcal{R}_4$ contains all these misspec-

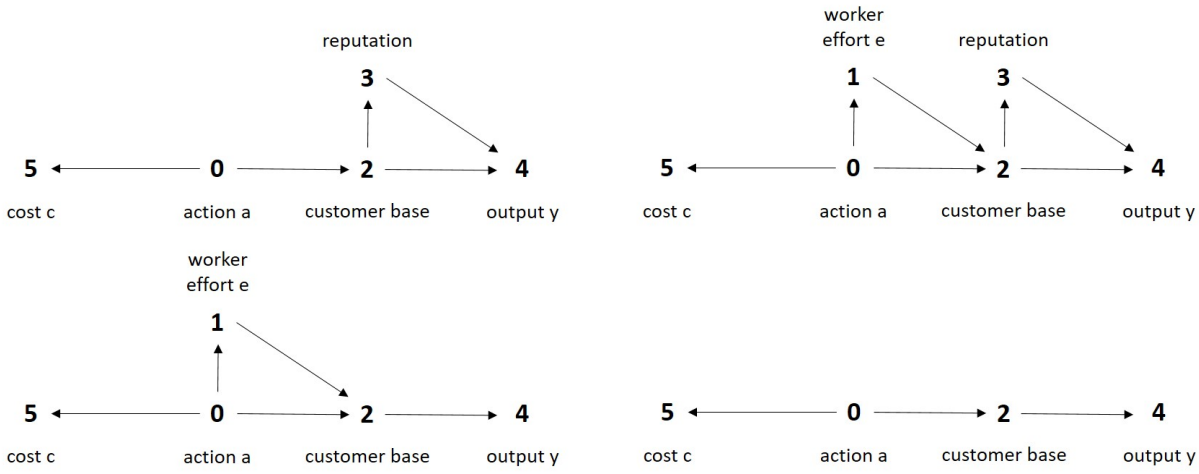

Figure 5b: Subjective model $\mathcal{R}_1$ (upper-left), subjective model $\mathcal{R}_2$ (upper-right), subjective model $\mathcal{R}_3$ (lower-left), and subjective model $\mathcal{R}_4$ (lower-right).

ifications. Applying Propositions 3 and 4, we can show that the set of nodes on fundamental active paths between action and outcomes is $\{0, 2, 4, 5\}$ in all these models, so that the agent's *IC* is identical under all these misspecification. Thus, in this environment, it does not matter which aspect of the principal's project the agent does not take into account.

Given the qualitative description of the production function, these misspecifications benefit the principle. Suppose the worker's effort $e \in \{0, 1\}$ increases in the agent's action $a \in \{0, 1\}$ (this is the case under an optimal incentive scheme if the production function is supermodular). If the agent's subjective model is misspecified, she does not take into account how inspiring the worker has a partial negative effect on sales, which relaxes the agent's *IC* as in Section 3.2. Thus, if the principal implements a transparent organization, he strictly prefers to keep the agent's model misspecified. Importantly, this increases the relative advantage of the transparent organization relative to the non-transparent one. There are cases where the principal would prefer the non-transparent organization if the agent has rational expectations, but prefers the transparent structure if he can keep her model misspecified (see the Online Appendix). In particular, a transparent structure might be strictly preferred even if the production function is not supermodular, which is the requirement in Winter's (2010) framework for positive incentive effects of transparency.

5.3 Narcissistic Leadership

An important question in psychology and managerial economics is whether narcissistic leaders are good or bad for organizational performance. A narcissistic personality exhibits a pronounced sense of self-importance, egocentricity, lack of empathy, and potentially arrogant

behaviors (Rosenthal and Pittinsky 2006). Psychologists propose that, within groups, narcissistic individuals often emerge as leaders, presumably due to their extraversion, which is recognized as an indicator of leadership quality (e.g., Grijalva et al. 2015). Indeed, many important politicians and managers exhibit a narcissistic personality (e.g., Watts et al. 2013). However, many psychologists also argue that narcissistic leadership has a negative impact on performance. Nevicka et al. (2011) find that narcissistic individuals – when they are in the position of a leader – ask for advice too infrequently and reduce the exchange of tacit information within a group. They suggest that "[...] narcissistic leaders, with their characteristic self-absorption and egocentrism, are biased to focus on their own information rather than to solicit unique information from other group members" (Nevicka et al. 2011, p. 1260).

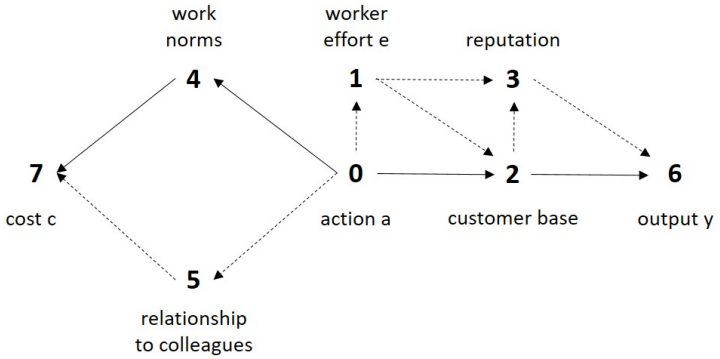

Figure 6: The objective model $\mathcal{R}^*$ (solid and dashed lines) and the egocentric agent's subjective model $\mathcal{R}$ (only solid lines) in the narcissistic leadership example.

Our framework allows for an incentives-based perspective on narcissistic leadership. We can show (a) that different features of a narcissistic personality can be traced back to the boundedly rational expectations of an egocentric individual, and (b) that the principal may have a preference for narcissistic leadership, so that he hires an agent who exhibits such a personality. Consider the objective model $\mathcal{R}^*$ in Figure 6. On the production side, it is the model $\mathcal{R}^*_T$ from the transparent organization from Figure 5a; on the cost side, it is the model from the specialist manager example from Figure 2b. A reflective agent who asks the worker about her job and colleagues about their needs at the workplace may gather enough information so that her subjective model is given by $\mathcal{R}^*$. In contrast, an egocentric agent may ignore the worker's knowledge and her colleagues' needs. Her subjective model $\hat{\mathcal{R}}$ then remains misspecified so that it lacks the nodes $\{1, 3, 5\}$ and the dashed links from $\mathcal{R}^*$ in Figure 6.

Let effort be binary, $a \in \{0, 1\}$. Suppose that both the misspecification in the production- and cost-function relax the incentive constraint that ensures high effort; we only have to use the specification of the production function from Subsection 5.2, and the specification of the cost function from Subsection 3.2 (i.e., the agent's effort has a positive impact on the two components 4 and 5, which in turn have a positive impact on the agent's costs). If the principal

wishes to implement high effort, he has a strict preference to hire an agent with subjective model $\hat{\mathcal{R}}$. In a standard setting, the *IC* is then be binding under the optimal equilibrium contract with $\mathcal{R} = \hat{\mathcal{R}}$. The egocentric agent then acts as if she exhibits a lack of empathy: She does not take into account that effort worsens the relationship to her colleagues; otherwise, she would choose low effort. Moreover, she exhibits a pronounced sense of self-importance. Since she does not take into account the partially negative impact of her action on sales, she overestimates her contribution to the final output. Both aspects are desirable from the principal's perspective.

Note that the probability model needed to get a strict preference for an egocentric agent indicates when we should expect narcissistic leadership. It occurs when the action that benefits the organization the most goes against the interests of some individuals, and when the production technology has partially negative consequences for the contractible output measure.

5.4 Relative Performance Evaluation

An optimal incentive scheme may condition not only on the agent's own performance, but also on the performance of others. The benefit of relative performance evaluation (RPE) is that it reduces the agent's exposure to common risks. Moreover, in many environments, it may be the only way to provide incentives when there exists no objective performance standard (Murphy 2001). However, an important drawback of RPE is that it encourages workers to sabotage each other's efforts (Lazear 1989, Gibbons and Murphy 1990) or to reduce helping and information sharing (Drago and Garvey 1998). Labor unions are often quite hostile towards RPE. Thus, many firms are reluctant to use it openly. We show that by using a camouflaged version of RPE, the principal can take advantage of its positive incentive effects and avoid the downsides. The idea is that, instead of using RPE openly, the principal uses "subjective" performance evaluation, which appears to the agent as being related only to her effort, but in reality draws on the performance of others.

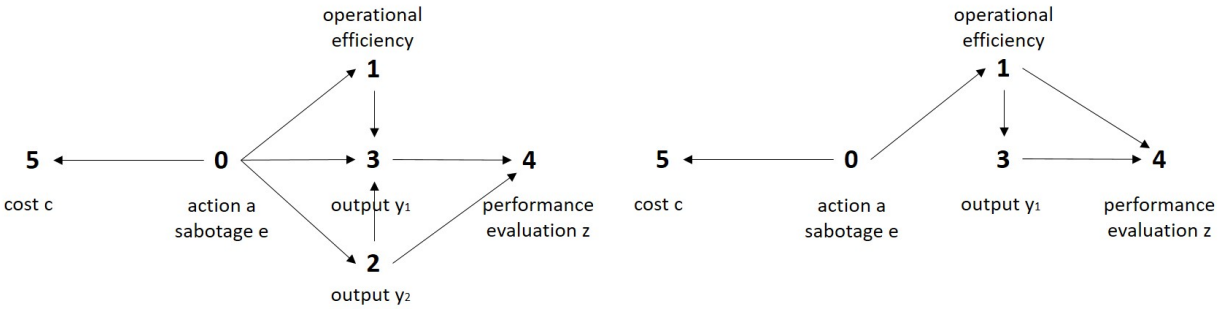

Figure 7a: Objective model $\mathcal{R}^*$ (left) and subjective model $\hat{\mathcal{R}}$ (right) in the performance evaluation application.

To illustrate this idea, we examine a simple model of RPE. Consider the objective model $\mathcal{R}^*$

in Figure 7a. The principal benefits from two outputs, the agent's output y_1 and her colleagues' output y_2; these are either low or high, $y_1, y_2 \in \{y_L, y_H\}$. Let $z \in \{-1, 0, 1\}$ be the agent's relative performance: It is good ($z = 1$) if the agent's output is larger than her colleagues' output, neutral ($z = 0$) in case of a tie, and negative ($z = -1$) if the agent's output is below her colleagues' output. The principal can condition the agent's wage on y_1 and z. The agent exerts effort $a \in \{0, 1\}$ and engages in sabotage $e \in \{0, 1\}$. Both efforts are costly. Sabotage ($e = 1$) ensures that the colleagues' output is low ($y_2 = y_L$).

The production function is as follows. The agent's effort a directly affects operational efficiency $x_1 \in \{0, 1\}$, $p(x_1 = 1 \mid a) = \beta_a^1$. With probability b there is a common shock that reduces the two outputs to y_L (this shock is not a node in $\mathcal{R}^*$ but incorporated through the links $0R^*3$ and $2R^*3$). The agent's output y_1 is high with probability 1 if and only if $x_1 = 1$ and there is no common shock; otherwise, it is low. Her colleagues' output y_2 is high with probability $\beta^2 \in (0, 1)$ if and only if there is no sabotage and no common shock; otherwise, it is low.[12]

Consider first the optimal incentive scheme $w(y_1, z)$ when the agent knows the objective model $\mathcal{R}^*$ and sabotage is not available. Standard arguments then show that the optimal incentive scheme which implements high effort, $w^*(y_1, z)$, uses RPE (see the Online Appendix). When output y_1 is high, the agent's wage is independent of z, $w^*(y_H, 1) = w^*(y_H, 0)$. However, if it is low, the agent's wage decreases in her colleagues' output, $w^*(y_L, 0) > w^*(y_L, -1)$; if both outputs are low, this indicates a common shock, in which case the punishment for low output is reduced. Naturally, this gives the agent an incentive to sabotage her colleagues' output. If sabotage is available to the agent and the costs of sabotage are small enough, $w^*(y_1, z)$ is no longer the optimal incentive scheme.

Next, assume that the principal can convince the agent that no relative performance evaluation takes place in the organization ("it does not fit our team culture"). Instead, there is a "subjective" evaluation of the agent's performance by an independent third party. The agent's pay can be tied to this evaluation. The corresponding subjective model is $\hat{\mathcal{R}}$ in Figure 7a. The agent perceives the performance evaluation z as being correlated with operational efficiency x_1 and her output y_1. Sabotage is perceived as worthless since it has no direct effect on x_1.

We show that if the agent's subjective model is given by $\hat{\mathcal{R}}$, then $w^*(y_1, z)$ is again the optimal incentive scheme that implements high effort. For this, we only need arguments from our Bayesian network methodology. If we fix $e = 0$, the true probability distribution factorizes according to model $\mathcal{R}^{[1]}$ in Figure 7b. By Proposition 2, $\mathcal{R}^{[1]}$ is equivalent to model $\mathcal{R}^{[2]}$ in Figure 7b, in which the colleagues' output has no effect on any of the other variables. Thus,

[12]Thus, the probability model is $p(x_1 = 1 \mid a) = \beta_a^1$, $p(y_2 = y_H \mid e = 0) = (1 - b)\beta^2$, $p(y_2 = y_H \mid e = 1) = 0$, $p(y_1 = y_H \mid x_1 = 1, y_2 = y_H, e = 0) = 1$, $p(y_1 = y_H \mid x_1 = 1, y_2 = y_L, e = 0) = \frac{(1-b)(1-\beta^2)}{(1-\beta^2)+b\beta^2}$, $p(y_1 = y_H \mid x_1 = 0, y_2, e) = 0$, $p(y_1 = y_H \mid x_1 = 1, y_2, e = 1) = 1 - b$. Note that the link between y_1 and y_2 captures the correlation between the two outputs induced by the common shock.

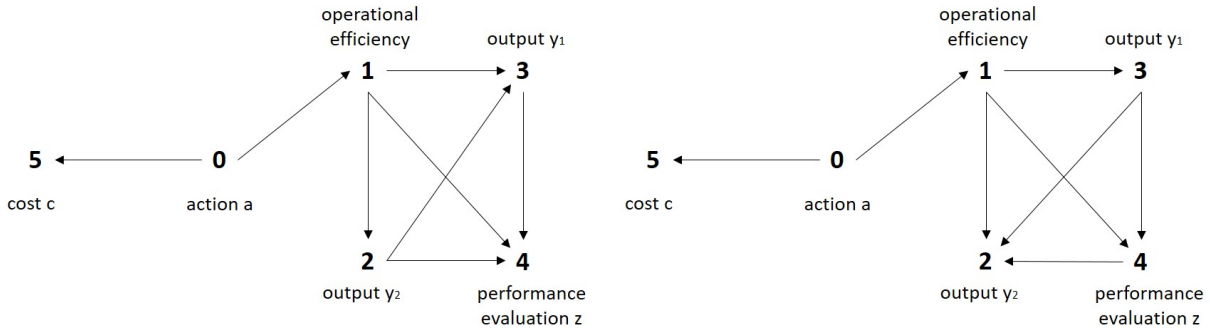

Figure 7b: Model $\mathcal{R}^{[1]}$ (left) and model $\mathcal{R}^{[2]}$ (right).

the agent with subjective model $\hat{\mathcal{R}}$ is behaviorally rational, provided that we keep $e = 0$ fixed. Consequently, $w^*(y_1, z)$ is optimal if the principal wants to implement high effort. In the corresponding personal equilibrium, the agent believes that the evaluation is related to operational efficiency (more precisely, z is *not* independent from x_1 conditional on y_1), while in the objective model z is independent from x_1 conditional on the two outputs y_1 and y_2. Incentives appear to the agent as being provided through subjective performance evaluation where the evaluator receives a private, noisy signal of her performance (as in Levin 2003).

The implication of this result is that RPE is more common than suggested by theory. In particular, its camouflaged version could also be applied in environments in which open RPE would have negative consequences for employees' collaboration. Thus, some applications of RPE are potentially hard to detect empirically since organizations are inclined to hide it.

5.5 Contract Design

An important question in contract theory is on which information the principal should condition the agent's wage. Holmström's (1979) sufficient statistic result states that all variables that are informative about the agent's effort should be used in the contract. In contrast, many real-world contracts appear as being incomplete relative to this benchmark. We show that there can be a trade-off between using more information in the contract and taking advantage of the agent's misspecification.

Consider the objective model $\mathcal{R}^*$ in Figure 8. The agent's job is to increase output y. To increase output, she has to increase operational efficiency x_1, which requires her to reprimand her subordinates from time to time. This in turn has a negative effect on employees' morale in the organization x_2. Morale has a positive effect on sales y, and a negative effect on turnover in the organization z. Both output y and turnover z are contractible. The principal's payoff is independent of z. If the principal conditions the agent's wage on both output y and turnover z, the agent takes her subordinates' mood into account when making decisions so that her model

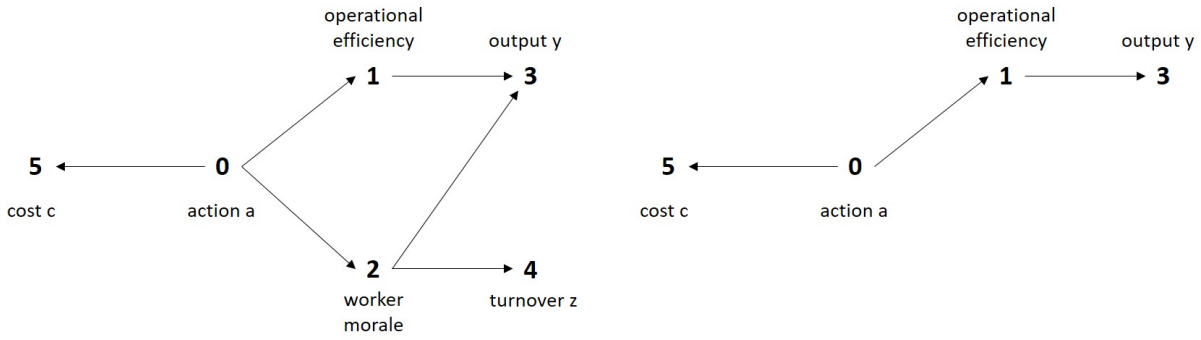

Figure 8: Objective model $\mathcal{R}^*$ (left) and the agent's subjective model $\hat{\mathcal{R}}$ (right) in the contractual completeness application.

equals $\mathcal{R}^*$.[13] Alternatively, the principal can keep the agent's model misspecified so that she ignores the effect of her effort on workers' morale. Her subjective model is then given by $\hat{\mathcal{R}}$ in Figure 8. In this case, the principal can condition the agent's wage only on output y.

To study the optimal equilibrium contract, we again use the setup from the production function example from Section 3.2 with binary effort $a \in \{0, 1\}$, $p(x_i = 1 \mid a) = \beta_a^i$ and $p(y_H \mid x_1, x_2) = \gamma_{x_1,x_2}$; in line with the story, we have $\beta_1^1 > \beta_0^1$, $\beta_1^2 < \beta_0^2$, and γ_{x_1,x_2} increases in both arguments. Turnover is high or low, $z \in \{z_H, z_L\}$, and we denote $p(z_H \mid x_2) = \xi_{x_2} \in (0, 1)$.

Suppose the principal wishes to implement high effort. If she conditions the agent's wage on turnover z, the agent's model becomes $\mathcal{R}^*$. The optimal incentive scheme $w^*(y, z)$ is then characterized by the first-order condition

$$\frac{1}{u'(w(y,z))} = \lambda + \mu\left[1 - \frac{p(y,z \mid a=0)}{p(y,z \mid a=1)}\right], \tag{17}$$

where λ and μ are positive constants. If $\xi_1 \neq \xi_0$, turnover z is informative about the agent's action so that indeed $w^*(y, z)$ varies in z. Next, suppose the principal keeps the agent's model misspecified so that the wage only varies in output y. The optimal incentive scheme $w^*(y)$ is then characterized by the first-order condition

$$\frac{1}{u'(w(y))} = \hat{\lambda} + \hat{\mu}\left[1 - \frac{p_{\hat{\mathcal{R}}}(y \mid a=0; \alpha=1)}{p_{\hat{\mathcal{R}}}(y \mid a=1; \alpha=1)}\right], \tag{18}$$

where $\hat{\lambda}$ and $\hat{\mu}$ are again positive constants. As in the marketer example, the *IC* is relaxed relative to the rational agent benchmark if the agent's model is misspecified. Thus, there is a trade-off between using more information and exploiting the agent's misspecification. If the informational content of turnover z is sufficiently small, i.e., if, all else equal, ξ_0 is close enough to ξ_1, the optimal contract is $(\hat{\mathcal{R}}, w^*(y), a = 1)$ so that it keeps the agent's model misspecified.

[13] Friebel et al. (2018) study such a case in a field-experiment with a large retail-chain.

In contrast, if the the importance of morale on final output small enough, i.e., if, all else equal, $\gamma_{x_1,0}$ is close enough to $\gamma_{x_1,1}$ for $x_1 \in \{0,1\}$, the optimal contract is $(\mathcal{R}^*, w^*(y,z), a = 1)$ and conditions on both output measures.

Auster (2013) captures a similar trade-off in a model with unawareness. In her framework, the agent is unaware of some outcomes $y' \in Y$. The principal's benefit from keeping the agent unaware of outcome y' is that he can then lower the wage $w(y')$ without affecting the participation constraint. The cost of not mentioning y' in the contract is that the information generated through y' cannot be used to tie the agent's wage to her effort. Thus, we show that a very similar trade-off can occur even when the agent is aware of all potential outcomes.

5.6 Further Applications

There exist many other design aspects of organizations that potentially matter for the agent's subjective model. For example, one further important aspect is the allocation of decision rights (e.g., Aghion and Tirole 1997). If information is dispersed in the organization, the extent to which this information is gathered or shared depends on which party has the formal authority to make binding decisions. There may exist a variety of opportunities to profitably keep another party's subjective model of the organization misspecified. In the Online Appendix, we offer an example. We adjust our framework so that it captures an incomplete contracting model in which the allocation of decision rights affects the agent's beliefs. Another design aspect becomes relevant when the principal does not exactly know the agent's subjective model. The question is then how he screens between different agent types who differ in their subjective model. In the Online Appendix, we consider a simple screening setting and generalize an important comparative static result from von Thadden and Zhao (2012) to our framework.

6 Conclusion

In this paper, we analyzed optimal contracting and organization when the agent has a misspecified model of the principal's project. To capture model misspecifications in a parsimonious manner, we applied Spiegler's (2016) Bayesian network framework to some classic settings in organizational economics. If the agent's subjective model can be represented by a perfect directed acyclic graph, she avoids neglect of correlation and has correct expectations on the equilibrium path. However, through the misspecification, the agent may incorrectly extrapolate how off-equilibrium actions map into outcomes, which affects the incentive compatibility constraint. Thus, the principal can strictly benefit from the agent's misperception, while the agent's model is validated by the data that she collects in equilibrium. Using results from the

Bayesian network literature, we characterized which misspecifications change the contracting problem, and under which circumstances different misperceptions have the same effect on the agent's incentives. We endogenized the misspecification in the agent's model by assuming that the principal can partially pick the project components that the agent (does not) take into account. This established a link between the organization of the agent's project and her subjective beliefs. Specifically, we showed that the option to keep the agent's model misspecified may result in technological inertia, workplace transparency, narcissistic leadership, camouflaged relative performance evaluation, and incomplete contracts.

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.

AUSTER, SARAH, AND NICOLA PAVONI (2018): "Optimal Delegation and Limited Awareness," Working Paper, Bocconi University.

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

BHASKAR, V., AND CAROLINE THOMAS (2019): "The Culture of Overconfidence," *American Economic Review: Insights*, forthcoming.

BLAKE, THOMAS, CHRIS NOSKO, AND STEVEN TADELIS (2015): "Consumer heterogeneity and paid search effectiveness: A large-scale field experiment," *Econometrica*, 83(1), 155–174.

BLOOM, NICHOLAS, BENN EIFERT, APRAJIT MAHAJAN, DAVID MCKENZIE, AND JOHN ROBERTS (2013): "Does Management Matter? Evidence from India," *Quarterly Journal of Economics*, 128(1), 1–51.

BURSZTYN, LEONARDO, ALESSANDRA GONZÁLEZ, AND DAVID YANAGIZAWA-DROTT (2018): "Misperceived Social Norms: Female Labor Force Participation in Saudi Arabia," Working Paper.

COWELL, ROBERT, A. PHILIP DAWID, STEFFEN LAURITZEN, AND DAVID SPIEGELHALTER (2007): *Probabilistic Networks and Expert Systems*, Springer, New York.

CYERT, RICHARD, AND JAMES MARCH (1963): *A Behavioral Theory of the Firm*, Englewood Cliffs, NJ: Prentice-Hall.

De la Rosa, Enrique (2011): "Overconfidence and Moral Hazard," *Games and Economic Behavior*, 73(2), 429–451.

Drago, Robert, and Gerald Garvey (1998): "Incentives for Helping on the Job: Theory and Evidence," *Journal of Labor Economics*, 16(1), 1–25.

Eliaz, Kfir, and Rani Spiegler (2018): "A Model of Competing Narratives," CEPR Discussion Paper No. DP13319.

Eliaz, Kfir, Rani Spiegler, and Heidi Thysen (2018): "Strategic Interpretations," Working Paper.

Fang, Hanming, and Giuseppe Moscarini (2005): "Morale Hazard," *Journal of Monetary Economics*, 52(4), 749–777.

Farzaneh-Far, Ramin, Jue Lin, Elissa Epel, William Harris, Elizabeth Blackburn, and Mary Whooley (2010): "Association of Marine Omega-3 Fatty Acid Levels With Telomeric Aging in Patients With Coronary Heart Disease," *Journal of the American Medical Association*, 303(3), 250–257.

Filiz-Ozbay, Emel (2012): "Incorporating unawareness into contract theory," *Games and Economic Behavior*, 76(1), 181–1941.

Friebel, Guido, Matthias Heinz, and Nick Zubanov (2018): "Making Managers Matter," Working Paper, Goethe-University Frankfurt.

Gervais, Simon, and Itay Goldstein (2007): "The Positive Effects of Biased Self-Perceptions in Firms," *Review of Finance*, 11(3), 453–496.

Gervais, Simon, J. B. Heaton, and Terrance Odean (2011): "Overconfidence, Compensation Contracts, and Capital Budgeting," *Journal of Finance*, 66(5), 1735–1777.

Grijalva, Emily, Peter Harms, Daniel Newman, Blaine Gaddis, and R. Chris Fraley (2015): "Narcissism and leadership: A meta-analytic review of linear and nonlinear relationships," *Personnel Psychology*, 68(1), 1–47.

Grossman, Sanford, and Oliver Hart (1983): "An Analysis of the Principal-Agent Problem," *Econometrica*, 51(1), 7–45.

Hanna, Rema, Sendhil Mullainathan, Joshua Schwartzstein (2014): "Learning Through Noticing: Theory and Evidence from a Field Experiment," *Quarterly Journal of Economics*, 129(3), 1311–1353.

HANNAN, MICHAEL, AND JOHN FREEMAN (1984): "Structural Inertia and Organizational Change," *American Sociological Review*, 49(2), 149–164.

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

HELLER, SARA, ANUJ SHAH, JONATHAN GURYAN, JENS LUDWIG, SENDHIL MULLAINATHAN, AND HAROLD POLLACK (2017): "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago," *Quarterly Journal of Economics*, 132(1), 1–54.

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

JUDGE, TIMOTHY, JOYCE BONO, REMUS ILIES, AND MEGAN GERHARD (2002): "Personality and Leadership: A Qualitative and Quantitative Review," *Journal of Applied Psychology*, 87(5), 765–780.

KOSKI, TIMO, AND JOHN NOBLE (2009): *Bayesian Networks: An Introduction*, Wiley Series in Probability, Wiley.

LARCOM, SHAUN, FERDINAND RAUCH, AND TIM WILLEMS (2019): "The Benefits of Forced Experimentation: Striking Evidence from the London Underground Network," *Quarterly Journal of Economics*, forthcoming.

LAZEAR, EDWARD (1989): "Pay Equality and Industrial Politics," *Journal of Political Economy*, 97(3), 561–580.

LEVIN, JONATHAN (2003): "Relational Incentive Contracts," *American Economic Review*, 93(3), 835–857.

MARCH, JAMES, HERBERT SIMON (1958): *Organizations*, Wiley, New York.

MIKLÓS-THAL, JEANINE, AND JUANJUAN ZHANG (2013): "(De)marketing to Manage Consumer Quality Inferences," *Journal of Marketing Research*, 50(1), 55–69.

MILGROM, PAUL (1988): "Employment Contracts, Influence Activities, and Efficient Organization Design," *Journal of Political Economy*, 96(1), 42–60.

MILGROM, PAUL, AND ILYA SEGAL (2002): "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70(2), 583–601.

MURPHY, KEVIN (2001): "Performance standards in incentive contracts," *Journal of Accounting and Economics*, 30(3), 245–278.

NEVICKA, BARBORA, FEMKE TEN VELDEN, ANNEBEL DE HOOGH, AND ANNELIES VAN VIANEN (2011): "Reality at Odds With Perceptions: Narcissistic Leaders and Group Performance," *Psychological Science*, 22(10), 1259–1264.

NULAND, SHERWIN (2004): *The Doctors' Plague: Germs, Childbed Fever, and the Strange Story of Ignac Semmelweis*, W. W. Norton Company.

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

ROSENTHAL, SETH, AND TODD PITTINSKY (2006): "Narcissistic Leadership," *The Leadership Quarterly*, 17(6), 617–633.

ROTEMBERG, JULIO, AND GARTH SALONER (2000): "Visionaries, Managers, and Strategic Direction," *RAND Journal of Economics*, 31(4), 693–716.

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

SHRIER, IAN, AND ROBERT PLATT (2008): "Reducing bias through directed acyclic graphs," *BMC Medical Research Methodology*, 8(70).

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 (2018): "Can Agents with Causal Misperceptions be Systematically Fooled?," *Journal of the European Economic Association*, forthcoming.

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.

WATTS, ASHLEY, SCOTT LILIENFELD, SARAH FRANCIS SMITH, JOSHUA MILLER, W. KEITH CAMPBELL, IRWIN WALDMAN, STEVEN RUBENZER, THOMAS FASCHINGBAUER (2013): "The Double-Edged Sword of Grandiose Narcissism: Implications for Successful and Unsuccessful Leadership Among U.S. Presidents," *Psychological Science*, 24(12), 2379–2389.

WINTER, EYAL (2010): "Transparency and incentives among peers," *RAND Journal of Economics*, 41(3), 504–523.

A Online Appendix

A.1 Existence of a Personal Equilibrium

We show that a personal equilibrium exists at any $\mathcal{R}$ 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

$$BR(p(a)) = \arg \max_{p'(a') \in \Delta(A)} \sum_{a' \in A} \sum_{y \in Y} \sum_{c \in C} p'(a') p_{\mathcal{R}}(y, c \mid a'; p(a))(u(w(y)) - c). \tag{19}$$

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 $p_{\mathcal{R}}(y, c \mid a'; p(a))$ are continuous in $p(a)$. Therefore, we also must have that $\sum_{a' \in A} \sum_{y \in Y} \sum_{c \in C} p'(a') p_{\mathcal{R}}(y, c \mid a'; p(a))(u(w(y)) - c)$ 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 Mathematical Details of Subsection 3.2

We derive $p_{\mathcal{R}}(y = 1 \mid a; \alpha)$ from the agent's subjective model $\mathcal{R}$ and the objective probability distribution. The equilibrium definition implies that $p_{\mathcal{R}}(x_1 \mid a; \alpha) = p(x_1 \mid a)$. So it remains to derive $p_{\mathcal{R}}(y \mid x_1; \alpha)$. Note that

$$\begin{aligned} p_{\mathcal{R}}(x_2 = 1 \mid x_1 = 1; \alpha) &= \frac{\alpha \beta_1^1 \beta_1^2 + (1-\alpha)\beta_0^1 \beta_0^2}{\alpha \beta_1^1 + (1-\alpha)\beta_0^1}, \\ p_{\mathcal{R}}(x_2 = 1 \mid x_1 = 0; \alpha) &= \frac{\alpha(1-\beta_1^1)\beta_1^2 + (1-\alpha)(1-\beta_0^1)\beta_0^2}{\alpha(1-\beta_1^1) + (1-\alpha)(1-\beta_0^1)}. \end{aligned}$$

With this we can calculate the equilibrium probability that the high output y_H realizes given that $x_1 = 1$ and $x_1 = 0$, respectively:

$$p_{\mathcal{R}}(y_H \mid x_1 = 1; \alpha) = \frac{\alpha \beta_1^1 \beta_1^2 + (1-\alpha)\beta_0^1 \beta_0^2}{\alpha \beta_1^1 + (1-\alpha)\beta_0^1} \gamma_{1,1} + \left(1 - \frac{\alpha \beta_1^1 \beta_1^2 + (1-\alpha)\beta_0^1 \beta_0^2}{\alpha \beta_1^1 + (1-\alpha)\beta_0^1}\right) \gamma_{1,0}$$

and

$$p_{\mathcal{R}}(y_H \mid x_1 = 0; \alpha) = \frac{\alpha(1-\beta_1^1)\beta_1^2 + (1-\alpha)(1-\beta_0^1)\beta_0^2}{\alpha(1-\beta_1^1) + (1-\alpha)(1-\beta_0^1)} \gamma_{0,1} + \left(1 - \frac{\alpha(1-\beta_1^1)\beta_1^2 + (1-\alpha)(1-\beta_0^1)\beta_0^2}{\alpha(1-\beta_1^1) + (1-\alpha)(1-\beta_0^1)}\right) \gamma_{0,0}.$$

Thus, we get

$$\begin{aligned} p_{\mathcal{R}}(y_H \mid a=1;\alpha) &= \beta_1^1\left(\gamma_{1,0}+\frac{\alpha\beta_1^1\beta_1^2+(1-\alpha)\beta_0^1\beta_0^2}{\alpha\beta_1^1+(1-\alpha)\beta_0^1}(\gamma_{1,1}-\gamma_{1,0})\right) \\ &\quad +(1-\beta_1^1)\left(\gamma_{0,0}+\frac{\alpha(1-\beta_1^1)\beta_1^2+(1-\alpha)(1-\beta_0^1)\beta_0^2}{\alpha(1-\beta_1^1)+(1-\alpha)(1-\beta_0^1)}(\gamma_{0,1}-\gamma_{0,0})\right) \end{aligned}$$

and

$$\begin{aligned} p_{\mathcal{R}}(y_H \mid a=0;\alpha) &= \beta_0^1\left(\gamma_{1,0}+\frac{\alpha\beta_1^1\beta_1^2+(1-\alpha)\beta_0^1\beta_0^2}{\alpha\beta_1^1+(1-\alpha)\beta_0^1}(\gamma_{1,1}-\gamma_{1,0})\right) \\ &\quad +(1-\beta_0^1)\left(\gamma_{0,0}+\frac{\alpha(1-\beta_1^1)\beta_1^2+(1-\alpha)(1-\beta_0^1)\beta_0^2}{\alpha(1-\beta_1^1)+(1-\alpha)(1-\beta_0^1)}(\gamma_{0,1}-\gamma_{0,0})\right). \end{aligned}$$

We now can use these terms to compute the incentive compatibility constraint *IC*.

A.3 Mathematical Details of Subsection 3.3

We examine the setting for the misspecification in the production function example from Section 3.2. Consider the first term from the *IC*,

$$\Delta(\alpha) = p_{\mathcal{R}}(y_H \mid a=1;\alpha) - p_{\mathcal{R}}(y_H \mid a=0;\alpha). \tag{20}$$

If $\Delta(\alpha)$ weakly increases in α, the *IC* is relaxed so that the principal can implement a higher probability of effort without changing incentives. If $\Delta(\alpha)$ weakly increases in α at all $\alpha \in [0,1]$, there is an optimal equilibrium contract that implements a pure action, $\alpha = 0$ or $\alpha = 1$. From equation (5) we get that $\Delta'(\alpha) \geq 0$ at all $\alpha \in [0,1]$ if

$$\frac{\beta_0^1\beta_1^1(\beta_1^2-\beta_0^2)}{[\alpha\beta_1^1+(1-\alpha)\beta_0^1]^2}(\gamma_{1,1}-\gamma_{1,0}) - \frac{(1-\beta_0^1)(1-\beta_1^1)(\beta_1^2-\beta_0^2)}{[\alpha(1-\beta_1^1)+(1-\alpha)(1-\beta_0^1)]^2}(\gamma_{0,1}-\gamma_{0,0}) \geq 0 \tag{21}$$

at all $\alpha \in [0,1]$. This condition is satisfied, for example, if $\beta_1^2 < \beta_0^2$ and $\beta_0^1 = 0$ or if $\beta_1^2 > \beta_0^2$ and $\gamma_{1,1} - \gamma_{1,0}$ is large enough relative to $\gamma_{0,1} - \gamma_{0,0}$.

A.4 Proof of Proposition 3 and 4

We first prove Proposition 4 and then Proposition 3. 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 4 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 for the proof of the "only if"-statement of Proposition 4 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 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 (2017b) 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 4. 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 3, we need two more results. We will use the following definitions. 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 $M \subset N^* \setminus H^*$ be a set of nodes connected through non-fundamental links. Suppose there are two nodes $i, j \in H^*$ with non-fundamental links to nodes in M. Then i and j are adjacent.*

Proof. Assume w.l.o.g. that i, j are on a fundamental active path between 0 and n (the argument for $n+1$ is identical). 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 $k \in H^*$ 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 $i, j \in H^*$ 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 3. It shows that all nodes that are not on a fundamental active path between action and outcome nodes can be made "unimportant" in the sense that they have no impact on outcomes. Formally, this means that we can find a DAG in $\mathcal{E}$ in which all links between one node in H^* and one node in $N^* \setminus H^*$ point towards the node in $N^* \setminus H^*$.

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

Proof. The proof proceeds by steps. **Step 1.** Consider any maximal set $M \subset N^* \setminus H^*$ of nodes connected through non-fundamental links and let $M^+ \subset H^*$ 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 $N_d \cap H^*$ 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 (2017b) 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 $M \subset N^* \setminus H^*$ of nodes connected by non-fundamental links in $\mathcal{R}^*$. □

Proof of Proposition 3. First, we show the "if"-statement. Assume that the agent's subjective DAG $\mathcal{R}$ is aware of all the nodes in H^*. Consider the DAG $\mathcal{G}^* \in \mathcal{E}$ in which all links with one end in H^* and the other in $N^* \setminus H^*$ point from H^* to $N^* \setminus H^*$. By Lemma 6, this DAG exists. From Proposition 2 it follows that $p_{\mathcal{G}^*}(x_{H^*}) = p(x_{H^*})$ for all distributions $p \in \Delta(X)$. Consider the subgraph $\mathcal{G} = (G, N)$ where G equals G^* restricted on N. Since none of the nodes in $N \setminus H^*$ impacts on any node in H^*, we have $p_{\mathcal{G}}(x_{H^*}) = p_{\mathcal{G}^*}(x_{H^*})$ for all $p \in \Delta(X)$. By construction, the DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent so that we have $p_{\mathcal{R}}(x_{H^*}) = p_{\mathcal{G}}(x_{H^*}) = p_{\mathcal{G}^*}(x_{H^*}) = p(x_{H^*})$ for all distributions $p \in \Delta(X)$, which proves the "if"-statement. Next, we show the "only if"-statement. Assume that there is one node $i \in H^*$ that is not in the agent's subjective model. Assume w.l.o.g. that i is on a fundamental active path τ between the action node 0 and the output node n. We find a probability distribution $p \in \Delta(X)$ 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. □

A.5 Imperfect Objective DAGs

Proposition 3 characterizes for perfect objective DAGs $\mathcal{R}^*$ which nodes must be in the agent's subjective model $\mathcal{R}$ so that the agent is behaviorally rational. We can partially extend Proposition 3 to imperfect objective DAGs $\mathcal{R}^*$. Note that one can make any imperfect DAG perfect by adding links between nodes that create v-colliders. If p is consistent with $\mathcal{R}^*$, it is consistent with any DAG that adds links to $\mathcal{R}^*$. If all added links disappear after taking out the nodes of interest, we can again use Proposition 3 to determine whether the agent is behaviorally rational under her subjective DAG $\mathcal{R}$.

We state this result formally. Let an imperfect DAG $\mathcal{R}^*$ and the agent's subjective DAG $\mathcal{R}$ be given. Define by $\mathcal{R}^*_m(N)$ a perfect DAG that is identical to $\mathcal{R}^*$ except that it contains additional links which have at least one end in the set of nodes $N^* \setminus N$. By construction, these links are not in the subjective DAG $\mathcal{R}$. We then define

$$H^*_m(N) = \{i \in N^* \mid i \text{ is part of a fundamental active path between 0 and } n \text{ or } n+1 \text{ in } \mathcal{R}^*_m(N)\}.$$

From Proposition 3 we immediately get that the agent is behaviorally rational if and only if N contains $H^*_m(N)$. Thus, we get the following result.

Corollary 3. *Let $\mathcal{R}^* = (N^*, R^*)$ be the (possibly imperfect) objective DAG and $\mathcal{R} = (N, R)$ the agent's subjective DAG with $N \subseteq N^*$. Suppose there exists a perfect DAG $\mathcal{R}^*_m(N) = (N^*, R^*_m(N))$ where $R^*_m(N)$ differs from R^* only in that $R^*_m(N)$ contains additional links with at least one end in $N^* \setminus N$. Then the agent is behaviorally rational if and only if her subjective DAG $\mathcal{R}$ contains all nodes from $H^*_m(N)$.*

We illustrate this result with our example from Section 3.2. Consider the objective DAG $\mathcal{R}^*$ and the subjective DAG $\mathcal{R}$ from Figure 2a. $\mathcal{R}^*$ is imperfect since it has a v-collider at node 3, so we cannot apply Proposition 3. We can make $\mathcal{R}^*$ perfect by adding a link from node 1 to node 2. This new DAG has the properties requested by the corollary, i.e., if we take out node 2, the additional link disappears and we again have the agent's subjective DAG $\mathcal{R}$. Now observe that in $\mathcal{R}^*_m(N)$ node 2 is on a fundamental active path between the action and the output node. Thus, the agent with subjective DAG $\mathcal{R}$ is not behaviorally rational.

A.6 Justifiability

In our framework, the agent has a fully specified model that makes predictions about outcomes for all actions $a \in A$. A natural question is then whether the optimal equilibrium contract is also optimal for the principal when evaluated from the agent's (potentially biased) perspective. If according to her subjective beliefs the principal should have offered an alternative contract, the agent may suspect that her subjective model $\mathcal{R}$ is not correct. This refinement is called "justifiability." It has first been defined in the unawareness literature, see Filiz-Ozbay (2012) and Heifetz et al. (2013). We can easily adapt it to our framework. The following definition captures "justifiability" and "partial justifiability" as additional refinements.

Definition 6. *An equilibrium contract $(\mathcal{R}, w^*(y), p^*(a))$ is justifiable if $w^*(y), p^*(a)$ is a solution*

to the maximization problem

$$\max_{w(y)\in W, p(a)\in\Delta(A)} \sum_{a\in A}\sum_{y\in Y} p(a)p_{\mathcal{R}}(y \mid a; p^*(a))(y - w(y))$$

subject to the constraints that for all a in the support of $p(a)$

$$a \in \arg\max \sum_{y\in Y}\sum_{c\in C} p_{\mathcal{R}}(y, c \mid a; p^*(a))(u(w(y)) - c),$$
$$\sum_{a\in A}\sum_{y\in Y}\sum_{c\in C} p(a)p_{\mathcal{R}}(y, c \mid a; p^*(a))(u(w(y)) - c) \geq \bar{U}.$$

An equilibrium contract $(\mathcal{R}, w^*(y), p^*(a))$ *is partially justifiable if* $w^*(y)$ *is a solution to this maximization problem when* $p(a) = p^*(a)$ *is given.*

An equilibrium contract $(\mathcal{R}, w(y), p(a))$ is justifiable if the choice of incentive scheme $w(y)$ and implemented action $p(a)$ maximizes the principal's expected payoff when evaluated according to the agent's beliefs $p_{\mathcal{R}}(y, c \mid a; p(a))$. This contract is partially justifiable if at least the incentive scheme $w(y)$ maximizes the principal's expected payoff, when evaluated according to the agent's beliefs, given that the principal wants to implemented action $p(a)$.

The optimality of an equilibrium contract can guarantee partial justifiability. Observe that the principal's maximization problem in (3) and the maximization problem in definition above are identical when $\mathcal{R}$ is perfect and we take the agent's subjective model $\mathcal{R}$ as well as the implemented action a^* as given. In this case, both the optimal equilibrium contract and the contract that is optimal according to the agent's beliefs are identical: both are defined by the objective equilibrium probabilities $p(y \mid a^*)$ in the principal's objective function and the agent's beliefs $p_{\mathcal{R}}(y, c \mid a; a^*)$ in the incentive compatibility and participation constraint.

To prove justifiability we have to rule out that, according to the agent's beliefs, the principal can benefit by implementing a different action. We can show that for several important special cases – namely, binary action and output sets – the optimality of an equilibrium contract implies justifiability. The next proposition summarizes our findings.

Proposition 5 (Justifiability)**.** *Let* $(\mathcal{R}, w^*(y), a^*)$ *be an optimal equilibrium contract that implements a pure action* a^* *and let* $\mathcal{R}$ *be perfect. Then the following holds:*

(a) This contract is partially justifiable.

(b) Suppose that A is a binary set and the principal strictly prefers $(\mathcal{R}, w^*(y), a^*)$ *to the optimal contract under the objective model* $\mathcal{R}^*$*. If* $\mathcal{R}$ *has a misspecification only in the cost function, or if* $\mathcal{R}$ *has a misspecification only in the production function and Y is a binary set, then this contract is justifiable.*

Proof. Statement (i) is proven by the arguments provided above. We prove statement (ii). Assume first that $\mathcal{R}$ has a misspecification only in the cost function. We denote $A = \{0, 1\}$ with the usual interpretation. Since $\mathcal{R}$ is perfect and the principal strictly prefers $(\mathcal{R}, w^*(y), p^*(a))$ to the optimal contract under the objective model $\mathcal{R}^*$, the equilibrium action is $a^* = 1$. We show that from the agent's perspective the principal cannot gain by implementing $a = 0$. By the principal's preference, incentive compatibility must be relaxed so that we have

$$\sum_{c \in C} p_{\mathcal{R}}(c \mid a = 0; a^*)c > \sum_{c \in C} p(c \mid a = 0)c. \tag{22}$$

We therefore have

$$\begin{aligned}
\sum_{y \in Y} p_{\mathcal{R}}(y \mid a = 1; a^*)(y - w^*(y)) &= \sum_{y \in Y} p(y \mid a = 1)(y - w^*(y)) \\
&> \sum_{y \in Y} p(y \mid a = 0)(y - u^{-1}(\sum_{c \in C} p(c \mid a = 0) + \bar{U})) \\
&> \sum_{y \in Y} p(y \mid a = 0)(y - u^{-1}(\sum_{c \in C} p_{\mathcal{R}}(c \mid a = 0; a^*) + \bar{U})),
\end{aligned}$$

which proves the result. Next, assume that $\mathcal{R}$ has a misspecification only in the production function and Y is a binary set. We denote $Y = \{0, 1\}$ with the usual interpretation. Since $\mathcal{R}$ is perfect and he principal strictly prefers $(\mathcal{R}, w^*(y), p^*(a))$ to the optimal contract under the objective model $\mathcal{R}^*$, the equilibrium action is $a^* = 1$ and $w_1^* > w_0^*$. Again, we show that from the agent's perspective the principal cannot gain by implementing $a = 0$. Denote by $\bar{w}$ the fixed wage that implements $a = 0$ at lowest costs (there is only a misspecification in the production function, so this value is the same under $\mathcal{R}$ and $\mathcal{R}^*$). The misspecification relaxes incentive compatibility. Since Y is binary we have $p(y = 1 \mid a = 0) > p_{\mathcal{R}}(y = 1 \mid a = 0; a^*)$. We therefore get

$$\begin{aligned}
\sum_{y \in Y} p_{\mathcal{R}}(y \mid a = 1; a^*)(y - w^*(y)) &= \sum_{y \in Y} p(y \mid a = 1)(y - w^*(y)) \\
&> \sum_{y \in Y} p(y \mid a = 0)(y - \bar{w}) \\
&> \sum_{y \in Y} p_{\mathcal{R}}(y \mid a = 0; a^*)(y - \bar{w}),
\end{aligned}$$

which proves the result. This completes the proof of statement (ii). □

A.7 Mathematical Details of Subsection 5.2

Since the worker's action is part of the probability model, we have to extend the definition of the equilibrium contract from Section 2. We first present this updated definition and then show by example that a misspecification in the agent's model may cause the principal to implement a transparent instead of a non-transparent organization.

Extended Equilibrium Contract. In the non-transparent organization, an extended contract $\{p(e), \bar{w}(y), (\mathcal{R}, w(y), p(a))\}$ is an extended equilibrium contract if (i) effort $p(e)$ is optimal for the worker given the agent's action $p(a)$ and her wage $\bar{w}(y)$, and (ii) $p(a)$ is a personal equilibrium at $\mathcal{R}$ and $w(y)$, given that the worker exerts effort $p(e)$. For the transparent setting, we only have to replace $p(e)$ by $p(e \mid a)$ in this definition.

Example. We define the probability model. The agent chooses action $a \in \{0, 1\}$ and the worker chooses effort $e \in \{0, 1\}$. The set of customers can be small or large, $x_2 \in \{0, 1\}$; the firm's reputation can be bad or good, $x_3 \in \{0, 1\}$; sales can be low or high, $y \in \{0, 1\}$; costs can be low or high $c \in \{0, 1\}$. For convenience, we abbreviate $c_A = \mathbb{E}[c \mid a = 1] - \mathbb{E}[c \mid a = 0]$ and $c_E = \mathbb{E}[c \mid e = 1] - \mathbb{E}[c \mid e = 0]$. Agent and worker are risk-neutral, protected by limited liability, and the value of their outside option is $\bar{U} = 0$. Let $w(y)$ ($\bar{w}(y)$) be the agent's (worker's) wage after outcome y.

Both customer base and reputation have a positive influence on sales, $p(y = 1 \mid x_2, x_3) = \beta_{24}x_2 + \beta_{34}x_3$. The agent's action and the worker's effort have a complementary effect on the customer base, $p(x_2 = 1 \mid a, e) = \beta_{012}(a + 1)(e + 1)$. Making cold-calls has a negative effect on firm reputation and the customer base has a positive effect on firm reputation, so we have $p(x_3 = 1 \mid e, x_2) = \beta_3 - \beta_{13}e + \beta_{23}x_2$. The worker always acts optimally given the incentives provided and the agent's equilibrium action (in case of a tie, agent 2 exerts high effort).

Suppose that the principal chooses a transparent structure and keeps the agent's model misspecified at $\mathcal{R}_{13}$. Assume that in equilibrium the agent chooses action $a = 1$ with probability $p(a = 1) = \alpha$. From the factorization formula we then get

$$p_{\mathcal{R}_{13}}(y = 1 \mid a; \alpha) = \sum_{x_2 \in X_2} p(x_2 \mid a) \sum_{e \in X_1} \sum_{x_3 \in X_3} p(e \mid x_2) p(x_3 \mid e, x_2) p(y = 1 \mid x_2, x_3), \tag{23}$$

where

$$p(e \mid x_2) = \frac{\sum_{a' \in A} p(a') p(e \mid a') p(x_2 \mid a', e)}{\sum_{e \in X_1} \sum_{a' \in A} p(a') p(e \mid a') p(x_2 \mid a', e)}. \tag{24}$$

Suppose that the worker exerts effort if and only the agent chooses action $a = 1$, i.e., $p(e \mid a) =$

a. Using equation (24) we can calculate

$$p(e = 1 \mid x_2 = 1) = \frac{\alpha 4\beta_{012}}{\alpha 4\beta_{012} + (1-\alpha)\beta_{012}} \tag{25}$$

and

$$p(e = 1 \mid x_2 = 0) = \frac{\alpha(1 - 4\beta_{012})}{\alpha(1 - 4\beta_{012}) + (1-\alpha)(1-\beta_{012})}. \tag{26}$$

Using the factorization in (23) we get

$$\begin{aligned} p_{\mathcal{R}_{13}}(y = 1 \mid a; \alpha) &= (\beta_{012} + 3a\beta_{012})[\beta_{24} + (\beta_3 + \beta_{23})\beta_{34} - \beta_{13}\beta_{34}p(e = 1 \mid x_2 = 1)] \\ &\quad +(1 - \beta_{012} - 3a\beta_{012})[\beta_3\beta_{34} - \beta_{13}\beta_{34}p(e = 1 \mid x_2 = 0)] \end{aligned} \tag{27}$$

The agent's subjective model then suggests that the effect of her action on sales

$$p_{\mathcal{R}_{13}}(y = 1 \mid a = 1; \alpha) - p_{\mathcal{R}_{13}}(y = 1 \mid a = 0; \alpha) \tag{28}$$

equals

$$3\beta_{012}\left[\beta_{24} + \beta_{23}\beta_{34} - \beta_{13}\beta_{34}\left(\frac{\alpha 4\beta_{012}}{\beta_{012} + 3\alpha\beta_{012}} - \frac{\alpha(1 - 4\beta_{012})}{1 - \beta_{012} - 3\alpha\beta_{012}}\right)\right]. \tag{29}$$

The term in round brackets strictly increases in α. Thus, the principal relaxes the incentive constraint by implementing $a = 1$ with higher probability.

Rational Agent. We examine how the principal's optimal wage schedule varies in the transparency structure when the agent is fully rational. The limited liability constraint implies that the optimal wage after low output is zero for both agents, $w(0) = \bar{w}(0) = 0$, regardless of the transparency structure, the agent's subjective model, and the efforts the principal wants to implement. If the agent chooses $a = 1$, the worker's effort has a positive effect on the expected output – despite its partial negative effect on firm reputation – when

$$2\beta_{012}(\beta_{24} + \beta_{23}\beta_{34}) - \beta_{13}\beta_{34} > 0. \tag{30}$$

In the following, we assume that this inequality is satisfied. Suppose that the agent's and worker's costs, c_A and c_E, are small enough such that the principal wishes to implement high effort from both employees. If he chooses the non-transparent structure $\mathcal{R}^*_{NT}$, the incentive constraint which ensures that the agent chooses $a = 1$ is then given by

$$2\beta_{012}(\beta_{24} + \beta_{23}\beta_{34})w(1) \geq c_A, \tag{31}$$

while the incentive constraint which ensures that the worker chooses $e = 1$ is

$$[2\beta_{012}(\beta_{24} + \beta_{23}\beta_{34}) - \beta_{13}\beta_{34}]\bar{w}(1) \geq c_E. \tag{32}$$

If the principal chooses the transparent structure $\mathcal{R}_T^*$, the incentive constraint for the worker remains the same, while for the agent it becomes

$$[3\beta_{012}(\beta_{24} + \beta_{23}\beta_{34}) - \beta_{13}\beta_{34}]w(1) \geq c_A. \tag{33}$$

Thus, it is cheaper to implement high effort from both employees under the non-transparent than under the transparent structure if

$$\beta_{012}(\beta_{24} + \beta_{23}\beta_{34}) - \beta_{13}\beta_{34} < 0. \tag{34}$$

Note that conditions (30) and (34) can be satisfied simultaneously. This means that it may be optimal for the principal to implement high effort from both employees *and* to choose the non-transparent structure. If the inequality in (34) is reversed, the production function is supermodular in the employees' efforts. The principal then optimally chooses a transparent structure when he wants both employees to exert high effort. Winter (2010) generalizes this statement to all supermodular production functions. In the following, we assume that both (30) and (34) are satisfied.

Agent with misspecified model. Next, we examine how the agent's optimal incentive scheme changes when subjective model is misspecified. Suppose again that the principal wants to implement high effort from both employees. As we showed in the main text, incentives do not change under the non-transparent structure. Thus, consider the transparent structure $\mathcal{R}_T^*$. The principal then optimally sets the worker's wage $\bar{w}(1)$ so that the incentive constraint in (32) is satisfied with equality. The worker then exerts high effort if and only if the agent chose $a = 1$. From (29) we get that the principal then implements $a = 1$ if

$$3\beta_{012}(\beta_{24} + \beta_{23}\beta_{34})w(1) \geq c_A. \tag{35}$$

Compare this constraint with that under the objective model in (33). Since the agent does not take into account the negative impact of the worker's effort on reputation, she overestimates the negative consequences from choosing $a = 0$ instead of $a = 1$, which in turn relaxes the incentive constraint. Next, compare this constraint with the incentive constraint under a non-transparent structure in (31). Through the misspecification, incentives are strictly larger under the transparent structure, even if (34) holds so that the production function is not supermodular.

A.8 Mathematical Details of Subsection 5.4

We derive the optimal incentive scheme $w^*(y_1, z)$ that implements high effort when the agent's model is the objective model $\mathcal{R}^*$. The agent's wage can take on the values $w(y_H, 1)$, $w(y_H, 0)$, $w(y_L, 0)$, or $w(y_L, -1)$. The optimal incentive scheme is characterized by the first-order conditions of the Lagrangian function

$$\begin{aligned}
L \;=\; & \beta_1^1(1-b)(1-\beta^2)[y_H + y_L - w(y_H,1)] + \beta_1^1(1-b)\beta^2[2y_H - w(y_H,0)] \\
& + ((1-\beta_1^1)(1-b)(1-\beta^2) + b)[2y_L - w(y_L,0)] + (1-\beta_1^1)(1-b)\beta^2[y_H + y_L - w(y_L,-1)] \\
& + \mu\,[\beta_1^1(1-b)(1-\beta^2)u(w(y_H,1)) + \beta_1^1(1-b)\beta^2 + u(w(y_H,0)) \\
& + ((1-\beta_1^1)(1-b)(1-\beta^2) + b)u(w(y_L,0)) + (1-\beta_1^1)(1-b)\beta^2 u(w(y_L,-1))] \\
& + \lambda\,[(\beta_1^1 - \beta_0^1)(1-b)(1-\beta^2)(u(w(y_H,1) - u(w(y_L,0)) \\
& + (\beta_1^1 - \beta_0^1)(1-b)\beta^2(u(w(y_H,0) - u(w(y_L,-1)) - \mathbb{E}[c \mid a=1] + \mathbb{E}[c \mid a=0]],
\end{aligned}$$

where μ and λ are the two Lagrange parameters for the *PC* and *IC*, respectively. The first-order conditions are

$$\begin{aligned}
\frac{1}{u'(w(y_H,1))} &= \mu + \lambda\frac{1}{\beta_1^1(1-b)}, \\
\frac{1}{u'(w(y_H,0))} &= \mu + \lambda\frac{1}{\beta_1^1(1-b)}, \\
\frac{1}{u'(w(y_L,0))} &= \mu - \lambda\frac{1-\beta^2}{(1-\beta_1^1)(1-b)(1-\beta^2)+b}, \\
\frac{1}{u'(w(y_L,-1))} &= \mu - \lambda\frac{1}{(1-\beta_1^1)(1-b)}.
\end{aligned}$$

Both constraints are binding in the optimum, so μ and λ are positive constants. Since u is concave, we get $w^*(y_H, 1) = w^*(y_H, 0) > w^*(y_L, 0) > w^*(y_L, -1)$.

A.9 Screening

Throughout the paper, we assumed that the principal knows the misspecification in the agent's subjective model and can tailor the contract to it. However, in many circumstances, he may not exactly know the agent's model. The question is then how the principal screens between different agent types who differ in their subjective model.

Let there be two agent types: the biased type 1 and the rational type 2. The biased type's default model $\hat{\mathcal{R}}$ is misspecified, while the rational type's default model equals the objective model $\mathcal{R}^*$. To screen between types, the principal offers two contracts, $(\mathcal{R}, w_1(y), p_1(a))$ and $(\mathcal{R}, w_2(y), a_2)$, where $\mathcal{R} \in \Gamma = \{\hat{\mathcal{R}}, \mathcal{R}^*\}$ must be identical in both contracts. If the principal

does not educate the agent, $\mathcal{R} = \hat{\mathcal{R}}$, the agent types keep their default model; if the principal educates the agents, $\mathcal{R} = \mathcal{R}^*$, both type's subjective model equals the objective model. Let λ be the chance that the agent is biased.

We characterize the optimal screening contract. If the principal educates the agents, we are back in the canonical principal-agent model and the optimal contract is given by a solution to (3) with $\mathcal{R} = \mathcal{R}^*$. Thus, suppose that the principal does not educate the agents. The optimal screening contract is then a solution to the problem

$$\max_{w_1(y),p_1(a),w_2(y),a_2} \lambda \sum_{a\in A}\sum_{y\in Y} p_1(a)p(y \mid a)(y - w_1(y)) + (1-\lambda)\sum_{y\in Y} p(y \mid a_2)(y - w_2(y)) \tag{36}$$

subject to the constraints

$$a \in \arg\max_{a'\in A} \mathbb{E}_{\hat{\mathcal{R}}}[u(w_1(y)) - c \mid a'; p_1(a)] \ \textit{for all} \ \ a \in \operatorname{supp}[p_1(a)], \tag{$IC1$}$$

$$a_2 \in \arg\max_{a'\in A} \mathbb{E}[u(w_2(y)) - c \mid a'], \tag{$IC2$}$$

$$\mathbb{E}_{\hat{\mathcal{R}}}[u(w_1(y)) - c \mid p_1(a); p_1(a)] \geq \bar{U}, \tag{$PC1$}$$

$$\mathbb{E}[u(w_2(y)) - c \mid a_2] \geq \bar{U}, \tag{$PC2$}$$

$$\mathbb{E}_{\hat{\mathcal{R}}}[u(w_1(y)) - c \mid p_1(a); p_1(a)] \geq \max_{a'\in A} \mathbb{E}_{\hat{\mathcal{R}}}[u(w_2(y)) - c \mid a'; p_1(a)], \tag{$IC12$}$$

$$\mathbb{E}[u(w_2(y)) - c \mid a_2] \geq \max_{a'\in A} \mathbb{E}[u(w_1(y)) - c \mid a']. \tag{$IC21$}$$

The constraints ($IC1$) and ($IC2$) ensure that both agent types choose the actions specified in the contract; ($PC1$) and ($PC2$) are participation constraints; the constraints ($IC12$) and ($IC21$) ensure that the agent types self-select into the right contracts.

As an example, we consider our marketer application from Subsection 3.2 with risk-averse agent and unlimited liability. Recall that the misspecification in the biased agent's model relaxes incentive compatibility so that the principal strictly benefits from her bias. Denote by $w_1^*(y)$ the wage scheme that maximizes the principal's profit when the agent is biased, and by $w_2^*(y)$ the wage scheme that maximizes the principal's profit when the agent is rational.[14] Observe that the rational agent earns an information rent under the incentive scheme $w_1^*(y)$. She earns more than $\bar{U}$ by not exerting effort, while the biased agent exerts effort and earns her reservation utility $\bar{U}$.

We get the following result. There exists a critical threshold $\lambda^* \in [0, 1)$ so that, if the chance of an biased agent exceeds λ^*, the principal strictly benefits from the biased agent's misspecification. The optimal screening contract then does not educate the agent, implements

[14]For the example, we assume that in both cases the principal wishes to implement high effort with certainty, and that $IC1$ is relaxed when $p_1(a = 1)$ increases.

high effort from the biased type, and low effort from the rational type. It optimally trades-off the principal's benefit from the biased type's misspecification and the information rent that must be paid to the rational type.[15] As λ decreases, taking advantage of the biased type's misperception becomes less and less attractive. Thus, if $\lambda < \lambda^*$, the optimal screening contract implements high effort from both agent types through incentive scheme $w_2^*(y)$, so that the principal does not benefit from the biased type's misspecification. These observations directly follow from the following general result.

Proposition 6. *(Screening) Consider the screening version of our framework. Suppose that the biased type's model $\hat{\mathcal{R}}$ is perfect and that the principal benefits from the misspecification if $\lambda = 1$. Then there exists $\lambda^* \in [0, 1)$ so that under any optimal screening contract the principal does not educate the agent and benefits from the misspecification if and only if $\lambda > \lambda^*$.*

This result is a generalization of Proposition 2 from von Thadden and Zhao (2012) to our framework. For an important class of screening problems, it shows a simple and intuitive comparative static result: Under an optimal screening contract, the principal benefits from a biased agent's misperception if the chance of having a biased agent is above a certain threshold; otherwise, he cannot take advantage of the misperception and the optimal contract is the same as in the canonical model. In von Thadden and Zhao (2012), the agent is unaware of her action set A and chooses a default action if the principal keeps her unaware. Our generalization therefore shows that the comparative static is relevant for settings in which the agent knows all actions and potential outcomes, and has correct expectations on the equilibrium path. It also holds for different kinds of misspecifications (e.g., those discussed in Subsection 3.2).

Proof. The proof is close to that of Proposition 2 in von Thadden and Zhao (2012). We write the objective function of the screening problem in (36) as

$$h(w_1, p_1(a), w_2, a_2; \lambda) = \lambda \sum_{a \in A} \sum_{y \in Y} p_1(a) p(y \mid a)(y - w_1(y)) + (1-\lambda) \sum_{y \in Y} p(y \mid a_2)(y - w_2(y)). \quad (37)$$

Note that h is absolutely continuous and differentiable everywhere in λ for each set of wage schedules and actions $(w_1, p_1(a), w_2, a_2)$. Denote by $V^t(w_t(y))$ the principal's expected profit from type t if he keeps type 1's model misspecified and type t chooses her effort under the wage schedule $w_t(y)$. Since $w_1^*(y)$ satisfies the constraints $(IC1)$ and $(PC1)$ we have

$$\sum_{a \in A} \sum_{y \in Y} p_1(a) p(y \mid a)(y - w_1(y)) \leq V^1(w_1^*(y)), \quad (38)$$

[15] Note that the principal cannot benefit from the biased agent's misspecification (which always would create an information rent) and implement high effort from both agent types.

and since $w_2^*(y)$ satisfies the constraints in $(IC2)$ and $(PC2)$ we have

$$\sum_{y\in Y} p(y \mid a_2)(y - w_2(y)) \leq V^2(w_2^*(y)). \tag{39}$$

Therefore, we have

$$\begin{aligned} \mid h_\lambda(w_1, p_1(a), w_2, a_2; \lambda)) \mid &= \left| \sum_{a\in A}\sum_{y\in Y} p_1(a)p(y \mid a)(y - w_1(y)) - \sum_{y\in Y} p(y \mid a_2)(y - w_2(y)) \right| \\ &\leq V^1(w_1^*(y)) + V^2(w_2^*(y)) \end{aligned} \tag{40}$$

for all $(w_1, p_1(a), w_2, a_2)$ and λ. Denote by $V(\lambda)$ the principal's value function, i.e., his maximal profit for given share λ. Theorem 2 of Milgrom and Segal (2002) implies that $V(\lambda)$ is absolutely continuous and that

$$V'(\lambda) = \sum_{a\in A}\sum_{y\in Y} p_1(a)p(y \mid a)(y - w_1(y)) - \sum_{y\in Y} p(y \mid a_2)(y - w_2(y)) \tag{41}$$

whenever this derivative exists. Suppose there is an λ so that $V'(\lambda) \leq 0$. Equation (41) then implies that the principal then earns weakly more from type 2 than from type 1 and for the optimal $(w_1, p_1(a), w_2, a_2)$ we have

$$h(w_1, p_1(a), w_2, a_2; \lambda) \leq \sum_{y\in Y} p(y \mid a_2)(y - w_2(y)) \leq V^2(w_2^*(y)). \tag{42}$$

In this case, it is optimal for the principal to educate type 1 and to offer $w_2^*(y)$. Observe that if it is optimal for the principal to educate type 1 at some $\hat{\lambda}$, it is optimal for him to educate type 1 for all $\lambda < \hat{\lambda}$, which proves the existence of a threshold $\lambda^* \in [0, 1]$. By assumption, we have $V^1(w_1^*(y)) > V^2(w_2^*(y))$. Hence, the continuity of $V(\lambda)$ implies $\lambda^* < 1$. □

A.10 Neglect of Correlation

In all applications, we assumed that the agent's subjective model $\mathcal{R}$ is perfect so that in equilibrium she correctly predicts her expected payoff. We now present an example where $\mathcal{R}$ is not perfect an where the agent exhibits neglect of correlation. Consider the objective and subjective model from Figure A1. They are identical except that the agent does not take into account a potential correlation between variable 1 and variable 2. We show that both the incentive compatibility and the participation constraint can be relaxed through this misspecification.

Recall the setting from Subsection 3.2 with binary action $a \in \{0, 1\}$; binary output and costs, $y \in \{y_L, y_H\}$ and $c \in \{c_L, c_H\}$; and binary variables $x_1, x_2 \in \{0, 1\}$. We assume that

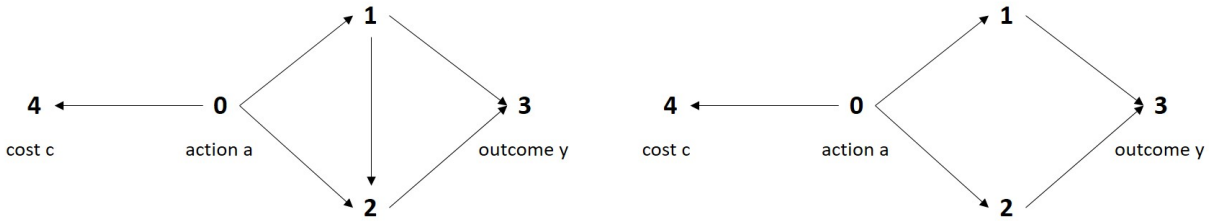

Figure A1: Objective model $\mathcal{R}^*$ (left) and subjective model $\mathcal{R}$ (right) in the neglect of correlation example.

$p(x_1 = 1 \mid a) = \beta_a^1$ and that there is a positive correlation between variable 1 and variable 2, $p(x_1 = 1 \mid a, x_1) = \beta_a^2 + \beta_{x_1}^{1,2}$. For convenience, we set $\beta_0^1 = \beta_0^2 = \beta_0^{1,2} = 0$. Finally, we assume that output is high if either x_1 or x_2 (or both) take on value one, $p(y_H \mid x_1, x_2) = \max\{x_1, x_2\}$.

We examine the impact of the misspecification on the optimal equilibrium contract. Suppose the principal wishes to implement $p(a = 1) = \alpha$. In equilibrium, the agent correctly asses the probability of a high outcome at node 1 and node 2 after high and low effort. So we have

$$p_{\mathcal{R}}(x_1 = 1 \mid a = 1; \alpha) = \beta_1^1, \tag{43}$$

$$p_{\mathcal{R}}(x_2 = 1 \mid a = 1; \alpha) = \beta_1^2 + \beta_1^1\beta_1^{1,2}. \tag{44}$$

The agent also understands the true relationship between the output y and outcomes x_1, x_2, that is, $_{\mathcal{R}}(y_H \mid x_1, x_2; \alpha) = \max\{x_1, x_2\}$. However, since the agent does not take into account the correlation between the nodes 1 and 2, she double count those instances in which both components take on the value one. So while the true probability after a high output after high effort equals

$$p(y_H \mid a = 1) = \beta_1^1 + (1 - \beta_1^1)\beta_1^2, \tag{45}$$

it is

$$p(y_H \mid a = 1; \alpha) = \beta_1^1 + (1 - \beta_1^1)\beta_1^2 + (1 - \beta_1^1)\beta_1^1\beta_1^{1,2} \tag{46}$$

in the agent's mind. The term $(1 - \beta_1^1)\beta_1^1\beta_1^{1,2}$ captures the double counting. The agent correctly predicts that output is low with certainty if she chooses low effort. Thus, the agent overestimates the effectiveness of her effort and her expected payoff, provided that $w(y_H) > w(y_L)$ and she exerts high effort with positive probability. Therefore, the misspecification in the agent's subjective model relaxes both the incentive compatibility and the participation constraint.

A.11 Authority and Supervision

A crucial design aspect of an organization is the allocation of decision rights. In this subsection, we demonstrate that the agent's subjective view on the organization may vary in how

decision rights are distributed. To this end, we consider a version of Aghion and Tirole's (1997) model of formal and real authority. In this model, both the principal and the agent gather information on profitable projects, and can make project suggestions. Which suggestion gets implemented depends on which party has formal authority: If the principal has formal authority, he decides which of the suggested projects is realized; if the agent has formal authority, the agent decides which project gets implemented. We show that the allocation of decision rights changes what the principal communicates about information gathering to the agent.

We examine a version of Aghion and Tirole (1997) in which the principal "supervises" the agent (all mathematical details are relegated to the end os this section). There are three potential projects and three states $1, 2, 3$ of the world that determine project availability. A project is denoted by (y, z), where $y \in \{0, y_L, y_H\}$ is the principal's payoff from this project, $z \in \{0, z_L, z_H\}$ is the agent's payoff, and $y_H > y_L > 0$ as well as $z_H > z_L > 0$. The "zero-project" is available in all states and offers a payoff of zero for both parties. The "principal-project" offers a high payoff y_H to the principal and a low payoff z_L to the agent; it is available in states 1 and 3. The "agent-project" offers y_L to the principal and z_H to the agent; it is available in states 2 and 3. Each party can suggest the principal- or agent-project only if it identifies the state of the world. The principal (agent) chooses effort $e \in [0, 1]$ (action $a \in [0, 1]$) and then identifies the state with probability e (a) and incurs convex expected costs of $g(e)$ ($g(a) = \mathbb{E}[c \mid a]$). If a party identifies the state, it suggests the project that yields the highest payoff for this party. The implementation decision then depends on the allocation of decision rights.

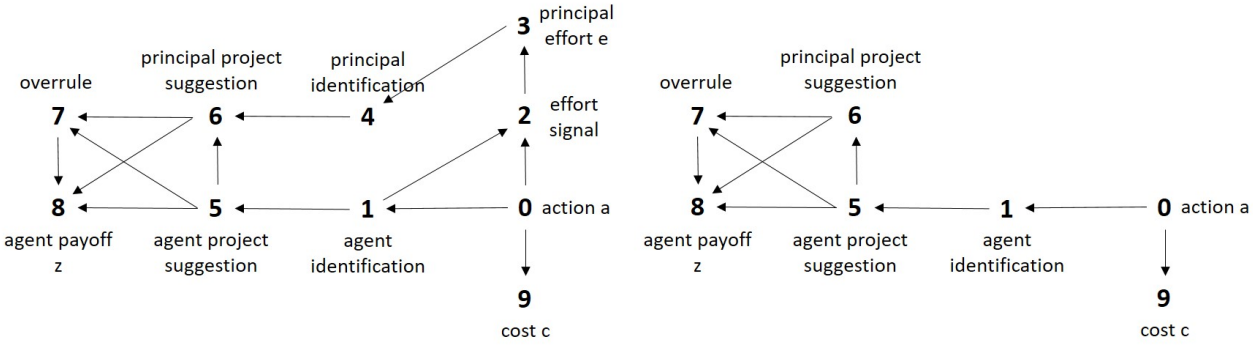

Figure A2: Objective model $\mathcal{R}^*$ (left) and subjective model $\hat{\mathcal{R}}$ (right) in the authority and supervision application.

Before the principal decides about his effort, he observes a signal $x_2 \in \{0, 1\}$ that depends on the agent's action a and whether the agent identified the state or not. If the agent identifies the state, the probability of a good signal $x_2 = 1$ is a; otherwise, it is ξa with $\xi \in (0, 1)$. So if the principal observes a bad signal $x_2 = 0$, this is an indication that the agent may not have identified the state. It is then optimal for him to exert more effort than after a good signal.

We adjust the definition of an equilibrium contract to the present setting. An extended

contract $\{p(e \mid x_2), (\mathcal{R}, w(z) = z, p(a))\}$ is an extended equilibrium contract if (i) effort $p(e \mid x_2)$ after signal x_2 is optimal for the principal given the agent's action $p(a)$, and (ii) $p(a)$ is a personal equilibrium at $\mathcal{R}$ and $w(z) = z$, given that the principal exerts effort $p(e \mid x_2)$.

Principal Formal Authority. We first consider the case when the principal keeps formal authority and implements the project suggestion that yields him the highest payoff. He then overrules the agent's suggestion in two cases: when both parties identify the state and they suggest different projects (i.e., in state 3), and when the principal identifies the state while the agent does not. The objective model $\mathcal{R}^*$ in Figure A2 captures the corresponding causal model.

Suppose the agent's subjective model is $\mathcal{R}^*$ so that she is rational. She then takes the principal's supervision into account. Exerting more effort a reduces the expected principal's effort e, which has two countervailing effects on the agent's payoff. First, it reduces the chance that the principal overrules the agent when she identifies the state by herself. Second, it reduces the chance that the principal comes up with a profitable project if the agent fails to identify the state by herself. The agent is aware of these two effects and chooses her action accordingly. Suppose that $\{e_1^f, e_0^f, (\mathcal{R}^*, w(z) = z, a^f)\}$ is the extended equilibrium contract that is optimal for the principal when the agent is rational, where e_1^f (e_0^f) is the effort that the principal exerts after a good (bad) signal, with $0 < e_1^f < e_0^f$, and $a^f > 0$.

Next, suppose that the agent does not take into account the principal's supervision so that her subjective model is given by model $\hat{\mathcal{R}}$ on the right of Figure A2. The principal may be able to effectively hide from the agent that he gets early signals on her performance. The design of the workplace may reinforce this belief, with the agent's workplace far away from the principal's office and few opportunities of direct interaction. The question is whether the principal benefits from this misperception. We compare the agent's expected marginal return to effort under the objective and subjective model at the original extended equilibrium contract with a^f, e_1^f, e_0^f,

$$\left.\frac{\partial \mathbb{E}[z \mid a]}{\partial a}\right|_{a^f, e_1^f, e_0^f} - \left.\frac{\partial \mathbb{E}_{\hat{\mathcal{R}}}[z \mid a; a^f]}{\partial a}\right|_{a^f, e_1^f, e_0^f} = (e_0^f - e_1^f)\left[a^f \frac{1}{3}(z_H - z_L) - (1 - a^f)\xi\left(\frac{2}{3}z_L + \frac{1}{3}z_H\right)\right]. \tag{47}$$

Recall that $e_0^f > e_1^f$. The terms in the squared brackets on the right-hand side capture the costs and benefits from keeping the agent's model misspecified. The costs are given by the term $a^f \frac{1}{3}(z_H - z_L)$. If the agent's model is misspecified, she ignores that by choosing a higher action she can reduce the probability of being overruled, which reduces her motivation to gather information. The benefits are given by the term $(1 - a^f)\xi(\frac{2}{3}z_L + \frac{1}{3}z_H)$. If the agent's model is misspecified, she does not take into account that by exerting more effort she lowers the chance that the principal comes up with an alternative suggestion, which helps if she does identify the

state by herself. This ignorance increases her motivation to gather information.

Both keeping the agent's model misspecified at $\hat{\mathcal{R}}$ and educating her about the objective model $\mathcal{R}^*$ can be optimal for the principal, depending on the parameters. To illustrate, assume that the agent's cost function $g(a)$ represents talent. A talented agent has low costs of information gathering, so a^f will be relatively large; a mediocre agent has large costs so that a^f will be relatively small. Observe from (47) that if a^f is large enough, effort motivation is higher under the objective model $\mathcal{R}^*$, while if a^f is small enough, the misspecified model $\hat{\mathcal{R}}$ maximizes effort incentives. Hence, it is optimal for the principal to inform a talented agent about his supervision, and to keep the mediocre agent's subjective model misspecified.

Again, we see that different misperceptions can have the same effect on incentives. Note that in model $\hat{\mathcal{R}}$ the nodes 6 and 7 are not on fundamental active paths, so the statistical information encapsulated in these nodes is not relevant for the agent's incentives under $\hat{\mathcal{R}}$. If we take out node 2 or node 3 from the objective model $\mathcal{R}^*$ (and the links to and from these nodes), incentives are the same as under $\hat{\mathcal{R}}$. Intuitively, if only node 2 is missing in the agent's subjective model (relative to $\mathcal{R}^*$), the agent takes into account that the principal exerts effort to come up with alternative suggestions, but erroneously thinks that this effort is independent from her action and information gathering; if only node 3 is missing, the agent admits that the principal gets an early signal about her progress, but ignores the influence of these signals on the principal's behavior. Thus, different misspecifications cause the same incentive effect.

Agent Formal Authority. We next consider the case when the agent holds formal authority so that she implements the project suggestion that yields her the highest payoff. The principal suggestion is preferred to the agent's suggestion only when the principal identifies the state while the agent does not. Again, the objective model $\mathcal{R}^*$ in Figure A2 captures this interaction, we only have to update the probability model.

We analyze how keeping the agent's subjective model misspecified at $\hat{\mathcal{R}}$ affects her incentives. Suppose that $\{e_1^r, e_0^r, (\mathcal{R}^*, w(z) = z, a^r)\}$ is the extended equilibrium contract that is optimal for the principal when the agent is rational, where e_1^r (e_0^r) is the effort that the principal exerts after a good (bad) signal, with $0 < e_1^r < e_0^r$, and $a^r > 0$. We compare the agent's expected marginal return to effort under the objective and subjective model at a^r, e_1^r, e_0^r,

$$\left.\frac{\partial \mathbb{E}[z \mid a]}{\partial a}\right|_{a^r, e_1^r, e_0^r} - \left.\frac{\partial \mathbb{E}_{\hat{\mathcal{R}}}[z \mid a; a^r]}{\partial a}\right|_{a^r, e_1^r, e_0^r} = -(e_0^r - e_1^r)(1 - a^r)\xi\left(\frac{2}{3}z_L + \frac{1}{3}z_H\right). \tag{48}$$

Thus, under agent formal authority, keeping the agent's model misspecified always increases her motivation to gather information. Since profitable project suggestions by the agent never get overruled, the only remaining channel of how the principal's effort affects the agent's utility is the insurance argument. Choosing a higher action reduces the principal's effort and thus the

chance that a profitable project is implemented when the agent does not identify the state. Not taking this into account increases the agent's effort motivation.

Under agent formal authority, the principal may motivate effort by keeping the agent's model misspecified, or discourage effort by educating her, depending on the payoff parameters. If y_L is sufficiently close to y_H, it is not important for the principal whether the principal or the agent project gets implemented. It is then optimal to maximize the agent's effort by not informing her about supervision. However, if y_L is close to zero, it is important for the principal that the principal project gets implemented. Assume that the principal's cost function $g(e)$ represents his time constraints. A busy principal has high costs, so e_1^r, e_0^r will be relatively small; a relaxed principal has low costs of information gathering so that e_1^r, e_0^r will be relatively large. When the principal is busy enough, it is optimal for him to keep the agent's model misspecified (to secure some chance that the principal project gets implemented in state 1). In contrast, if the principal is sufficiently relaxed, it is optimal for him to discourage effort by informing the agent about supervision in order to get his way.

Mathematical Details. We write out the full probability model for the case of principal formal authority. Denote by $\delta \in \{1,2,3\}$ the state. Suppose the agent chooses action a^f, while the principal exerts effort e_1^f after a good signal $x_1 = 1$, and effort e_0^f after a bad signal $x_1 = 0$. We define X_i for the remaining nodes i: $X_1 = X_2 = X_4 = X_7 = \{0,1\}$, $X_5 = X_6 = \{(y,z,\delta) : (y,z)$ is available in state $\delta\}$, $X_8 = \{0, z_L, z_H\}$. We then have $p(a = a^f) = 1$, $p(x_1 = 1 \mid a) = a$, $p(x_2 = 1 \mid x_1 = 1, a) = a$, $p(x_2 = 1 \mid x_1 = 0, a) = \xi a$, $p(e = e_1^f \mid x_2 = 1) = 1$, $p(e = e_0^f \mid x_2 = 0) = 1$, $p(x_4 = 1 \mid e) = e$, $p(x_5 = (0,0,1) \mid x_1 = 0) = 1$, $p(x_5 \mid x_1 = 1) = \frac{1}{3}$ for $x_5 \in \{(y_H, z_L, 1), (y_L, z_H, 2), (y_L, z_H, 3)\}$, $p(x_6 = (0,0,1) \mid x_4 = 0, x_5) = 1$, $p(x_6 = x_5 \mid x_1 = 1, x_5) = 1$ for $x_5 \in \{(y_H, z_L, 1), (y_L, z_H, 2)\}$, $p(x_6 = (y_H, z_L, 3) \mid x_1 = 1, x_5 = (y_L, z_H, 3)) = 1$, $p(x_6 \mid x_1 = 1, x_5 = (0,0,1)) = \frac{1}{3}$ for $x_6 \in \{(y_H, z_L, 1), (y_L, z_H, 2), (y_H, z_L, 3)\}$, $p(x_7 = 1 \mid x_5 = (0,0,1), x_6 \neq (0,0,1)) = 1$, $p(x_7 = 1 \mid x_5 \neq (0,0,1), x_6 = (0,0,1)) = 0$, $p(x_7 = 1 \mid x_5 = x_6 = (y,z,\delta)) = 0$, $p(x_7 = 1 \mid x_5 = (y_L, z_H, 3), x_6 = (y_H, z_L, 3)) = 1$, $p(z \mid x_5 = (y,z,\delta), x_6, x_7 = 0) = 1$ and $p(z \mid x_5, x_6 = (y,z,\delta), x_7 = 1) = 1$. For the case of agent formal authority, only $p(x_7 \mid x_5, x_6)$ has to be adjusted accordingly.

We characterize the principal's equilibrium effort. Suppose that the principal has formal authority and the agent chooses action a^f. The principal's expected payoff from effort e after signal $x_2 \in \{0,1\}$ is then

$$\begin{aligned}
\mathbb{E}[V \mid x_2 = 1, e] &= e\left(\frac{2}{3}y_H + \frac{1}{3}y_L\right) + (1-e)\frac{a^f}{a^f + (1-a^f)\xi}\left(\frac{1}{3}y_H + \frac{2}{3}y_L\right) - g(e), \\
\mathbb{E}[V \mid x_2 = 0, e] &= e\left(\frac{2}{3}y_H + \frac{1}{3}y_L\right) + (1-e)\frac{a^f}{a^f + (1-\xi a^f)}\left(\frac{1}{3}y_H + \frac{2}{3}y_L\right) - g(e).
\end{aligned}$$

Under agent formal authority, when the agent exerts effort a^r, these values are

$$\begin{aligned}
\mathbb{E}[V \mid x_2 = 1, e] &= e\left(1 - \frac{a^r}{a^r + (1-a^r)\xi}\right)\left(\frac{2}{3}y_H + \frac{1}{3}y_L\right) + \frac{a^r}{a^r + (1-a^r)\xi}\left(\frac{1}{3}y_H + \frac{2}{3}y_L\right) - g(e), \\
\mathbb{E}[V \mid x_2 = 0, e] &= e\left(1 - \frac{a^r}{a^r + (1-\xi a^r)}\right)\left(\frac{2}{3}y_H + \frac{1}{3}y_L\right) + \frac{a^r}{a^r + (1-\xi a^r)}\left(\frac{1}{3}y_H + \frac{2}{3}y_L\right) - g(e).
\end{aligned}$$

From these equations, we derive $e_0^f > e_1^f$ and $e_0^r > e_1^r$.

Next, we characterize the agent's equilibrium effort under the subjective and objective model. Note from the subjective model $\hat{\mathcal{R}}$ that nodes 6 and 7 are not on fundamental active paths, so we can ignore them in the calculations below. Suppose that the principal has formal authority and exerts effort e_1^f (e_0^f) after a good (bad) signal. The agent's expected payoff from action a when the equilibrium action is a^f is then

$$\begin{aligned}
\mathbb{E}_{\hat{\mathcal{R}}}[z \mid a; a^f] &= a\left[\frac{1}{3}z_L + \frac{2}{3}z_H - \frac{1}{3}(a^f e_1^f + (1-a^f)e_0^f)(z_H - z_L)\right] \\
&\quad + (1-a)\left[(\xi a^f e_1^f + (1-\xi a^f)e_0^f)\left(\frac{1}{3}z_H + \frac{2}{3}z_L\right)\right] - g(a).
\end{aligned}$$

For the objective model, we get $\mathbb{E}[z \mid a]$ from $\mathbb{E}_{\hat{\mathcal{R}}}[z \mid a; a^f]$ by replacing a^f by a (i.e., the agent takes into account how her action impacts on the principal's effort). From this we get equation (47). Under agent formal authority, when her equilibrium action is a^r and the principal exerts effort e_1^r (e_0^r) after a good (bad) signal, the agent's expected payoff from action a is

$$\mathbb{E}_{\hat{\mathcal{R}}}[z \mid a; a^r] = a\left(\frac{2}{3}z_H + \frac{1}{3}z_L\right) + (1-a)(\xi a^r e_1^r + (1-\xi a^r)e_0^r)\left(\frac{1}{3}z_H + \frac{2}{3}z_L\right) - g(a).$$

From this we get the statements in the main text.