Schumacher, Heiner; Thysen, Heidi

Conference Paper

Equilibrium Contracts and Boundedly Rational Expectations

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Contract Theory, No. G09-V1

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Equilibrium Contracts and Boundedly Rational Expectations*

Heiner Schumacher† Heidi Christina Thysen‡

University of Leuven London School of Economics

February 14, 2017

Abstract

We study an informed-principal framework with moral hazard in which the principal chooses the variables the agent is aware of. The agent fits a causal model connecting these variables to the objective probability distribution. The principal may keep her unaware of some variables so that she incorrectly extrapolates how non-equilibrium actions map into outcomes. This framework captures models of contracting with unaware agents, shrouded attributes, and overconfidence in a unified manner. We provide a general characterization of the variable set the agent has to be aware of so that in equilibrium she anticipates the true relationship between actions and outcomes.

Keywords: Informed Principal, Moral Hazard, Unawareness, Bayesian Networks
JEL Classification: D03, D82, D86

*Preliminary and incomplete version. We are grateful to Yair Antler, Kfir Eliaz, and Ran Spiegler for their valuable comments and suggestions. The usual disclaimer applies.

†Corresponding Author. University of 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.

"The world is my representation." — Arthur Schopenhauer

1 Introduction

At the beginning of a contractual relationship, an agent may not know everything about her new business. As consumer she may be unaware of certain details in the small print or hidden add-on fees; as employee she may be unaware of the contingencies she runs into if she takes certain actions or the effectiveness of her effort. She may learn all these details when she gathers practical experience with the new business. However, the principal may also be able to influence the learning process, e.g., by restricting access to vital information or discouraging experimentation. As a result the agent may not understand correlations that otherwise would change her behavior. In this paper, we analyze to what extent the principal can profit from the agent's lack of knowledge on a permanent basis – that is, by structuring the learning process in a way such that even an infinite number of observations does not lead to rational expectations.

To capture misperceptions that affect the agent's equilibrium behavior, we use Spiegler's (2016a) Bayesian network model and combine it with an informed-principal framework. This modeling strategy allows us to have distorted expectations without imposing parameter assumptions on the agent's beliefs. The key difference to previous models of contracting with boundedly rational agents is that the agent's observations match her expectations in equilibrium. The agent is never "surprised" by the outcome in the sense that she misjudges its probability of occurrence. However, she may incorrectly use her data to extrapolate how non-equilibrium actions affect the outcome distribution.

In our model, the principal's project can be described by a network of stochastic relationships between different variables. These variables include at least the agent's action a, the output y, and her costs of action c;[1] they may comprise other variables that impact on outcomes through different channels (such as "add-on usage", "sales", or "customer relationships"). A link $j \to k$ means that variable j impacts on variable k. The network can be represented as a directed, acyclic graph (DAG) $\mathcal{R}^*$, see the left graph in Figure 1 for an example. The true probability distribution p over these variables factorizes according to the DAG. For example, the left graph in Figure 1 implies

$$p(x^*) = p(x_0)p(x_1 \mid x_0)p(x_2 \mid x_0, x_1)p(x_3 \mid x_0, x_1, x_2)p(x_4 \mid x_1, x_3)p(x_5 \mid x_0), \qquad (1)$$

where $x^* = (x_0, ..., x_5)$; x_0 is the action a, x_4 is the output y, and x_5 is the cost c. Initially, the agent may not be aware of all nodes.[2] We denote her subjective DAG by $\mathcal{R}$

[1] Our model collapses to the canonical principal-agent framework if action, output, and costs are the only project variables.

[2] In this paper, we use the terms "variables", "nodes", and "components" synonymously.

which is $\mathcal{R}^*$ restricted on the variables she is aware of. The agent then fits her subjective model to p so that her belief $p_{\mathcal{R}}$ factorizes according to $\mathcal{R}$. For instance, if the agent's subjective DAG is given by the right graph in Figure 1, she is unaware of the second and third node. Her subjective belief over the variables $x = (x_0, x_1, x_4, x_5)$ is then given by

$$p_{\mathcal{R}}(x) = p(x_0)p(x_1 \mid x_0)p(x_4 \mid x_1)p(x_5 \mid x_0). \tag{2}$$

The principal knows the objective DAG $\mathcal{R}^*$ and can make the agent aware of any node she is unaware off. In the example, he can make her aware of the second and/or the third node. If the principal makes her aware of both nodes, the agent's subjective DAG corresponds to the objective DAG $\mathcal{R}^*$ so that she holds rational beliefs. A contract in our model specifies an incentive scheme $w(y)$ and the agent's subjective DAG $\mathcal{R}$. The agent's action maximizes her expected utility for given subjective belief $p_{\mathcal{R}}$. This belief may depend on the agent's action. We therefore follow Spiegler (2016a) and formalize the agent's equilibrium action as personal equilibrium.

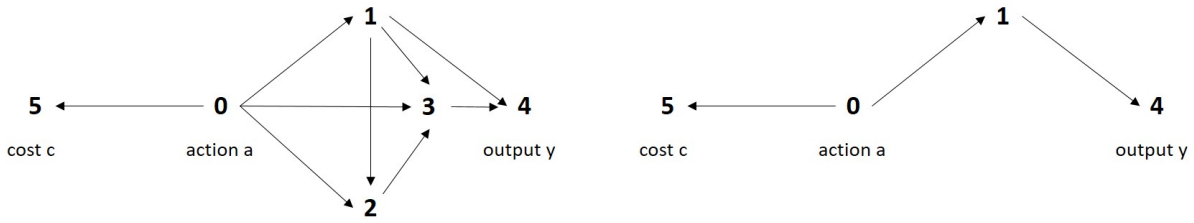

Figure 1: An objective DAG $\mathcal{R}^*$ (left) and the agent's subjective DAG $\mathcal{R}$ (right).

In Section 4, we use our framework to study a number of models from the literature on contracting with boundedly rational agents. Specifically, we examine moral hazard with overconfident agents (Gervais and Goldstein 2007, De la Rosa 2011, Gervais et al. 2011, Spinnewijn 2015), contracting with unaware agents (Von Thadden and Zhao 2012, Auster 2013), and competition when firms can shroud add-on prices (Gabaix and Laibson 2006, Armstrong and Vickers 2012). These models adopt very different modeling strategies and make different assumptions on the agent's knowledge and reasoning. We show that they can be represented in a unified framework in which the agent knows all actions and outcomes.

Many results from the original models can be recovered in our equilibrium framework. If the agent is unaware of some components of the production function, an "incentive effect" may arise – the agent overestimates the importance of her effort so that the principal can reduce incentives or implement higher effort. Depending on the probability distribution p, the agent may also underestimate the importance of her actions in which case the principal has strict a preference for making her aware of the true production function. If the agent is unaware of some components of the cost function, then in the absence of

incentives she may choose an action that does not minimize her costs. Again, depending on the true probability distribution, it may be strictly optimal for the principal to (not) make the agent aware of the true cost function.

We obtain all results despite the agent knowing the true distribution over outcomes *given her equilibrium action*. To illustrate this difference to the original models, consider the shrouded attributes model of Gabaix and Laibson (2006). In the original version, the myopic agent is surprised by the increase in her expenses (since she did not anticipate the use of the add-on). In our version, the agent knows the exact equilibrium distribution over the prices she is going to pay. Nevertheless, she does not understand how substitution effort affects the price distribution and thus continues to use the add-on. The absence of surprises in our equilibrium greatly enhances the appeal of the original models. It shows that their results do not rely on the one-shot nature of the principal-agent interaction and therefore are (to some extent) robust to learning.

We then turn to a general analysis and ask which components of the principal's project the agent has to be aware off so that she acts rationally, regardless of the probability distribution p and the incentive scheme $w(y)$. In the examples we discuss in Section 4, we choose the objective DAG $\mathcal{R}^*$ so that the omission of one component leads to non-rational equilibrium behavior. However, the question is how general these findings are. Does the agent have to be aware of all project components to understand the true relationship between action and outcomes?

In Sections 5 and 6, we study this question by extending the concept of "behavioral rationality" from Spiegler (2016a) to our framework and applying results from the Bayesian network literature. The agent is behaviorally rational if for all distributions p and incentive schemes $w(y)$ she correctly anticipates how outcomes vary with actions. For a large subclass of DAGs, so-called "perfect DAGs", we provide a complete characterization of the set of nodes H^* that the agent has to be aware off in order to be behaviorally rational. All DAGs that we use in Section 4 belong to this class; moreover, Spiegler (forthcoming) shows that perfect DAGs are the outcome of a learning process in which the agent sequentially uses large datasets to compute the true model. We show that the set H^* can be a strict subset of all project components. This implies that the agent may understand the true relationship between actions and outcomes even if she is not aware of all variables and thus has a misspecified model of the environment. However, if she is unaware of nodes in H^*, then for an open set of probability distributions p and incentive schemes $w(y)$ she acts behaviorally irrational in equilibrium. The principal then exploits unawareness or has a strict incentive to make the agent aware of certain components. In the example above, the agent is behaviorally rational as long as she is aware of the set $H^* = \{0, 1, 3, 4, 5\}$. Being unaware of component 2 does not change the optimal incentive scheme for the principal; while being unaware of component 3 does (for some

distributions p).

The remainder of the paper is organized as follows. In Section 2, we relate our contribution to the previous literature. Section 3 describes the model. In Section 4, we examine several contracting models from the literature in our framework. In Section 5, we explain the main results and in Section 6 we provide their proofs. Section 7 concludes. Lengthy proofs are relegated to the Appendix.

2 Related Literature

Bayesian Networks. Bayesian networks and directed acyclic graphs 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] Spiegler (2016a) introduced Bayesian networks to model agents with boundedly rational expectations. He showed that DAGs can be used to capture a variety of different inference errors such as reverse causation, coarseness and mis-attribution biases. Moreover, he characterizes when an agent who omits a single link from the objective DAG is behaviorally rational when all nodes in the objective DAG share a link. Spiegler (forthcoming) proposes a natural extrapolation procedure to deal with missing data. The users of the resulting dataset are unaware of the nature of the extrapolation. He shows that a DAG can be justified as the output of the extrapolation procedure over an ordered dataset if and only if the DAG is perfect. Spiegler (2016b) analyzes when an agent with a subjective DAG can systematically be fooled by providing biased forecasts. He demonstrates that a subjective DAG induces unbiased forecasts if and only if the DAG is perfect. Finally, Eyster and Rabin (2014) use DAGs to model the observational structure in a social learning setting and use this to analyze the extent to which imitation is rational.

Contracting with Boundedly Rational Agents. A growing literature studies contracting between rational principals and boundedly rational agents; see Kőszegi (2014) for an extensive survey. Our paper provides a common framework for two strands of this literature. The first strand analyzes versions of the moral-hazard model with overconfident agents; see, for example, Gervais and Goldstein (2007), De la Rosa (2011),

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

Gervais et al. (2011), Spinnewijn (2015). The second strand examines the extent to which firms can exploit consumers who do not fully understand all aspects of the contract (but could in principle be educated through the contract); see Gabaix and Laibson (2006), Armstrong and Vickers (2012), and Kosfeld and Schüwer (forthcoming). We show that several results from these two literatures can be obtained even if the agent correctly anticipates the consequences of her equilibrium action. Thus, they are to some extent robust to experience and learning.

Unawareness. We borrow the term "unawareness" from a rich literature that builds state-space models of unawareness and applies them to contract theoretical settings. In this literature, unawareness refers to the fact that an agent may not know all contingencies that could potentially happen. Dekel et al. (1998) showed that it is impossible to incorporate non-tivial unawarness in a standard state space model. Heifetz et al. (2006, 2013) and Galanis (2013) therefore develop state-space models that capture non-tivial unawareness. Several papers use the state-space model from Heifetz et al. (2006) to examine the impact of unawareness on the principal-agent relationship. Filiz-Ozbay (2012) study whether insurance companies exploit or educate the consumer if she does not anticipate all events that may cause a damage. Von Thadden and Xiao (2012) and Auster (2013) study moral hazard problems with a fully aware principal and an unaware agent. The former model assumes that the agent is unaware of some actions a, while the latter assumes that the agent is unaware of some potential outcomes y. In our setting, the agent is aware of all actions and potential outcomes. She may be unaware of variables that have no direct impact on her utility, but that stochastically influence final outcomes. In this case, she forms beliefs according to a misspecified model so that she holds biased beliefs about how non-equilibrium actions affect the distribution over project outputs or effort costs. Despite these differences, we recover some of the results from von Thadden and Xiao (2012) and Auster (2013) in our framework.

Informed Principals. Starting with Myerson (1983) and Maskin and Tirole (1990, 1992) a large literature studies principal-agent models in which the principal has private information. Informed-principal models with moral hazard on the side of the agent are considered, for example, in Inderst (2001), Bénabou and Tirole (2003), Martimort and Sand-Zantman (2006), Kaya (2010), Wagner et al. (2015), and Karle et al. (2016). In these model, the agent has rational expectations. The shape of the offered contract thus not only influences the agent's incentives, but may also signal the principal's private information to her. In contrast, signaling does not occur in our framework. Instead, the contract defines the set of variables the agent uses to fit her subjective model.

3 The model

Basic Framework. We consider a standard principal-agent problem and add a causal structure to the principal's project and the agent's costs. The principal proposes a contract to the agent. The agent then chooses and action $a \in A$, where $A \subset \mathbb{R}$. One action out of A is the agent's rejection of the contract in which case she enjoys the value of her outside option $\bar{U}$, while the principal earns zero. If the agent does not reject the contract, her action stochastically influences the project's output and her effort costs. Let Y be the finite set of possible outputs and $p(y \mid a)$ the probability of output y when the agent chooses action a; let C be the finite set of possible costs and $p(c \mid a)$ the probability of cost c when the agent chooses a. Her utility from wage w is given by the function $u(w)$. The principal's contract specifies – among other things that we explain below – the wage conditional on output, $w(y)$. When the outcome is y and the agent's costs are c, the principal's utility is $V = y - w(y)$ and the agent's utility is $U = u(w(y)) - c$.

Causal Structure. We explicitly model how the agent's effort affects the output and her costs. Her effort influences one or more project components.[4] The interaction of these components in turn determine the distribution over outcomes and costs. Following Spiegler (2016a) we model the causal structure of the principal's project as a directed acyclic graph (DAG). The outcome of each node $i \in N^* = \{0, ..., n, n+1\}$ of the project is captured by a variable $x_i \in X_i$, where X_i is a finite set that contains at least two elements. Node 0 is the action ($x_0 = a$, $X_0 = A$), node n is the outcome ($x_n = y$, $X_n = Y$), and node $n+1$ is the agent's cost ($x_{n+1} = c$, $X_{n+1} = C$).[5] The state is a vector $x^* = (x_0, x_1, ..., x_{n+1})$ and the set of all states is $X^* = \times_{i \in N^*} X_i$. The causal structure is given by an irreflexive, asymmetric and acyclic binary relation R^* over N^*. We denote it by $\mathcal{R}^* = (N^*, R^*)$. One may read iR^*j for $i, j \in N^*$ as "node i impacts on node j." The set of nodes that influence i is defined as $R^*(i) = \{j \in N^* \mid jR^*i\}$. The probability distribution over states naturally factorizes according to R^* via the formula

$$p(x^*) = \prod_{i \in N^*} p(x_i \mid x_{R^*(i)}). \tag{3}$$

Beliefs and Decisions. The principal is aware of the true causal structure of his project. We will therefore call $\mathcal{R}^*$ the "objective DAG." In contrast, the agent may miss out important nodes and the links to and from these nodes. She has her own subjective causal model, described by the DAG $\mathcal{R} = (N, R)$. N is a subset of N^*. It contains at least the action a, outcome y, and costs c. The relation R equals R^* restricted on N:

[4] Like sales, production, R&D, corporate culture or the relationships to employees and stakeholders.

[5] In the following, we use these variable labels interchangeably. We also implicitly assume that set of variables that influence the output is disjoint from the set of variables that influence effort costs.

We have iRj for $i, j \in N$ if and only if iR^*j. Denote by $x = (x_i)_{i \in N}$ the corresponding state vector and $X = \times_{i \in N} X_i$. The agent's DAG maps the objective distribution p to her subjective beliefs, i.e., the agent fits her causal model to the data generated by p. Thus, her beliefs will be given by the factorization formula

$$p_{\mathcal{R}}(x) = \prod_{i \in N} p(x_i \mid x_{R(i)}). \tag{4}$$

The agent chooses the action a to maximize her expected utility given the incentive wage $w(y)$ and her subjective belief $p_{\mathcal{R}}$. As Spiegler (2016a) shows, her behavior potentially influences the evaluation of her actions since $p_{\mathcal{R}}(y, c \mid a)$ may depend on $p(a)$. The agent's action choice therefore must be formalized as a personal equilibrium.

Definition 1 *A mixed action $p(a)$ with full support on A is an ε-perturbed personal equilibrium at DAG $\mathcal{R}$ and wage schedule $w(y)$ if*

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

for every action $a \in A$ with $p(a) > \varepsilon$. An action $p(a)$ is a personal equilibrium at $\mathcal{R}$ and $w(y)$ if there exists a sequence $\varepsilon^k \to 0$ and a sequence $p^k(a) \to p(a)$ of perturbations of $p(a)$, such that, for every k, $p^k(a)$ is an ε^k-perturbed personal equilibrium at $\mathcal{R}$ and $w(y)$.

Contract and Equilibrium. The principal can make the agent aware of one or more nodes through the contract. He essentially chooses the set of components N the agent is aware of (and thus her subjective DAG $\mathcal{R}$). Denote by $\hat{N}$ the agent's default set of nodes, i.e., the nodes she is aware of if the principal does not make her aware of any additional nodes. Let $\hat{\mathcal{R}}$ be the corresponding default DAG. The principal can choose any N that satisfies $\hat{N} \subseteq N \subseteq N^*$. If $\hat{N} = N^*$ (or $\mathcal{R} = \mathcal{R}^*$), principal and agent share the same beliefs and we are back in the standard model. Denote by Γ the set of subjective DAGs $\mathcal{R}$ that the principal could choose for the agent. A contract is then given by the pair $(\mathcal{R}, w(y))$. The agent's PE strategy σ defines for every contract $(\mathcal{R}, w(y)) \in \Gamma \times \mathbb{R}^{|Y|}$ a personal equilibrium $p(a) \in \Delta(A)$ at $\mathcal{R}$ and $w(y)$. An equilibrium of the game consists of the agent's PE strategy σ and a contract $(\mathcal{R}, w(y))$ that maximizes the principal's expected profit for given σ. For convenience, we assume that σ selects for each contract the PE that maximizes the principal's profit.

4 Optimal Contracts for Agents with Boundedly Rational Expectations

Our goal in this section is two-fold. First, we demonstrate that several different models from the literature on contracting with boundedly rational agents can be represented in our framework. Second, we show that some results from these model can be generalized in the sense that they can occur in equilibrium even if the agent correctly anticipates the payoff consequences of her equilibrium action.

4.1 The Incentive Effect – Biased Expectations about the Production Function

We start by analyzing a simple version of our model in which the agent has an incomplete understanding of how effort translates into outcomes. The agent can exert either high or low effort a. Effort affects the outcome y through two nodes, 1 and 2. Initially, the agent is unaware of node 2. The principal can choose whether to make her aware of node 2 or not. Figure 2 shows the objective DAG $\mathcal{R}^*$ and the agent's subjective DAG $\hat{\mathcal{R}}$ if she is kept unaware of node 2. In the following, we analyze the circumstances under which the principal has a strict incentive (not) to keep the agent unaware.

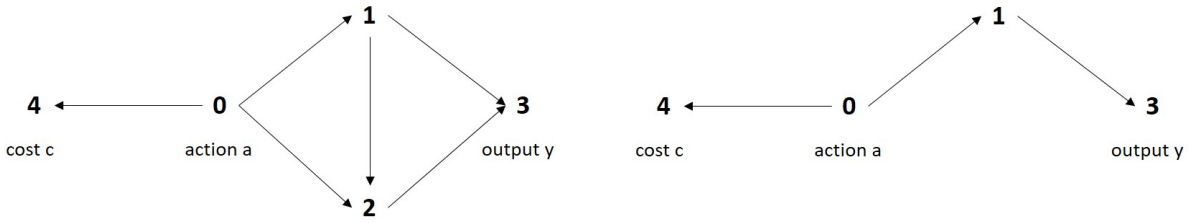

Figure 2: Objective DAG $\mathcal{R}^*$ (left) and subjective DAG $\hat{\mathcal{R}}$ of the unaware agent (right) in the biased expectations about production function example.

Framework and Probability Model. The agent has a strictly concave utility function u. All variables are binary: low and high effort $a \in \{0,1\}$, the variables $x_1, x_2 \in \{0,1\}$ and the outcome $y \in \{0,1\}$. High effort ($a = 1$) creates costs of $c > 0$ for the agent, while low effort ($a = 0$) produces no costs. The agent's outside value is $\bar{U} = 0$. We use the factorization formula (3) to calculate the probability of high output after effort a,

$$p(y = 1 \mid a) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} p(x_1 \mid a) p(x_2 \mid a, x_1) p(y = 1 \mid x_1, x_2). \tag{5}$$

This is also the agent's subjective success probability if she is aware of the objective DAG $\mathcal{R}^*$. Suppose that she is unaware of the second component. Assume that in equilibrium she exerts effort a^* with certainty. Using again the factorization formula, we can calculate

the agent's subjective probability of success from effort a,

$$p_{\hat{\mathcal{R}}}(y = 1 \mid a; a^*) = \sum_{x_1 \in X_1} p(x_1 \mid a) p_{\hat{\mathcal{R}}}(y = 1 \mid x_1), \tag{6}$$

where in equilibrium we have

$$p_{\hat{\mathcal{R}}}(y = 1 \mid x_1) = \sum_{x_2 \in X_2} p(x_2 \mid a^*, x_1) p(y = 1 \mid x_1, x_2). \tag{7}$$

Observe that on the equilibrium path, $a = a^*$, the true and the subjective success probability are identical. However, if the agent is unaware of the second component, she may over- or underestimate the success probability when deviating to another effort level $a \neq a^*$. Assume the agent exerts high effort in equilibrium, $a^* = 1$. Then her subjective success probability from low effort is smaller under unawareness,

$$p_{\hat{\mathcal{R}}}(y = 1 \mid a = 0, a^* = 1) < p(y = 1 \mid a = 0), \tag{8}$$

if for all $x_1 \in X_1$ we have

$$\sum_{x_2 \in X_2} p(x_2 \mid a^* = 1, x_1) p(y = 1 \mid x_1, x_2) < \sum_{x_2 \in X_2} p(x_2 \mid a = 0, x_1) p(y = 1 \mid x_1, x_2). \tag{9}$$

This inequality has an intuitive interpretation. Suppose the second component has a positive (negative) effect on the success probability, i.e., $p(y = 1 \mid x_1, x_2)$ increases (decreases) in x_2; then the unaware agent underestimates the success probability after low effort if effort has a negative (positive) effect on the second component. In other words, the agent overestimates the effectiveness of effort if she is unaware of a channel of causality in which high effort has a negative impact on the final outcome. We say assumption $SA+$ is satisfied if the inequality in (9) holds; and assumption $SA-$ is satisfied if (9) holds with reversed inequality.

Equilibrium Contracts. Suppose the principal wishes to implement high effort. If he chooses to make the agent aware of the second component, $\mathcal{R} = \mathcal{R}^*$, the second-best optimal contract is given by the solution to the maximization problem

$$\max_{w_1, w_0} p(y = 1 \mid a = 1)(1 - w_1) + p(y = 0 \mid a = 1)(-w_0) \tag{10}$$

$$s.t.$$

$$[p(y = 1 \mid a = 1) - p(y = 1 \mid a = 0)][u(w_1) - u(w_0)] \geq c, \tag{IC^*}$$

$$p(y = 1 \mid a = 1) u(w_1) + p(y = 0 \mid a = 1) u(w_0) \geq 0. \tag{PC^*}$$

Standard arguments show that there exists a unique solution to this problem and that both constraints are binding. If the principal chooses to keep the agent unaware of the

second component, $\mathcal{R} = \hat{\mathcal{R}}$, the second-best optimal contract is given by the solution to the maximization problem

$$\max_{w_1,w_0} p(y=1 \mid a=1)(1-w_1) + p(y=0 \mid a=1)(-w_0) \tag{11}$$

$$s.t.$$

$$[p(y=1 \mid a=1) - p_{\hat{\mathcal{R}}}(y=1 \mid a=0; a^*=1)][u(w_1) - u(w_0)] \geq c, \tag{$\hat{IC}$}$$

$$p(y=1 \mid a=1)u(w_1) + p(y=0 \mid a=1)u(w_0) \geq 0. \tag{$\hat{PC}$}$$

Again, this problem has a unique solution which is binding at both constraints. Note that the two maximization problems only differ in the incentive compatibility constraints, IC^* and $\hat{IC}$. Specifically, the incentive compatibility is relaxed through unawareness if the agent with subjective DAG $\hat{\mathcal{R}}$ overestimates the effectiveness of effort, i.e., when assumption $SA+$ is satisfied. In this case, the principal can implement high effort with fewer incentives than when the agent were aware of the second component. He then strictly benefits from the agent's unawareness. In contrast, incentive compatibility is tightened if the unaware agent underestimates the effectiveness of effort, that is, when assumption $SA-$ is satisfied. To implement high effort, the principal then optimally educates the agent about the second component. Since the agent's PE strategy selects the equilibrium action that maximizes the principal's profit, we get the following result.

Proposition 1 (Biased Expectations about Production Function) *Consider the principal-agent problem with biased expectations about the production function. Suppose the optimal contract under awareness, $\mathcal{R} = \mathcal{R}^*$, implements high effort. If the production function satisfies $SA+$, the equilibrium contract is unique, keeps the agent unaware, $\mathcal{R} = \hat{\mathcal{R}}$, and implements high effort; if it satisfies $SA-$, the equilibrium contract is unique, makes the agent aware of all nodes, $\mathcal{R} = \mathcal{R}^*$, and implements high effort.*

If the principal wishes to implement low effort, $a = 0$, it does not matter whether the agent is aware or unaware of the second component. In both cases, the principal offers a fixed-wage that equalizes the participation constraint, which in equilibrium is unaffected by the agent's expectations.

Relationship to previous models. Our framework generates an "incentive-effect" that also has been found in several models with overconfident agents; see De la Rosa (2011), Gervais et al. (2011), and Spinnewijn (2015). In all models, the agent overestimates the effectiveness of her effort so that the principal can implement the same effort with reduced incentives and thereby lower agency costs. The other models use the assumption that the agent overestimates the success probability of her equilibrium action. In our framework, the agent knows the true probability of success when she

chooses her equilibrium effort. However, depending on the probability distribution p, she may underestimate the success probability of non-equilibrium actions.

The model in this subsection also resembles that of Auster (2013). In both models, the agent is unaware of some consequences of her effort; and, depending on the probability distribution over outcomes, the principal may strictly prefer to keep the agent unaware or to make her aware of certain project components. The key difference is that in Auster (2013) the agent is unaware of some outcomes 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. Such a trade-off does not occur in our framework. The agent knows all outcomes in Y and the equilibrium distribution over Y. Consequently, the participation constraint is unaffected by unawareness. As we have shown, the principal can nevertheless benefit from the agent's unawareness. We therefore created a version of Auster's (2013) model in which the agent's experiences match her expectations.

4.2 Incentives for Unaware Agents – Biased Expectations about Effort Costs

We analyze a version of our model in which the agent has biased expectations about the costs of effort. Effort is a continuous variable $a \in [\underline{a}, \overline{a}]$ and the total costs of effort increase in a. Initially, the agent does not fully understand how effort translates into costs. Figure 3 shows the objective DAG $\mathcal{R}^*$ and the agent's subjective DAG $\hat{\mathcal{R}}$ if she is kept unaware. In the case of unawareness, an intermediate "default action" $a^d \in (\underline{a}, \overline{a})$ appears to her as the least costly action (instead of $\underline{a}$). We analyze under what circumstances the principal wants to make the agent aware of the true model.

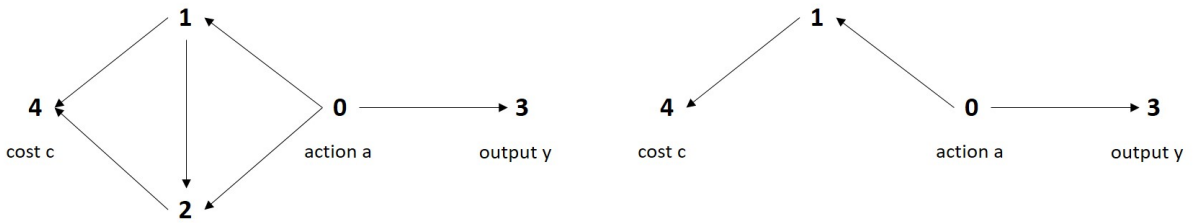

Figure 3: True DAG $\mathcal{R}^*$ (left) and subjective DAG $\hat{\mathcal{R}}$ of the unaware agent (right) in the biased expectations about effort costs example.

Framework and Probability Model. If the agent exerts effort $a \in [\underline{a}, \overline{a}]$, the probability of outcome $y \in \{0, 1\}$ is $p(y \mid a)$. We assume that $p(y = 1 \mid a)$ is strictly concave in a and that the agent knows $p(y \mid a)$. Her total *expected* costs of effort are given by

ka^β. Effort affects costs through two components, 1 and 2.[6] Both variables are binary, $x_1, x_2 \in \{0, 1\}$. The costs of effort are either high, $c = 1$, or low, $c = 0$. The probability of high costs $p(c = 1 \mid x_1, x_2)$ increases in both variables. If the agent is unaware of the second component and chooses a^* in equilibrium, her subjective probability of high costs after choosing action a equals

$$p_{\hat{\mathcal{R}}}(c = 1 \mid a; a^*) = \sum_{x_1 \in X_1} p(x_1 \mid a) \sum_{x_2 \in X_2} p(x_2 \mid a^*) p(c = 1 \mid x_1, x_2). \tag{12}$$

To capture the notion of a default action, we put some more structure on the joint probability distribution p. Assume that for some $a^d \in (\underline{a}, \overline{a})$ we have

$$p(x_1 = 1 \mid a) = \beta_1 + \beta_2 (a - a^d)^2, \tag{13}$$

where β_1, β_2 are small enough so that $p(x_1 = 1 \mid a) \in [0, 1]$ for all actions a. Thus, the probability of $x_1 = 1$ increases in effort if the agent works "too much" ($a > a^d$) and decreases in effort if she works "too little" ($a < a^d$). Let the probability of high costs be given by $p(c = 1 \mid x_1, x_2) = k\Delta(1 + x_1 + x_2)$ for some $\Delta > 0$ with $3k\Delta < 1$. We can calculate[7] that

$$p_{\hat{\mathcal{R}}}(c = 1 \mid a; a^*) = k(a^*)^\beta - k\Delta\beta_2 (a^* - a^d)^2 + k\Delta\beta_2 (a - a^d)^2. \tag{14}$$

From this equation one can observe how the agent's misspecified model affects her expectations about costs. The agent knows the total costs at the equilibrium action a^*, but incorrectly extrapolates costs to other actions.

Equilibrium Contracts. The principal can choose whether to make the agent aware of the second component or not. If he chooses to make her aware of it, $\mathcal{R} = \mathcal{R}^*$, the optimal incentive scheme is given by a solution of the following maximization problem:

$$\max_{a, w_1, w_0} p(y = 1 \mid a)(1 - w_1) + p(y = 0 \mid a)(-w_0) \tag{15}$$

$$s.t.$$

$$a = \left[\frac{\frac{d}{da} p(y = 1 \mid a)(u(w_1) - u(w_0))}{k\beta} \right]^{\frac{1}{\beta - 1}} \tag{IC^*}$$

$$p(y = 1 \mid a) u(w_1) + p(y = 0 \mid a) u(w_0) - k(a)^\beta \geq 0. \tag{PC^*}$$

[6] For example, effort costs may originate from two sources, "annoyance of others" and "stress." Both annoyance of others and stress increase the probability of high physical costs. Initially the agent exclusively focuses on not to annoy her co-workers; that is, she is unaware of the fact that she can influence her stress level through her effort.

[7] See the Appendix for details.

If the principal chooses to keep her unaware of it, $\mathcal{R} = \hat{\mathcal{R}}$, the optimal incentive scheme is given by a solution of the maximization problem

$$\max_{a,w_1,w_0} p(y=1 \mid a)(1-w_1) + p(y=0 \mid a)(-w_0) \tag{16}$$

$$s.t.$$

$$a = \left[\frac{\frac{d}{da} p(y=1 \mid a)(u(w_1) - u(w_0))}{2\Delta k \beta_2} \right] + a^d. \tag{$\hat{IC}$}$$

$$p(y=1 \mid a)u(w_1) + p(y=0 \mid a)u(w_0) - k(a)^\beta \geq 0. \tag{$\hat{PC}$}$$

Note that again the only difference between the two maximization problems is the incentive constraint. Using the two maximization problems, we can characterize the equilibrium contract.

Proposition 2 (Biased Expectations about Costs) *Consider the principal-agent problem with biased expectations about the cost function. There exist values* $0 < \underline{k} < k^d < k_1 \leq k_2 < k_3 < \overline{k}$ *so that*

(i) if $k > \overline{k}$*, the agent does not work for the principal in equilibrium;*

(ii) if $k \in [k_3, \overline{k}]$*, the unique equilibrium contract makes the agent aware of the second component,* $\mathcal{R} = \mathcal{R}^*$*, and offers a fixed wage,* $w_1 = w_0 = k\underline{a}^\beta$*;*

(iii) if $k \in (k_2, k_3)$*, the unique equilibrium contract makes the agent aware of the second component,* $\mathcal{R} = \mathcal{R}^*$*, and offers an incentive wage scheme,* $w_1 > w_0$*;*

(iv) if $k \in [k_1, k_2]$*, an equilibrium contract either makes the agent aware of the second component and offers an incentive wage scheme, or it keeps the agent unaware and pays a fixed wage of* $w_1 = w_0 = k(a^d)^\beta$*;*

(v) if $k \in [k^d, k_1)$*, the unique equilibrium contract keeps the agent unaware,* $\mathcal{R} = \hat{\mathcal{R}}$*, and offers a fixed wage of* $w_1 = w_0 = k(a^d)^\beta$*;*

(vi) if $k \in (\underline{k}, k^d)$*, the unique equilibrium contract keeps the agent unaware,* $\mathcal{R} = \hat{\mathcal{R}}$*, and offers an incentive wage scheme,* $w_1 > w_0$*;*

(vii) if $k \in [0, \underline{k}]$*, any equilibrium contract offers an incentive wage scheme,* $w_1 > w_0$*.*

This result is intuitive. In the regions *ii* and *iii*, the costs of effort are large enough so that the principal does not want the agent to exert her default effort a^d or more. It is better for the principal to implement the minimal effort $\underline{a}$ (in region *ii*), or some incentives to implement relatively a small effort level $a < a^d$ (in region *iii*). To do this, the principal has to make the agent aware of the second component. In the regions *v* and *vi*,

effort costs are small enough to exploit the agent's unawareness. The first-best action is then close or equal to a^d so that the optimal incentive scheme under unawareness is relatively close to the first-best contract under awareness. In contrast, the optimal contract under awareness would entail relatively large agency costs.

Relationship to von Thadden and Zhao (2012). Proposition 2 replicates and extends the first main result from von Thadden and Zhao (2012). The original model assumes that the agent is unaware of her available actions. As long as the principal does not make her aware, she chooses the default action a^d. We can drop this assumption. In our version of the model, the agent is unaware of an important factor that drives her effort costs. Therefore, she incorrectly extrapolates how effort costs change with the chosen action. As a consequence, the principal may choose to provide effort incentives *and* keep the agent unaware of the objective DAG (see region *vi*). This could not happen in the original model where effort incentives were only effective if the agent was aware of her available actions.

4.3 Shrouded Attributes

Next, we consider an extension of our model in which two principals compete for the agent. They sell a base good that comes with an add-on. The agent chooses the principal, $a \in \{0, 1, 2\}$, where $a = 0$ represents "no trade", and substitution effort, $e \in \{0, 1\}$. If the agent exerts substitution effort ($e = 1$), she incurs costs of $c > 0$ and only uses the chosen principal's base good. If the agent does not exert substitution effort ($e = 0$), she uses the base-good and with positive probability also the add-on. Usage of the add-on can result in additional costs, depending on the contract. The agent is either aware or unaware ("myopic") of the add-on component. The objective DAG is equivalent to $\mathcal{R}^*$ in Figure 2, where node 1 is the base good and node 2 the add-on component. The two principals can choose whether to make the myopic agent aware through the contract. In the following, we examine under what circumstances an equilibrium exists in which both principals keep the myopic agent unaware.

Framework and Probability Model. There are two outcomes, a high-cost outcome y_h and a low-cost outcome y_l.[8] The agent's action choice (a, e) affects two components: "base good usage" $x_1 \in \{0, 1\}$ and "add-on usage" $x_2 \in \{0, 1\}$. If the agent chooses some principal's contract and does not exert substitution effort, she always uses the base

[8]For convenience, we assume that principals have no direct costs or benefits from interacting with the agent, i.e., $y_h = y_l = 0$.

good,[9] $x_1 = 1$, and with probability $f < 1$ she also uses the add on, $x_2 = 1$. If the agent chooses some contract and exerts substitution effort, she only uses the base good. Principal k charges π_k for the base good and $\hat{\pi}_k \in [0, \bar{\pi}]$ for the add-on. If the agent uses both goods, she pays the price for the high-cost outcome $w_k(y_h) = -\pi_k - \hat{\pi}_k$; otherwise, she only pays the price for the low-cost outcome $w_k(y_l) = -\pi_k$. The costs of substitution effort are smaller than the highest expected costs of using the add-on, $c < f\bar{\pi}$. The agent's outside value is $\bar{U} < -\bar{\pi}$.

The agent is myopic with probability α. Initially, the myopic type is unaware of the add-on component, $\mathcal{R} = \hat{\mathcal{R}}$. Each principal can make the myopic agent aware of it by offering a contract with $\mathcal{R} = \mathcal{R}^*$. If both principals keep her unaware, she only observes the goods' prices in the low-cost state $w_k(y_l)$ for $k \in \{1, 2\}$. We call this assumption GL. It captures that whenever the agent observes the add-on price, she also understands the add-on component (i.e., the benefits from substitution effort).[10] If at least one principal makes the myopic agent aware of the add-on component, she observes the prices in both states, $w_k(y_l)$ and $w_k(y_h)$ for $k \in \{1, 2\}$. The aware agent knows that the probability of a low-cost state increases in substitution effort,

$$p(y = y_l \mid a \in \{1, 2\}, e = 0) = 1 - f < 1 = p(y = y_l \mid a \in \{1, 2\}, e = 1). \tag{17}$$

If the myopic agent is kept unaware of the add-on component, then in equilibrium she sees no value in exerting substitution effort,

$$p_{\hat{\mathcal{R}}}(y = y_l \mid a \in \{1, 2\}, e = 0) = p_{\hat{\mathcal{R}}}(y = y_l \mid a \in \{1, 2\}, e = 1) = 1 - f. \tag{18}$$

Next, we examine what contracts the two firms will offer in equilibrium.

Equilibrium Contracts. We show that there can be an equilibrium in which both principals keep the myopic agent unaware of the add-on component. Suppose that principals charge the maximal price for the add-on, $\hat{\pi}_k = \bar{\pi}$. The aware agent then exerts substitution effort, while the myopic agent sees no benefit from doing so, see the equality in (18). The competition between the two principals ensures that all the gains from ripping off the myopic agent are transfered to the aware agent through a subsidy in the price of the base good, i.e., $\pi_k = -\alpha f\bar{\pi}$. This renders it unprofitable for principals to educate the myopic agent. If one principal makes the myopic agent aware of the add-on component, she contracts with the other principal, purchases the base-good at the subsidized price, and substitutes away from the add-on. This strategy is profitable for the agent if the share of myopic agents (and thus the subsidy in the base price) is sufficiently large. Educating the agent therefore does not necessarily pay off for a principal.

[9] Here $x_1 = 0$ only occurs if the agent rejects both contracts.

[10] In the appendix, we examine the model if assumption GL is dropped.

Proposition 3 (Shrouded Attributes) *Consider the shrouded attributes model and assume that GL holds. If* $\alpha > \frac{c}{f\bar{\pi}}$*, there exists a symmetric "shrouding" equilibrium in which both principals offer the contract* $(\hat{\mathcal{R}}, w_k(y))$ *with* $w_k(y_l) = \alpha f\bar{\pi}$ *and* $w_k(y_h) = \alpha f\bar{\pi} - \bar{\pi}$*. In this equilibrium, the aware agent chooses substitution effort, while the myopic agent chooses no substitution effort.*

Relationship to the shrouded attributes literature. Proposition 3 replicates Gabaix and Laibson's (2006) main result in our framework. The key difference between the two formulations is that in our model the myopic agent correctly anticipates that with positive probability she will pay $\bar{\pi}$ more than indicated by the base good price. However, she does not understand what creates the variation in the final price and how she could avoid it. In the original model, the myopic agent does not anticipate the high price from base good and add-on usage.

The connection between a shrouded add-on price and the unawareness of the substitution effort benefits is crucial. In Gabaix and Laibson (2006), these two features of the transaction are linked to each other. Whenever the agent observes the add-on price, she also knows about the benefits of substitution effort (assumption GL). The result in Proposition 3 does not hold if GL is dropped. Suppose that principals could unshroud the price without making the agent aware of the add-on component. Then they could offer the myopic agent a better deal without the risk that the agent learns how to substitute away from the add-on. This mechanism undermines the exploitative equilibrium. We can show that if GL does not hold, then in any equilibrium the price for the add-on is below $\frac{c}{f}$ so that no agent exerts substitution effort (see the Appendix for details). This difference to the original model is due to the fact that in equilibrium the agent always knows the distribution over outcomes.

4.4 Positive Effects of Biased Expectations in Teams

In our last example, we again consider biased expectations about the production function. There are now two agents. Agent 1 exerts effort $a \in [0, 1]$; agent 2 observes a and then chooses her effort $e \in [0, 1]$. The success of the project depends on both efforts, and the production technology exhibits a complementarity. The objective DAG is displayed in Figure 4. Node 2 captures the complementarity, node 3 is a "confounding factor" and node 4 is agent 2's effort. Initially, agent 1 is unaware of node 3 and will therefore overestimate the importance of her effort for the final outcome. Agent 2 is aware of all components and therefore unbiased. We analyze to what extent the principal and the agents may benefit from agent 1's unawareness.

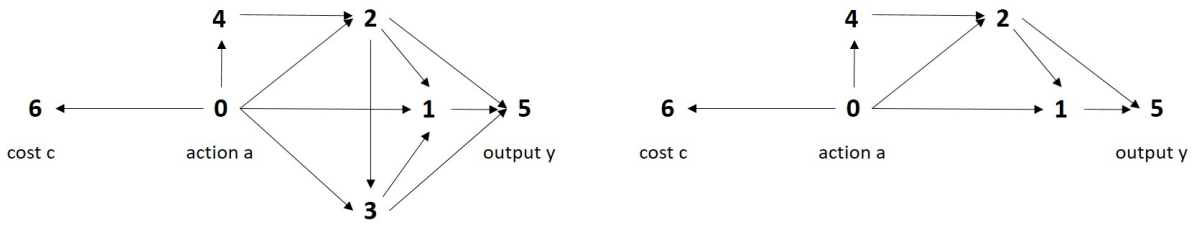

Figure 4: Objective DAG $\mathcal{R}^*$ (left) and subjective DAG $\hat{\mathcal{R}}$ of the unaware agent (right) in the biased expectations in teams example.

Framework and Probability Model. The team project's outcome is binary, $y \in \{0,1\}$. Agents are risk-neutral, protected by limited liability and the value of their outside option is $\bar{U} = 0$. Denote by $w_{i,y}$ agent i's wage when the outcome is y. The agents' efforts affect the outcome through three components $i = 1,2,3$ with binary values $x_i \in \{0,1\}$. If $x_3 = 1$, the high output materializes for sure; if $x_3 = 0$, the high outcome occurs with probability $s_1x_1 + s_2x_2$ where $s_1 + s_2 < 1$; for the influence of the agents' efforts on the components $i = 1,2,3$ we assume $p(x_1 = 1 \mid a, x_2, x_3 = 1) = 1$, $p(x_1 = 1 \mid a, x_2, x_3 = 0) = \alpha_1 a + \alpha_2 x_2$, $p(x_2 = 1 \mid a, e) = \varepsilon ae$, and $p(x_3 = 1 \mid a, x_2) = \beta$. The costs of effort for agent 1 are deterministic and given by $c(a) = \frac{\gamma}{2}a^2$ (same for the second agent). Agent 2 will always optimally respond to agent 1's effort choice. We can derive her reaction function as

$$e^*(a) = \frac{1}{\gamma}(w_{2,1} - w_{2,0})(1-\beta)(s_2 + s_1\alpha_2)\varepsilon a. \tag{19}$$

The principal may or may not make agent 1 aware of the confounding factor. If he keeps her unaware of it, $\mathcal{R} = \hat{\mathcal{R}}$, she may not correctly anticipate the true connection between her effort and the distribution over outputs. Abbreviate $p_1(x_2, a) = p(x_1 = 1 \mid x_2; a)$ and $p_3(x_1, x_2, a) = p(x_3 = 1 \mid x_1, x_2, a)$. If her equilibrium action is $a^* > 0$, then agent 1's subjective probability of a high output after choosing effort a equals[11]

$$\begin{aligned} p_{\hat{\mathcal{R}}}(y = 1 \mid a; a^*) \;=\; & \varepsilon a e^*(a) p_1(1,a) p_3(1,1,a^*) + (1 - \varepsilon a e^*(a)) p_1(0,a) p_3(1,0,a^*) \\ & + s_1[\varepsilon a e^*(a) p_1(1,a)(1 - p_3(1,1,a^*)) \\ & + (1 - \varepsilon a e^*(a)) p_1(0,a)(1 - p_3(1,0,a^*))] \\ & + s_2 \varepsilon a e^*(a) p_1(1,a)[(1 - p_3(1,1,a^*)) + 1]. \end{aligned} \tag{20}$$

Observe that the probability $p(x_3 = 1 \mid x_1, x_2; a)$ captures how agent 1 attributes the effect of the confounding factor to the two components 1 and 2 that she can directly influence through her action.

Equilibrium Contracts. Limited liability implies that both agents' wage is zero after a low output, regardless of whether the principal makes agent 1 aware of the confounding

[11] See the appendix for details.

factor or not. Suppose that he offers the wage schedule $w_{1,1}$, $w_{2,1}$ to the two agents. If agent 1 is aware of the confounding factor, the first-order condition for optimal effort is

$$a = \frac{w_{1,1}s_1\alpha_1\gamma(1-\beta)}{\gamma^2 - 2w_{2,1}w_{1,1}\varepsilon^2(1-\beta)^2(s_1\alpha_2+s_2)^2}. \tag{21}$$

If agent 1 is unaware of the confounding factor, the first-order condition for optimal effort becomes

$$\begin{aligned} a \;=\; & \frac{1}{\gamma}w_{1,1} \times [2\varepsilon e^*(a)(p_1(1,a)p_3(1,1,a^*) - p_1(0,a)p_3(1,0,a^*)) \\ & +\alpha_1(1-\beta)p_3(1,0,a^*) + \varepsilon a e^*(a)\alpha_1(1-\beta))(p_3(1,1,a^*) - p_3(1,0,a^*)) \\ & +s_1(2\varepsilon e^*(a)(p_1(1,a)(1-p_3(1,1,a^*)) - p_1(0,a)(1-p_3(1,0,a^*))) \\ & +\alpha_1(1-\beta)(1-p_3(1,0,a^*)) + \varepsilon a e^*(a)\alpha_1(1-\beta)(p_3(1,0,a^*) - p_3(1,1,a^*)) \\ & +s_2(\varepsilon e^*(a)(2p_1(1,a) + \alpha_1 a(1-\beta))(2-p_3(1,1,a^*)))]. \end{aligned} \tag{22}$$

We can show that if the cost parameter γ is sufficiently large, then for each wage $w_{1,1} \in [0,1]$ there is a unique personal equilibrium effort a^* and this effort strictly increases in $w_{1,1}$. For an open set of parameter values the principal strictly prefers to keep agent 1 unaware of the confounding factor; and because of the complementarity both agents may benefit from that.

Proposition 4 (Biased Expectations in Teams) *Consider the principal-agent problem with biased expectations in teams. If the costs of effort (the parameter γ) are sufficiently large, then for an open set of parameter values $(\alpha_1, \alpha_2, s_1, s_2, \beta)$ the principal strictly prefers to keep agent 1 unaware, $\mathcal{R} = \hat{\mathcal{R}}$. Both agents can benefit from agent 1's unawareness.*

The analysis in this subsection replicates the results from Gervais and Goldstein (2007). Both models show that the team incentive problem can be alleviated if one agent is overoptimistic regarding the effectiveness of her effort. In particular, even the biased agent may benefit from her misjudgment if the complementarity between the agents' efforts is sufficiently strong. In contrast to the original model, the biased agent in our version correctly anticipates the equilibrium distribution over outcomes. However, her unawareness about the confounding factor causes her to overestimate the importance of her effort for the final output.

Interestingly, such a bias does not occur for all components in $\mathcal{R}^*$. Suppose that agent 1 is initially unaware of agent 2 and her contribution to the project, node 4, but aware of all other components, $\hat{N} = N^* \setminus \{4\}$. We can show that in this case agent 1 chooses the same effort as if she knew all project components. In fact, this is true for all probability distributions p that factorize according to $\mathcal{R}^*$. Thus, the agent may be unaware of

important aspects of the project and still act rationally in equilibrium. The principal then cannot gain by keeping the agent unaware or educating her. In the next section, we generalize this finding.

5 Behavioral Rationality

In this section, we present a result that characterizes the subset of components $H^* \subset N^*$ the agent has to be aware of in order to behave rationally for all probability distributions p (that are consistent with the objective DAG $\mathcal{R}^*$) and incentive schemes $w(y)$. This means that the agent can ignore all components in the set $N^* \setminus H^*$, i.e., for any given p and $w(y)$, she always acts as if she knew the objective DAG $\mathcal{R}^*$. She still may have a misspecified model, but she correctly anticipates how different actions map into outcomes. This result will get us a corollary that identifies for a large class of DAGs whether the principal can profit from the agent's initial lack of awareness or has a strict incentive to educate the agent. The first step is to translate the concept of behavioral rationality from Spiegler (2016a) to our framework.

Definition 2 *The agent with subjective DAG $\mathcal{R}$ is behaviorally rational if, for any 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 state the characterization, we have to introduce a number of graph-theoretical concepts. Let the DAG be given by $\mathcal{R} = (N, R)$. The skeleton of (N, R), denoted by $(N, \tilde{R})$, is obtained by making the DAG undirected. We have $i\tilde{R}j$ if iRj or jRi. 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 is called the v-structure. A DAG is called perfect if it has an empty v-structure. In the following, we will assume that $\mathcal{R}$ is a perfect DAG (all example DAGs in this paper are perfect). The v-structure of a DAG is essential for its causal properties. In the proof of our main result, we will make heavy use of the following definition and result from the Bayesian network literature.

Definition 3 *Two DAGs $\mathcal{R}$ and $\mathcal{G}$ are equivalent if $p_{\mathcal{R}}(x) = p_{\mathcal{G}}(x)$ for every $p \in \Delta(X)$.*

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

To illustrate, let the objective DAG be given by $\mathcal{R}^*$ from Figure 2. Suppose that the DAG $\mathcal{G}$ is identical to $\mathcal{R}^*$ except that the link $1R^*2$ is missing in $\mathcal{G}$. Then $\mathcal{G}$ has a v-collider at node 3 (and a different skeleton). This new DAG is not equivalent to the

original DAG $\mathcal{R}^*$. There exists a probability distribution p so that the agent's subjective belief $p_{\mathcal{G}}$ would differ from p if her subjective DAG is given $\mathcal{G}$. In contrast, if $\mathcal{G}$ is the same as $\mathcal{R}^*$ with the link $1R^*2$ reversed to $2G^*1$, then $\mathcal{G}$ is equivalent to $\mathcal{R}^*$. We then have $p_{\mathcal{G}}(x) = p(x)$ for all $x \in X$, regardless of p.

We need a few more definitions. A subset of nodes $K \subset N$ is called ancestral if for all nodes $j \in K$ we have $R(j) \subset K$. A path τ of length d from node j to node k is a sequence of nodes $\tau_0, \tau_1, ..., \tau_d$ so that $\tau_0 = j$, $\tau_d = k$, and $\tau_{h-1}\tilde{R}\tau_h$ for all $h \in \{1, ..., d\}$. The length of the shortest path between j and k is called the distance between these nodes and denoted by $d(j, k)$. 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 (this means that nothing influences the agent's action). In each of these DAGs, all active paths between the action node 0 and any node j point towards j. 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 will be a close connection between fundamental links and the nodes that can be removed while still maintaining behavioral rationality.

Definition 4 *Consider two nodes $j, k \in N^*$. If jGk for all $\mathcal{G} = (G, N^*) \in \mathcal{E}$, then the link jGk is called fundamental link and denoted by jEk.*

An intuition for fundamental links is that they capture an empirically relevant direction of causality. Specifically, they describe how the agent's action impacts on other variables. For simple DAGs, we can identify all fundamental links. Consider again the DAG $\mathcal{R}^*$ from Figure 2. Since the action node is ancestral the three links pointing from node 0 to other nodes are fundamental ($0R^*1$, $0R^*2$, and $0R^*4$). Given this, the two links pointing into the outcome node 3 ($1R^*3$ and $2R^*3$) also must be fundamental. If we would turn around one or both of them, we would have a v-collider since there is no link between the action node 0 and the outcome node 3. Hence, the resulting DAG would not be equivalent to $\mathcal{R}^*$. The remaining link $1R^*2$ is not fundamental. This reasoning can be generalized to all perfect DAGs. Our first main result shows that nodes that are connected by fundamental links in perfect DAGs exhibit characteristics that are easy to identify.

Proposition 6 (Characterization of Fundamental Links) *Let $\mathcal{R}^*$ be a perfect DAG and consider two adjacent nodes $j, k \in N^*$. The link jR^*k is fundamental if and only if at least one of the following conditions is satisfied:*

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

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

From this result we can derive an algorithm that identifies all fundamental links in a perfect DAG. First, find for each node j the distance to the action node, $d(0, j)$. Links between nodes of differing distance are fundamental links. Then check the links between nodes j, k that are of equal distance to the action node. Such a link is fundamental if and only if there exists a third node l so that there is a fundamental link between l and j, but no link between l and k. We now 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 τ is active and all the links between neighboring nodes in τ are fundamental.*

Consider again DAG $\mathcal{R}^*$ from Figure 2. The path $\tau = \{0, 1, 3\}$ is a fundamental active path since both links $0R^*1$ and $1R^*3$ are fundamental. In contrast, the active path $\tau' = \{0, 1, 2, 3\}$ is not fundamental since the link $1R^*2$ 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 has to be aware of 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^* obviously have no influence on output or costs, so the agent can safely ignore them. By Proposition 5, the agent knows the true statistical relationship between actions and outcomes if she is aware of the nodes H^*. This is our second main result.

Proposition 7 (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^*.*

Using Proposition 6 and 7 we can identify all nodes the agent needs to be aware of in order to behave rationally. To illustrate, consider the example DAG $\mathcal{R}^*$ from Figure 1. Using Proposition 6, we can derive that all links are fundamental, except $1R^*2$ and $3R^*2$. Thus, in this DAG, the fundamental active paths between action and outcomes are $\tau = \{0, 1, 4\}$, $\tau' = \{0, 3, 4\}$ and $\tau'' = \{0, 5\}$. The set of nodes on fundamental active paths is therefore given by $H^* = \{0, 1, 3, 4, 5\}$. By Proposition 7, the agent may be unaware of component 2 and still act as if she knew the complete project, regardless of the probability distribution p and the incentives scheme $w(y)$.

We can now connect our main results to the contract-theoretic context of this paper. The general question is whether the principal can benefit from the agent's initial unawareness of certain project components. Indeed, if we have a sufficiently rich environment (at least two actions, output an cost levels) and the action affects the output through two different channels of causality, and we can find a component $j \in N^*$ and probability distributions p so that the principal strictly prefers to keep the agent unaware of p; and probability distributions so that the principal strictly prefers to make the agent aware of j. Thus, Proposition 7 can be interpreted as a general description of all environments in which it is profitable for the principal to exploit or educate the agent.

Corollary 1 *Let $\mathcal{R}^*$ be a perfect DAG and assume that $\min\{C\} < \max\{Y\}$. Suppose $\mathcal{R}^*$ contains two fundamental paths τ, τ' between the action node 0 and the output node n as well as nodes j, j' so that j (j') is part of τ (τ'), but not part of τ' (τ). Let the agent's default DAG be given by $\hat{\mathcal{R}} = (N^* \setminus \{j\}, \hat{R})$. Then there exists an open set of probability distributions in $\Delta(X^*)$ so that any equilibrium contract keeps the agent unaware of component j and an open set of probability distributions in $\Delta(X^*)$ so that any equilibrium contract makes the agent aware of component j.*

6 Proof of the Main Results

In this section, we proof Proposition 6 and Proposition 7. We first observe that in a perfect DAG we can use the distance between the action node 0 and two adjacent nodes j, k to determine whether the link between j and k is fundamental. It turns out that this link is fundamental if j and k differ in their distance to node 0.

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

Proof. Let $j, k \in N^*$ be adjacent nodes in $\mathcal{R}^*$ and assume w.l.o.g. that $d(0, j) = d(0, k) - 1$. First, suppose $d(0, j) = 0$ so that $j = 0$. Since node 0 is ancestral, we must have jGk in every DAG $\mathcal{G} \in \mathcal{E}$. Next, suppose $d(0, j) = d > 0$. Since $\mathcal{R}^*$ is perfect and node 0 is ancestral, there exists an active path of length d from node 0 to node j. Denote by l the direct ancestor of j on this path. There cannot exist a link between l and k, otherwise we would have $d(0, j) = d(0, k)$, a contradiction. Thus, we must have jGk in every DAG $\mathcal{G} \in \mathcal{E}$, otherwise we would have a v-collider at node j. Therefore, jR^*k is a fundamental link. ∎

Lemma 2 *Let $j, k \in N^*$ and jR^*k. If there exists a node $l \in N^*$ such that lEj and $l \notin R^*(k)$, then jEk.*

Proof. If there is a fundamental link from node l to node j, then jR^*k implies that we cannot have $kR*l$, otherwise we would have a directed cycle. Node k and node l are therefore not adjacent. Hence, if kGj in some DAG $\mathcal{G} \in \mathcal{E}$, there would be a v-collider at j, a contradiction. Consequently, the link jR^*k must be fundamental. ■

Using Lemma 1 and Lemma 2, we already can proof the "if"-statement of Proposition 6. For the "only if"-statement we need two more results. The first provides a condition under which a link is not fundamental.

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

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

The second result for the proof of the "only if"-statement demonstrates that in a perfect DAG $\mathcal{R}^*$ we can find for any node j an equivalent DAG $\mathcal{G} \in \mathcal{E}$ in which there is no non-fundamental link that points to j. Spiegler (2016b) shows a similar result, i.e., that for each node j in a perfect DAG there exists an equivalent DAG in which j is anchestral. The complication in our case is that the action node 0 must remain ancestral (fundamental links cannot be turned around).
For all j in N* there exists a DAG G in E in which all non-fundamental links adjacent to j points away from j.

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

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 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 $j, k \in N^{[\kappa]}$ there exists a path between j and k consisting of non-fundamental links). **Step 1.** We show that all nodes in a given set $N_d^{[\kappa]}$ have the same parents in N_{d-1}. Define by $R^*(j \mid N_{d-1})$ the parents of node j in set N_{d-1}. Consider two nodes $j, k \in N^{[\kappa]}$ that are connected through the non-fundamental link jR^*k. Since $\mathcal{R}^*$ is perfect, we must have $R^*(k \mid N_{d-1}) \subset R^*(j \mid N_{d-1})$. Since jR^*k is non-fundamental, we must have $R^*(j \mid N_{d-1}) \subset R^*(k \mid N_{d-1})$ so that $R^*(j \mid N_{d-1}) = R^*(k \mid N_{d-1})$. The result follows from the fact that, by definition, all nodes in $N_d^{[\kappa]}$ are connected through non-fundamental links. **Step 2.** Consider two links $j \in N_d^{[\kappa]}$ and $j' \in N_d^{[\kappa']}$ with $\kappa \neq \kappa'$ that are adjacent. Assume w.l.o.g. that jR^*j'. By definition, jR^*j' is a fundamental link. Since all nodes in $N_d^{[\kappa']}$ are connected through non-fundamental links, we must have jEk' for all $k' \in N_d^{[\kappa']}$. This implies that there cannot exist nodes $k \in N_d^{[\kappa]}$ and $k' \in N_d^{[\kappa']}$ so that $k'R^*k$. Otherwise, we would have $k'Ek$ and $k'Ej$ for all $j \in N_d^{[\kappa]}$, a contradiction. **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 $j \in N^*$ and assume w.l.o.g. that $j \in N_d^{[\kappa]}$. 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 (2016b) implies that there exists a DAG $\mathcal{Q}^{[\kappa]}$ in which node j 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 $\mathcal{Q}^*$. Using Step 1 to Step 3, we can show that there are no v-colliders or cycles in $\mathcal{Q}^*$, which proves the result. ■

6.1 Proof of Proposition 6

The "if"-statement follows from Lemma 1 and Lemma 2. We prove the "only if"-statement. Consider any two adjacent nodes $j, k \in N^*$ with jR^*k and $d(0, j) = d(0, k)$. First, suppose that for any node $l \in R^*(j)$ the link lR^*j is not fundamental. By Lemma 4 there exists a DAG $\mathcal{G} \in \mathcal{E}$ in which all non-fundamental links are turned away from node j. Take such a DAG and update it by reversing the link between j and k. By construction, this new DAG is also an element of $\mathcal{E}$. Hence, the link jR^*k is not fundamental. Next, suppose that for any node $l \in R^*(j)$ with a fundamental link lR^*j we also have $l \in R^*(k)$. By Lemma 4, we can find a DAG $\mathcal{G} \in \mathcal{E}$ in which all non-fundamental links are turned away from node j. In this DAG, we have $G(j) \subset G(k)$. From Lemma 3 it then follows that the link jR^*k is not fundamental. This completes the proof.

6.2 Proof of Proposition 7

Before we can proof that behavioral rationality implies the knowledge of all nodes on fundamental active paths, we need two more results.

Lemma 5 *Let $j, k \in N^*$ and $d(0, j) = d(0, k)$. If both nodes are part of a shortest path between 0 and n (or between 0 and $n + 1$), then they are adjacent.*

Proof. Assume w.l.o.g. that the nodes j, k are part of a shortest path between 0 and n. Assume by contradiction that they are not adjacent. Denote by $\tau^{[j]}$ ($\tau^{[k]}$) the shortest active path between 0 and n on which node j (k) lies. Choose those nodes $l^{[j]}$ and $l^{[k]}$ so that (i) $l^{[j]}$ ($l^{[k]}$) lies on path $\tau^{[j]}$ ($\tau^{[k]}$), (ii) $d(l^{[j]}, n) = d(l^{[k]}, n) = d$, (iii) $l^{[j]}$ and $l^{[k]}$ are not adjacent, and (iv) d is minimal. Such nodes exist by the assumption on j and k. Denote by $m^{[j]}$ ($m^{[k]}$) the child of $l^{[j]}$ ($l^{[k]}$) on path $\tau^{[j]}$ ($\tau^{[k]}$). If $m^{[j]}$ and $m^{[k]}$ are identical, then $l^{[j]}$ and $l^{[k]}$ must be adjacent, otherwise there would be a v-collider, a contradiction. Suppose that $m^{[j]}$ and $m^{[k]}$ are not identical. By the assumption in (iv), $m^{[j]}$ and $m^{[k]}$ must be adjacent. Assume w.l.o.g. that $m^{[j]} R^* m^{[k]}$. Then, by the assumption of a shortest active path, we must have $l^{[k]} R^* m^{[j]}$ to avoid the v-collider. But then the nodes $l^{[j]}$ and $l^{[k]}$ must be adjacent, a contradiction. Iterating these arguments shows that the nodes j and k must be adjacent. ■

The next result is crucial for the proof of Proposition 7. It shows that all nodes that are not on a fundamental path between action and outcome nodes can be made "unimportant" in the sense that they have no impact on the 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. To prove Lemma 6 we only need to show that for all nodes $l \in N^* \setminus H^*$ it holds that if there are nodes $j, k \in H^*$ such that $j \tilde{R}^* l$ and $k \tilde{R}^* l$, then we have $j \tilde{R}^* k$. Let $l \in N^* \setminus H^*$ and $j, k \in H^*$. We have to distinguish between five different cases. **Case 1.** Suppose that $j R^* l$ and $k R^* l$. Since $\mathcal{R}^*$ is perfect, we have $j \tilde{R}^* k$. **Case 2.** Suppose that j and k are part of a shortest fundamental path between 0 and n (or 0 and $n+1$). Then according to Lemma 5 we have $j \tilde{R}^* k$. **Case 3.** Suppose that w.l.o.g. node k is not part of a shortest path, $l R^* j$ and $l R^* k$. Then either we have $j R^* k$ – in which case we are done – or there exists a directed fundamental path from k to j. Therefore, node l must be a parent to the direct ancestor of j on the directed path from k to j as $\mathcal{R}^*$ is acyclic and perfect. By reapplying this argument, we find that node l must be a direct ancestor of the direct descendant of k. Denote by h the direct descendant of k on the directed fundamental path between k and j. Since node k is not on a shortest fundamental path, there exists a node $m \in H^*$ such that $m E k$ and not $m R^* h$. Since $\mathcal{R}^*$ is perfect, we have a contradiction. **Case 4.** Suppose that w.l.o.g. node k is not part of a shortest

fundamental path, jR^*l and lR^*k. There cannot be a fundamental directed path from k to j, since that would create a directed cycle. Thus, we have jR^*k. **Case 5.** Suppose that w.l.o.g. node k is not part of a shortest fundamental path, lR^*j and kR^*l. Since node l does not have any fundamental links into H^*, we know that we can find a DAG in $\mathcal{E}$ such that we end up in one of the other three cases. This concludes the proof of Lemma 6. ■

We now can prove Proposition 7. First, we show the "if"-statement. Assume that the agent 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 5 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 $p_{\mathcal{R}}(x_{H^*}) = p(x_{H^*})$ 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 $j \in H^*$ the agent is unaware of. 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 τ be an active fundamental path between the nodes 0 and n with $k \in \tau$. Let h be the h'th node in τ. Consider a probability distribution with the following properties: $p(x_j \mid x_{R^*(j)}) = p(x_j)$ for all $j \notin \tau$ and $p(x_h \mid x_{R^*(h)}) = p(x_h \mid x_{h-1})$. Clearly, such a distribution can have the desired property. This completes the proof.

7 Conclusion

We examined a principal-agent framework in which the principal chooses the variables the agent is aware of. The agent then fits a causal model which comprises these variables to her observations. Her beliefs may depend on her actions. Therefore, we followed Spiegler (2016) and modeled the agent's behavior as personal equilibrium: her equilibrium action is optimal given her beliefs and her beliefs correspond to the empirical distribution over the variables she is aware of, given her equilibrium action. This framework captures a number of models from the literature on contracting with boundedly rational agents in a uniform framework. We demonstrated that several results from this literature (such as the incentive effect or the shrouding of add-on prices) can be recovered in our framework. This implies that the mechanisms behind these results can hold even if the agent collects an infinite amount of data on the variables she is aware of.

Our framework builds on the Bayesian network terminology of Verma and Pearl (1991), Cowell et al. (2007) and Pearl (2009). We therefore can use results from this literature to derive a number of general results. Specifically, we examine which project components

the agent has to be aware of so that she understands the true relationship between her actions and outcomes – even if she only has the data that are generated by her equilibrium action. We show that the agent can be unaware of important project components an still behave rationally. She then has a misspecified model, but correctly anticipates how non-equilibrium actions affect outcomes. However, if the agent is unaware of components that have an immutable causal interpretation, then for some objective probability distributions the principal can profitably exploit this unawareness; and for some probability distributions he has a strict preference for educating the agent.

References

Armstrong, Mark, and John Vickers (2012): "Consumer Protection and Contingent Charges," *Journal of Economic Literature*, 50(2), 477–493.

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

Bénabou, Roland, and Jean Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70(3), 489–520.

Cowell, Robert, A. Philip Dawid, Steffen Lauritzen, and David Spiegelhalter (2007): *Probabilistic Networks and Expert Systems*, Springer, New York.

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

Dekel, Eddie, Barton Lipman, and Aldo Rustichini (1998): "Standard State-Space Models Preclude Unawareness," *Econometrica*, 66(1), 159–173.

Eyster, Erik, and Matthew Rabin (2014): "Extensive Imitation is Irrational and Harmful," *Quarterly Journal of Economics*, 129(4), 1861–1898.

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.

Gabaix, Xavier, and David Laibson (2006): "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, 121(2), 505–540.

Galanis, Spyros (2013): "Unawareness of theorems," *Economic Theory*, 52(1), 41–73.

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.

Heifetz, Aviad, Martin Meier, and Burkhard Schipper (2006): "Interactive unawareness," *Journal of Economic Theory*, 130(1), 78–94.

Heifetz, Aviad, Martin Meier, and Burkhard Schipper (2013): "Unawareness, beliefs, and speculative trade," *Games and Economic Behavior*, 77(1), 100–121.

Inderst, Roman (2001): "Incentive schemes as a signaling device," *Journal of Economic Behavior and Organization*, 44(4), 455–465.

Karle, Heiko, Heiner Schumacher, and Christian Staat (2016): "Signaling Quality through Increased Incentives," *European Economic Review*, 85, 8–21.

Kaya, Ayca (2010): "When does it pay to get informed," *International Economic Review*, 51(2), 533–551.

Kosfeld, Michael, and Ulrich SChüwer (forthcoming): "Add-on Pricing in Retail Financial Markets and the Fallacies of Consumer Education," *Review of Finance*.

Kőszegi, Botond (2014): "Behavioral Contract Theory," *Journal of Economic Literature*, 52(4), 1075–1118.

Koski, Timo, and John Noble (2009): *Bayesian Networks: An Introduction*, Wiley Series in Probability, Wiley.

Martimort, David, and Wilfried Sand-Zantman (2006): "Signalling and the design of delegated management contracts for public utilities," *RAND Journal of Economics*, 37(4), 763–782.

Maskin, Eric, and Jean Tirole (1990): "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica*, 58(2), 379–409.

Maskin, Eric, and Jean Tirole (1992): "The Principal-Agent Relationship with an Informed Principal, II: Common Values," *Econometrica*, 60(1), 1–42.

Pearl, Judea (2009): *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Shrier, Ian, and Robert Platt (2008): "Reducing bias through directed acyclic graphs," *BMC Medical Research Methodology*, 8(70).

Spiegler, Ran (2016a): "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 131(3), 1243–1290.

Spiegler, Ran (2016b): "Can Agents with Causal Misperceptions be Systematically Fooled?," unpublished manuscript.

Spiegler, Ran (forthcoming): "Data Monkeys: A Procedural Model of Extrapolation from Partial Statistics," *Review of Economic Studies*.

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.

Verma, Thomas, and Judea Pearl (1991): "Equivalence and Synthesis of Causal Models," *Uncertainty in Artificial Intelligence*, 6, 255–268.

Wagner, Christoph, Tymofiy Mylovanov, and Thomas Tröger (2015): "Informed-principal problem with moral hazard, risk neutrality, and no limited liability," *Journal of Economic Theory* 159, 280–289.

Appendix

Mathematical Details from Subsection 4.2. We plug the parameter assumptions into the probability framework and get

$$\begin{aligned} p_{\hat{\mathcal{R}}}(c=1 \mid a;a^*) &= (\beta_1+\beta_2(a-a^d)^2)[p(x_2=1 \mid a^*)3k\Delta + p(x_2=0 \mid a^*)2k\Delta] \\ &\quad +(1-\beta_1-\beta_2(a-a^d)^2)[p(x_2=1 \mid a^*)2k\Delta + p(x_2=0 \mid a^*)k\Delta]. \end{aligned}$$

Simplifying this expression gives us

$$p_{\hat{\mathcal{R}}}(c=1 \mid a;a^*) = k\Delta p(x_2=1 \mid a^*) + k\Delta + k\Delta(\beta_1+\beta_2(a-a^d)^2). \tag{23}$$

Using $p_{\hat{\mathcal{R}}}(c=1 \mid a;a^*) = k(a^*)^\beta$ we therefore have

$$p(x_2=1 \mid a^*) = \frac{1}{\Delta}(a^*)^\beta - 1 - (\beta_1+\beta_2(a^*-a^d)^2). \tag{24}$$

We now can plug this expression back into (23) so that we get the desired expression for $p_{\hat{\mathcal{R}}}(c=1 \mid a;a^*)$. ∎

Proof of Proposition 2. The proof proceeds in steps. **Step 1.** The first-best effort level for given k maximizes $p(y=1 \mid a) - ka^\beta$. Since $p(y=1 \mid a)$ is concave, it is uniquely characterized by the first-order condition

$$\frac{d}{da}p(y=1 \mid a) - k\beta a^{\beta-1} = 0 \tag{25}$$

and strictly decreases in k. **Step 2.** We introduce some definitions. Denote by $w^{sb}(y \mid \mathcal{R},k,a)$ the second-best optimal wage schedule that implements the PE equilibrium effort a if the agent is aware of $\mathcal{R}$ and the cost-parameter is k. The corresponding first-best optimal wage schedule is a fixed wage $w^{fb}(y \mid \mathcal{R},k,a) = ka^\beta$. Define by k^d the cost level for which a^d is the first-best action. Define by k_3 the cost level for which action $\underline{a}$ solves (25). Let $\overline{k}$ be the cost level so that $p(y=1 \mid \underline{a}) - \overline{k}\underline{a}^\beta = 0$. Note that $k_3 < \overline{k}$ and that it does not pay off for the principal to employ the agent if $k > \overline{k}$. Let $V(\mathcal{R},sb,k,a)$ be the principal's profit from the second-best wage schedule $w^{sb}(y \mid \mathcal{R},k,a)$, and $V(\mathcal{R},fb,k,a)$ his profit from the corresponding first-best wage schedule. **Step 3.** We show that for all $k \in [k^d,\overline{k}]$ the second-best contract for an unaware agent specifies a fixed-wage of $w^d = k(a^d)^\beta$. Fix some $k \in [k^d,\overline{k}]$. Note that when $\mathcal{R} = \hat{\mathcal{R}}$, then in a PE the agent exerts effort $a^* > a^d$ or does not work for the principal when $w_1 > w_0$. In both cases, we have $V(\hat{\mathcal{R}},fb,k,a^*) > V(\hat{\mathcal{R}},sb,k,a^*)$ since the agent is risk-averse, and $V(\hat{\mathcal{R}},fb,k,a^d) > V(\hat{\mathcal{R}},fb,k,a^*)$ since $p(y=1 \mid a)$ is concave and the cost function convex. If the agent is unaware, then principal can implement a^d with the first-best wage schedule. Hence, we have $V(\hat{\mathcal{R}},sb,k,a^d) > V(\hat{\mathcal{R}},sb,k,a^*)$,

which proves the claim. Using similar arguments, we can show that whenever $k \in [k_3, \overline{k}]$ the unique optimal contract makes the agent aware of the second component, $\mathcal{R} = \mathcal{R}^*$, and offers a fixed wage, $w_1 = w_0 = k\underline{a}^\beta$. **Step 4.** We show that there exists a value $k_2 < k_3$ so that whenever $k \in (k_2, k_3)$ the unique optimal contract makes the agent aware of the second component, $\mathcal{R} = \mathcal{R}^*$, and offers an incentive wage scheme, $w_1 > w_0$. By definition we have $w^{sb}(y \mid \mathcal{R}^*, k_3, \underline{a}) = w^{fb}(y \mid \mathcal{R}^*, k_3, \underline{a})$. By Step 1, we have $V(\hat{\mathcal{R}}, sb, k_3, a^d) < V(\mathcal{R}^*, sb, k_3, \underline{a})$. Note that both $V(\hat{\mathcal{R}}, sb, k, a^d)$ and $V(\mathcal{R}^*, sb, k, \underline{a})$ are continuous in k. The result then follows from Step 3. **Step 5.** We show that there exists a value $k_1 \in (k^d, k_2]$ such that, whenever $k \in [k^d, k_1)$, the unique optimal contract keeps the agent unaware, $\mathcal{R} = \hat{\mathcal{R}}$, and offers a fixed wage of $w_1 = w_0 = k(a^d)^\beta$. At $k = k^d$ the first-best wage schedule is identical to the second-best wage schedule when the agent is unaware, so that $V(\hat{\mathcal{R}}, sb, k^d, a^d) = V(\mathcal{R}^*, fb, k^d, a^d)$. Since the agent is risk-averse we have $V(\mathcal{R}^*, fb, k^d, a^d) > \max_a V(\mathcal{R}^*, sb, k^d, a)$. The claim then follows from the fact that both $V(\hat{\mathcal{R}}, sb, k, a^d)$ and $\max_a V(\mathcal{R}^*, sb, k, a)$ are continuous in k. **Step 6.** We show that there exists a value $\underline{k}$ such that, whenever $k \in (\underline{k}, k^d)$, the unique optimal contract keeps the agent unaware, $\mathcal{R} = \hat{\mathcal{R}}$, and offers an incentive wage scheme, $w_1 > w_0$. Assume that $k < k^d$ and consider the second-best contract for the unaware agent $w^{sb}(y \mid \hat{\mathcal{R}}, k, a)$. Note from the maximization problem in (16) that, starting from a fixed wage, the provision of incentives creates only second-order costs (in terms of the agent's risk-premium), but first-order gains. Hence, the second-best optimal contract for the unaware agent specifies positive incentives. The rest of the proof is analogous to that in Step 5. The proofs of the statements in *iv* and *vii* are straightforward and therefore omitted. ■

Proof of Proposition 3. We consider a symmetric equilibrium in which the agent chooses each principal with equal probability in case of indifference. Suppose that principal $k \in \{1, 2\}$, offers contract contract $(\hat{\mathcal{R}}, w_k(y))$ with $w_k(y_h) = \alpha f\bar{\pi} - \bar{\pi}$ and $w_k(y_l) = \alpha f\bar{\pi}$. We follow the proof in Gabaix and Laibson (2006) and show that there is no alternative contract that yields principal k a strictly positive profit. **Step 1.** Suppose that principal k offers a contract with $\mathcal{R} = \hat{\mathcal{R}}$. By assumption GL, lowering the price $\hat{\pi}_k$ is not profitable for the principal. The aware agent anticipates $\hat{\pi}_k = \bar{\pi}$ and thus exerts substitution effort. If principal k charges $\pi_k > -\alpha f\bar{\pi}$, his profit is zero because both agent types contract with the other principal. If principal k charges $\pi_k \leq -\alpha f\bar{\pi}$, his expected profit is either zero or strictly negative. **Step 2.** Suppose that principal k offers a contract with $\mathcal{R} = \mathcal{R}^*$. If principal k charges $\hat{\pi}_k \leq \frac{c}{f}$, the agent trades with principal k only if $-\pi_k - f\hat{\pi}_k \geq \alpha f\bar{\pi} - c > 0$, where the last inequality follows from the assumption on α. Thus, principal k trades with the agent only at a negative expected profit. If he charges $\hat{\pi}_k > \frac{c}{f}$, the agent trades with him only if $-\pi_k - c \geq \alpha f\bar{\pi} - c > 0$. So there is trade only if principal k charges a negative price for his base good, which

rules out profitable trade. This completes the proof. ■

Additional Results from Subsection (4.3). Suppose that GL does not hold. The agent always observes prices, i.e., payoffs in the two cost-states. We prove the following statement.

Proposition 8 (Shrouded Attributes, GL does not hold) *Consider the shrouded attributes model and assume that GL does not hold. Then in each equilibrium both principals offer a contract $(\mathcal{R}, w_k(y))$ with $w_k(y_l) = f\hat{\pi}_k$, $w_k(y_h) = f\hat{\pi}_k - \hat{\pi}_k$, and $\hat{\pi}_k \leq \frac{c}{f}$. No agent type chooses substitution effort in such an equilibrium.*

Standard arguments show that there is no profitable deviation from the suggested allocation. It remains to show that there is no equilibrium in which an agent type exerts substitution effort. Assume by contradiction that such an equilibrium exists. Then one principal charges $\hat{\pi}_k > \frac{c}{f}$ and some agent type chooses this contract. Competition ensures that principals cannot earn positive profits. We thus have $\pi_k = 0$ and some agent type's payoff in equilibrium is $-c$. A principal can then profitably offer a contract with $\pi_k = \varepsilon$ and $\hat{\pi}_k = 0$. If ε is sufficiently small, this contract would at least some agent type and generate a positive expected profit. This completes the proof. ■

Mathematical Details from Subsection 4.4. We abbreviate $p(x_2 = 1 \mid a, e^*(a)) = p_2(a)$ and $p(y = 1 \mid x_1, x_2, x_3) = p_y(x_1, x_2, x_3)$. When the agent is kept unaware of node 3 and in equilibrium exerts action a^*, her belief about the success probability after action a equals

$$\begin{aligned} p_{\hat{\mathcal{R}}}(y = 1 \mid a; a^*) &= p_2(a)p_1(1,a)[p_3(1,1,a^*)p_y(1,1,1) + (1 - p_3(1,1,a^*))p_y(1,1,0)] \\ &+ (1 - p_2(a))p_1(0,a)[p_3(1,0,a^*)p_y(1,0,1) + (1 - p_3(1,0,a^*))p_y(1,0,0)] \\ &+ p_2(a)(1 - p_1(1,a))[p_3(0,1,a^*)p_y(0,1,1) + (1 - p_3(0,1,a^*))p_y(0,1,0)] \\ &+ (1 - p_2(a))(1 - p_1(0,a))[p_3(0,0,a^*)p_y(0,0,1) + (1 - p_3(0,0,a^*))p_y(0,0,0)], \end{aligned} \tag{26}$$

which can be simplified to equation (20). Abbreviate $p_1(x_2, x_3, a) = p(x_1 \mid x_2, x_3; a)$. We

then can calculate

$$\begin{aligned}
p_3(1,1,a^*) &= \frac{p_1(1,1,a^*)p(x_3=1 \mid x_2=1;a^*)}{p_1(1,1,a^*)p(x_3=1 \mid x_2=1;a^*)+p_1(1,0,a^*)p(x_3=0 \mid x_2=1;a^*)} \\
&= \frac{\beta}{\beta+(\alpha_1 a^*+\alpha_2)(1-\beta)}, \\
p_3(1,0,a^*) &= \frac{p_1(0,1,a^*)p(x_3=1 \mid x_2=0;a^*)}{p_1(0,1,a^*)p(x_3=1 \mid x_2=0;a^*)+p_1(0,0,a^*)p(x_3=0 \mid x_2=0;a^*)} \\
&= \frac{\beta}{\beta+\alpha_1 a^*(1-\beta)}, \\
p_3(0,1,a^*) &= \frac{(1-p_1(1,1,a^*))p(x_3=1 \mid x_2=1;a^*)}{(1-p_1(1,1,a^*))p(x_3=1 \mid x_2=1;a^*)+(1-p_1(1,1,a^*))p(x_3=0 \mid x_2=1;a^*)} \\
&= 0 = p_3(0,0,a^*).
\end{aligned}$$

Insert this into (26). Equation (20) then follows from our assumptions on p. ■

Proof of Proposition 4. To be completed. ■