Causal Inference as an Organizational Problem and Organizational Culture as a Solution*

Michael D. Ryall
University of Toronto

Olav Sorenson
Yale University

April 19, 2019

Abstract

Research on organizations generally presumes that managers have the ability to direct the organization towards some goals. But that presumption depends crucially of the ability of the manager to understand how particular actions or directives might influence organizational outcomes, a problem of causal inference. We develop a formal model of this inference problem and use it to demonstrate that managers can only solve it under very specific conditions. However, organizational culture can extend the ability of the manager to understand the expected results of incentives or policies by mitigating problems that arise when the individual beliefs of other members of the organization influence organizational outcomes and when managers cannot observe those personal beliefs.

INTRODUCTION

Organizations have long been thought to solve a problem of coordination, allowing groups of people to cooperate to produce something that no single one of them could on their own (Parsons, 1960). One sees this idea across the social sciences. Among early sociological writings, Weber (1968) viewed organizations as tools, means to an ends, and analyzed the common structural features that allowed them to operate effectively. Early organizational theorists, such as the administrative and scientific management schools focused on how to optimize the structures of organizations to coordinate the activities of their employees (Taylor, 1911; Simon, 1945). Economic theories of the firm similarly argue that organizations facilitate joint production, particularly that which might not occur otherwise because individuals trying to

*PRELIMINARY AND INCOMPLETE.

collaborate through trade would find it too difficult to split the jointly-created value of their interdependent activities (Coase, 1937; Alchian and Demsetz, 1972).

But the mere existence of an organization, even a formal one, does not guarantee coordination among those affiliated with it. Many, for example, have noted that organizational members vary in their goals and that they may attempt to divert the resources of the organization to achieve them. Political perspectives on organizations highlight the fact that decisions within a firm and firm behavior arise from internal negotiations among those who hold important information and control valuable resources or among those who oversee various subunits of the organization (March, 1962; Pfeffer and Salancik, 1974). In economics, a vast literature on agency theory analyzes the ways in which the interests of employees might deviate from those of the owners of the firm, leading employees to pursue their own ends (Ross, 1973; Jensen and Meckling, 1976).

We call attention to another potential impediment to coordination, the absence of a clear connection between the behavior of the members of the organization and the consequences of their actions for organizational outcomes. Even if everyone has a common goal, coordination requires that the individuals involved have a sufficient understanding of what each must do to achieve that end. In essence, they must understand the causal relationships between internal actions and organization-level outcomes. When organizational members also vary in their interests, an inadequate understanding of these causal relationships exacerbates the difficulty of dealing with any divergent goals. Solutions to agency problems and internal politics have generally been of two forms, either using incentives to align interests or vesting some with authority over the actions of others (e.g., Grossman and Hart, 1986; Milgrom and Roberts, 1992). But determining what behaviors an organization would want to encourage or mandate requires a clear understanding of the relationship between individual-level behaviors and firm-level outcomes.

Others have also called attention to the difficulty of understanding the connections between organizational actions and outcomes, most prominently the classic literature on bounded rationality (March and Simon, 1958; March, 1978). But March and Simon and their followers have been more concerned with the cognitive limits of humans—their ability

to collect and process the information required. Organizations, by dividing goals into subgoals, restricting the information necessary to achieve those subgoals, and by providing a structure for aggregating them, have been seen as a means of ameliorating these constraints (Simon, 1945, 1962). By contrast, the problem of causal identification is one that arises when the data from which to make causal inferences is inadequate to the task. Such inadequacy may be due to unknown causes, confounding causal relationships between objects of interest, and the inability to employ experimental manipulations. Thus, there is no a-priori reason to imagine that the sorts of organizational structures which serve to ameliorate the limits of human cognition have any effect on the identification problem – even with unlimited computational capacity, incomplete data may well befuddle the exercise.

The existence of this inference problem from the perspective of an organizational manager should not surprise social scientists. Much of the focus of scholarly empirical research and methodology over the past two decades has been in terms of trying to understand whether particular factors have causal effects on outcomes of interest (e.g., Morgan and Winship, 2007; Angrist and Pischke, 2009). Even with methodological advances, such as the use of instrumental variables, causal inference is a maddeningly difficult endeavor in social settings. What *is* surprising is that this recognition of the difficulty of causal identification in scholarly research has had little influence on our theoretical understanding of it's effect on social and organizational behavior. Instead, the almost universal – but usually implicit – assumption has been that managers and organization members grasp the essential causal relationships from actions and policies to organizational outcomes. For example, foundational theoretical work on fine-tuning formal organizational structure to ameliorate problems of bounded rationality (sociology) or on refining incentive mechanisms to induce superior profitability (economics) *presuppose* a sufficiently accurate understanding of the effects of such interventions.

In this paper, we take a step back and ask: under what conditions is it possible – at least in principle – for organizational or mechanism designers to have the requisite knowledge to implement effective interventions of these kinds? If managers cannot discern the effects of their interventions from the information available (even without running into constraints on

their cognitive processing abilities), then the intervention exercise is pointless from the start. We develop this question in depth by building a mathematical representation of the causal inference problem in the context of an organization. Given information on some relevant factors but not on others, can a manager isolate the effects of of an employee's behavior on organizational outcomes?[1] Our analysis focuses on what is theoretically possible for the manager, say, using advanced statistical techniques on a complete set of high fidelity data. This presents a best-case scenario from the manager's perspective: if causal identification fails under our generous assumptions, it most certainly fails in more restrictive settings.

Our analysis applies the Bayesian network approach to modeling real-world causal systems. Over the past thirty years, major advances have been made in our understanding of causality and of the conditions necessary to infer the existence of causal relationships using this approach (e.g., Pearl, 1988, 2009). These advances have been influential on the empirical side of sociology, both in terms of helping researchers understand whether and how various empirical designs allow for causal inference (Morgan and Winship, 2007; Elwert, 2013; Elwert and Winship, 2014) and in terms of suggesting novel approaches to causal identification (e.g., Elwert and Christakis, 2006).

This paper proceeds in two main steps. The first is an econometric exercise in which we analyze the identification problem with respect to organizations. We begin with a set of variables that ultimately affect an organizational performance measure of some interest. Some of these variables may represent internal factors (employee actions), some external (competitor actions), some known (intermediate work product), and some hidden (private information of the employee). The system of variables generates data according to a specific pattern of causal interactions. Based upon the structure of these interactions, we determine the extent to which it is possible for a manager to use the data generated to isolate the effects of an employee's behavior on organizational outcomes. Our main result shows that the existence of unknown factors stymies causal inference only in organizations with a very special causal structure. This peculiarity suggests that, at least from a purely mathematical

[1]For expositional clarity, we refer to the person trying to understand these effects as the "manager" and those carrying out the activities of the organization as the "employees." Nevertheless, our findings extend to communities and informal structures without a hierarchy.

perspective, identification is generally possible. Unfortunately, the problematic structure seems common – perhaps even pervasive – in the real world.

Having uncovered an issue that presents serious challenges to real-world managers in the first part of the paper, we move on in the second part to explore some ways in which this issue maybe may be mitigated. Interestingly, we find that two components of organizational culture – (i) joint commitment to a collective belief, and (ii) organizational routines – have the ability to alter the pattern of influence relationships and open the way to accurate inference on the part of the manager. The discovery that culture has the power to reveal important information about these otherwise unknowable features of an organization's workings is new.

Although organizational culture has long been seen as a means of coordinating the members of an organization (Ouchi, 1981; O'Reilly and Chatman, 1996), the usual interpretation has been that a strong culture – either through selection or social influence (enculturation) – leads individuals to sublimate their own interests and values to those of the group. But these values are fundamentally unobservable and, even if they become congruent with those of the collective, they still interact with employee beliefs in determining behavior. Organizational culture, in the sense of common values, therefore, does not and cannot solve the causal inference problem.

The cultural components that can solve the causal inference problem must be public, observable. Our formulation of collective beliefs, for example, follows recent developments in social ontology (Gilbert, 2014). Having a collective belief in our model means that the members of the organization jointly commit to emulate, through the actions of each, a single entity operating according to that belief.[2] Members of the organization need not *actually* hold the belief. Instead, they choose their actions such that behavior at the organizational level emulates that of an individual actor holding the collective belief (Gilbert, 2014).[3] Re-

[2]If *all* members of the organization actually did hold the same beliefs, that would lead to similar results. But formal models of enculturation processes suggest that such convergence would require long periods of time and could prove sensitive to employee turnover and environmental drift (Harrison and Carroll, 1991; March, 1991; Harrison and Carroll, 2006).

[3]For example, Miller Valentine Group, an integrated real estate company, claims that it is a corporation that believes in accountability (https://mvg.com/about-us/corporate-beliefs/). By Gilbert (2014), this does not mean that all employees literally believe in accountability. Rather, what is required is that each, in their various roles across the company, behave in such a way that the organizational behavior of Miller Valentine Group emulates an actor that believes in accountability.

placing private beliefs with a collective belief as the driver of individual actions solves the identification problem. Managers can observe actions and compare them to behaviors consistent with the public standard.

Our definition of organizational routines proceeds along more traditional lines. Routines represent rules and procedures for action—"if, then" operations understood by most members of the organization. Once again, the reason that adopting these routines solves the causal identification problem stems from the fact that they replace unknown beliefs of organizational members with concrete routines (which have an explicitly public nature) as the essential drivers of activities.

Organizational culture may, of course, contribute to organizational performance in other ways. To the extent that employees do adopt the values of the organization or at least to the extent that they feel a stronger sense of community, the members of strong culture organizations may exert more effort and exhibit higher levels of commitment to the firm (e.g., O'Reilly, 1989). But our analysis suggests that culture may also have value for a hitherto unrecognized benefit: By enabling managers to understand better the causal connections inherent in the organization and its environment, organizational culture allows those managers to more accurately anticipate the consequences of their directives, thereby permitting them to direct the activities of their employees more effectively.

Although our model uses the language of a manager and employee throughout, both the issue of causal identification and the idea that culture can solve this problem extend to situations where no one has been vested with formal authority. In a small community, for example, each individual interested in forwarding the goals of the group must anticipate how others will act and how their own actions will interact with those of other members of the community in realizing them. We elaborate on how these results can extend to such informal organizations in the discussion section.

THE INFERENCE PROBLEM

Organizations can take a variety of forms, from the formal for-profit entities that dominate the production and distribution of goods and services to the non-profit organizations that support various communities and even to informal groups of individuals. Across all of these diverse forms, the individuals involved, the members, attempt to coordinate their actions to forward group-level goals. That coordination, in turn, requires that the members of these organizations – particularly managers responsible for coordinating the activities of multiple team members – have a sense of the causal relationships between their actions and the organization-level outcomes.

To understand better the nature of this inference problem, we develop a formal model based upon Bayesian network theory. Over the last thirty years, our understanding of causal identification has grown by leaps and bounds. An important aspect of that progress has been the effort that Judea Pearl and others have put into developing a mathematical language for representing and analyzing stochastic causal systems (e.g., Pearl, 1988, 2009). The Bayesian network depicts the system as a directed, acyclic graph (DAG), in which the nodes represent random variables of interest and the directed edges indicate direct influence relationships. The instantiation of a node's direct causes determines the probability distribution by which its own state is generated. A variable's direct causes directly influence its stochastic behavior. Variables that directly influence a node's direct causes indirectly influence the node, and so on. These models are parameterized by associating each node with an appropriate conditional probability table (as we illustrate momentarily).

Building on this infrastructure, our model treats the organization as a system of relationships between variables. In any organization, a variety of factors interact to produce the outcomes associated with the activities of the organization. Some of these variables reflect the internal operations of the organization, such as the actions of its members, the consequences of those actions, and the goals of the collective. Some factors are external to the organization, residing in the environment surrounding it. Examples of environmental factors might include sources of private information to members, the actions and prefer-

ences of buyers and suppliers, the institutional entities surrounding the organization, and the competitive and cooperative actions of other organizations.

For the purposes of our formal model, the essential requirement is to classify these factors in terms of managerial awareness. We assume that managers have information on the nature of a collection of factors, which we refer to as *known variables*. This would include the potential states of those variables and their statistical behavior (perhaps resulting from a sufficient period of data collection). By contrast, managers do not have extensive information on a separate collection of factors, which we term *unknown variables*. In the extreme case, they may not even be aware of the existence of these factors.

As an aside, we note that agency problems – a divergence in the goals of the manager and the employee – are a special case of our model. Managers cannot observe the personal utility functions of employees and they may have difficulty observing employees' actions. As such, they would be represented as unknown variables. Interestingly, the extent to which unknown variables prove problematic in the assessment of the causal effects of known variables upon each other depends upon the pattern of extant influence relations as summarized by the system's DAG.

Together, known and unknown variables are assumed to form a comprehensive system: the addition of more variables does not increase the explanatory accuracy of the model. Even though it is possible to include deterministic relationships in our setup, we assume all variable outcomes are generated by strictly positive distributions on their states. This has no material effect on our analysis but does allow us to avoid tedious mathematical details. Moreover, this seems reasonable since uncertainty is a fundamental feature of organizational behavior given the complexity of such systems in the real world. To the extent that the Bayesian network provides a complete enumeration of the factors that determine an organization's behavior and an accurate summary of the influence relations between them, it represents the best that even someone omniscient could achieve in understanding its dynamics. We refer to this system of known and unknown variables linked by their influence relations as the *objective model* of the organization.

The managerial problem is as follows. One of the variables represents organizational per-

formance – some states of which the managers wishes to induce with maximum likelihood by manipulating some of the variables that directly or indirectly influence it. The manager knows the joint probability distribution on the known variables but nothing about the behavior of the hidden variables. We want to know: is this knowledge sufficient to compute the effect of manipulating a given (known) variable on the expected outcome of the performance variable?

Three Simple Cases

Before developing a general answer to the econometric identification question, we illustrate all the essential ideas with three stripped-down, simple cases. In each case, we pit a manager trying to assess the effects of a single employee's actions on organizational performance. The system consists of three known and one unknown variables. The known variables are: the employee's action, denoted A (e.g., selecting the number of orders to fill simultaneously); an intermediate factor influenced by the employee's action, denoted Ω (e.g., the accuracy of order fulfillment), and performance, denoted Π (e.g., revenues from orders shipped less the cost of labor and shipment corrections). The unknown variable is the state of some factor relevant to the employee's choice of action, known by the employee but not the manager, denoted Θ (i.e., the state of Θ is the employee's private information).

We assume that the manager has the power to order the employee take a particular action and, when so ordered, the employee obeys. She knows the joint probability distribution on the known variables as generated by the unperturbed system (i.e., prior to any intervention on her part). We wish to determine under what conditions, if any, this knowledge is sufficient to allow the manager to identify which employee action maximizes the expectation of desired performance. Note well that this is a best-case scenario for the manager: she has perfect knowledge of the statistical behavior of the known variables and a perfect ability to direct the employee's actions. If she cannot succeed under these conditions, then she certainly fails in more realistic situations involving faulty data and less responsive employees.

Figure 1 then provides a graphical representation of the first case we wish to examine. We adopt the convention that dashed circles represent the manager's unknown variables, those

not known, while the solid circles denote the known variables, those known. To keep things simple, assume all the variables are binary; e.g., $A = \{0, 1\}$. Arbitrary states are indicated by small letters; e.g., $a \in A$, $\omega \in \Omega$, etc. In all three of the cases covered in this section, assume the manager wants to maximize the probability that $\pi = 1$.

The arrows indicate the *influence relationships* between these variables: for example $A \to \Omega$ means that, for at least one pair of states a and a', the associated probabilities on the states of Ω differ—different actions lead to different probability distributions on Ω. Throughout our analysis, causal relationships are intended to represent real world primitives: the stochastic behavior of a variable is fully determined by the states of its direct causes. Thus, a variable's behavior can be described by a local stochastic law, also a primitive of the model, elaborated by a conditional probability table as illustrated below. We denote the local stochastic law that determines A as ρ_A and write $\rho_A(a|\theta)$ to indicate the probability that the employee takes action a given private information θ.

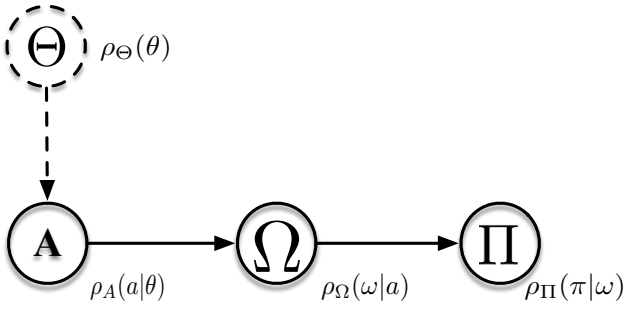

Figure 1: A graphical model of the first case.

For example, the employee might be the marketing professional in a firm whose action options, A, are whether or not to offer a discount to a customer segment. The immediate consequence of the discount decision then influences the market share attained in that segment, indicated by Ω (say, high or low). The market share outcome, in turn, influences organizational performance, in this case profit, Π (again, high or low).[4] In this version of the example, performance uncertainty allows for factors outside the model that have independent

[4]The local law ρ_Π captures the process governing the relationship between the intermediate consequences of employee actions and organizational outcomes. Treating the outcomes of interest as a single performance variable does not imply single-dimensional measures of performance: the states may represent multidimensional outcomes. In all cases, we assume the manager can at least partially order the states of the performance variable according their desirability.

effects beyond market share (e.g., macroeconomic factors). What about Θ? In this case, Θ is a factor that influences the behavior of the marketing professional but has no implications for any other variables. Thus, it may represent employee qualities (his "type") such as whether he is hard-working or lazy, highly or poorly skilled, etc.; or it could represent an aspect of his subjective state of mind, such as his prior beliefs about the efficacy of discounts; or it could represent an external event, such as whether or not he was a victim of road rage on the way in to work. The dashed lines indicate that Θ is "unknown" - a classification about which we say more below.

Can the manager manage in this case? When does the manager have a sufficiently good understanding of the effects of employee actions on organizational outcomes to implement changes that lead to the desired outcomes (or at least to a higher probability of them)? We assume the manager can implement a very effective intervention: she induces the employee to select a specific action, a, simply by ordering him to do so. This best-case assumption, which eliminates many real-world problems such as incentive misalignment or a misunderstanding of directions on the part of the employee, is appropriate given our focus on the inference problem that *precedes* any problems of implementation or mechanism design.[5]

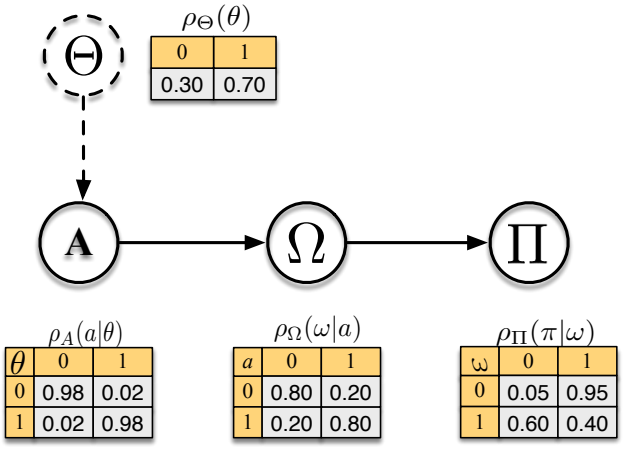

Figure 2: Parameters for the first case.

The qualitative pattern of causal relations illustrated by the DAG in Figure 1 can be

[5]As we point out later, it is easy to generalize our results by switching the manager's decision variable from employee actions to a policy choice that, in turn, influences employees' actions.

augmented by a set quantitative stochastic laws to create a complete *causal system*. Figure 2 illustrates a such a system. A complete causal system generates a joint probability distribution (hereafter, *JPD*) on the variables, which can be computed in a straightforward fashion. In this case specifically, for each profile of outcomes, (θ, a, ω, π), it can be shown that

$$P(\theta, a, \omega, \pi) = \rho_{\Pi}(\pi|\omega)\rho_{\Omega}(\omega|a)\rho_{A}(a|b)\rho_{\Theta}(\theta), \tag{1}$$

where P is the JPD on system-level outcomes (i.e., the various outcome configurations of all four variables). We call P the *status quo* distribution because it describes the stochastic behavior of the system absent any intervention by the manager.

In this example, there are 16 possible system-level outcomes (= 2^4). Consider Table 1. Part (a) reports the status quo distribution P generated by the system articulated in Figure 2. It shows the probability of each configuration of all four variables in the system. From this quantification of P, any implied probabilistic quantity can be computed. To take a few examples, $P(a = 1) = 0.692$, $P(a = 0|\theta = 0) = 0.98$ and $P(\omega = 1|\pi = 0) = 0.95$. Part (b) of Table 1 reports the *marginal distribution on known variables* (hereafter, *MDK*), or $P(a, \omega, \pi)$. Each row indicates the probability of one of the eight possible outcome configurations of the known variables. We assume the manager knows the content of Part (b) – perhaps through a long tenure in that job, or from data acquired as a result of a formal TQM program, or even from being aware of the variable and assigning prior probabilities to its states.

By classifying Θ as *unknown*, we mean that the manager is incapable of forming beliefs about the probabilistic relationship between it and the other, known, variables. This may be due to "ambiguity" (an inability to form probabilities about a factor of which the manager is aware, i.e., in the sense of Knight) or it may be that the manager simply has no awareness of the factor at all. Note well, therefore, the "known" variable designation is broader than the class of "observed" variables. The known variables represent precisely that collection of factors upon which the manager can make some assessment about their joint probabilistic behavior.

For example, suppose Θ represents something about the employee's private beliefs and

motivations (what economists would refer to as his "type"), where $\theta = 1$ is a person motivated to behave in accordance with the performance goals of the organization and $\theta = 0$ is an easily disgruntled type who often enjoys undermining performance. Then, intuition suggests that by *observing* a, the manager leans something about the employee's motivation. For example, the manager may have a good sense of the relative mix of Θ types in the general population and from this, given observation of an actual action a by the employee, the ability to compute $P(\theta|a)$. This is perfectly fine and allowed by our theory. However, if this conditions do hold, then Θ counts as a *known* variable. If the manager is unable to say anything about the probabilistic character of Θ, then she really cannot – in any precise sense – infer the probabilities of its states by observing employee actions.

(a) Complete JPD

Θ	A	Ω	Π	P
0	0	0	0	0.0118
0	0	0	1	0.2235
0	0	1	0	0.0353
0	0	1	1	0.0235
0	1	0	0	0.0001
0	1	0	1	0.0011
0	1	1	0	0.0029
0	1	1	1	0.0019
1	0	0	0	0.0006
1	0	0	1	0.0105
1	0	1	0	0.0017
1	0	1	1	0.0011
1	1	0	0	0.0069
1	1	0	1	0.1304
1	1	1	0	0.3293
1	1	1	1	0.2196

(b) Resulting MDK

A	Ω	Π	P
0	0	0	0.0123
0	0	1	0.2340
0	1	0	0.0370
0	1	1	0.0246
1	0	0	0.0069
1	0	1	0.1315
1	1	0	0.3322
1	1	1	0.2215

(c)
$$P(\pi = 1) = .6116$$
$$P(\pi = 1|a = 1) = .5100$$
$$P(\pi = 1|a = 0) = .8400$$

Table 1: Part (a) shows the full joint probability distribution generated by the causal system in Figure 2. Part (b) shows the marginal distribution on observables implied by P. Part (c) reports the probability that $\pi = 1$ as well as the conditional probabilities that $\pi = 1$ given $a = 1$ and $a = 0$, respectively.

The marginal distribution on known variables (b) is computed by "integrating out" the

unknown variable:

$$P(a, \omega, \pi) = \sum_{\theta \in \Theta} P(\pi|\omega)P(\omega|a)P(a|\theta)P(\theta). \tag{2}$$

In terms of causal knowledge (i.e., what the manager knows about the DAG), we simply assume she knows that the employee's action has some effect on performance. That is, minimally, she knows there is a directed path of some kind from A to Π but, beyond that, she may be ignorant of the true causal relations as summarized by the DAG.[6]

Part (c) of the table calculates probabilistic quantities of obvious interest to the manager. Unperturbed (by any managerial intervention), the probability that the desired performance result occurs is $P(\pi = 1) = .6116$. The manager can order the employee to do $a = 1$ or $a = 0$. From the MDK shown in Part (b), the manager can compute the probabilities that $\pi = 1$ conditional upon the alternative actions: $P(\pi = 1|a = 1) = .5100$ and $P(\pi = 1|a = 0) = .8400$. This suggests that performance can be improved by ordering the employee to do $a = 0$. However, the manager may be concerned, having heard somewhere that, "correlation does not imply causation." As it turns out, this concern is generally well-founded. What we have shown at this point is the connection between the status quo causal system and the JPD and MDK it generates. We have yet to show what happens to the behavior of the system when the manager makes an intervention. This is the task to which we now turn.

Recall, the question of interest is whether the MDK is sufficient to isolate the effects of the employee's actions on performance. If the answer is yes, then the manager who knows the MDK knows precisely what happens to performance when she orders the employee to take one action or the other. In this example, it turns out that the quantities calculated in Part (c) *do* provide a precise quantification of the effects of the employee's actions on performance. Perhaps this seems obvious: since we have assumed the manager knows causality runs from A to Π and not visa versa, why not simply compute $P(\pi|a)$ to quantify the effects of A on

[6]She may well know more. The point is that she correctly believes A has an effect on Π. That is, even though $P(A|\Pi)$ is easy enough to compute from the JPD on observables, she does not take this to imply a reverse-causality story in which Π causes A. In dynamic settings, it may well be that Π in period t does affect A in period $t+1$. Later, when we present the general setup, this would be accommodated via the use of time-stamped variables.

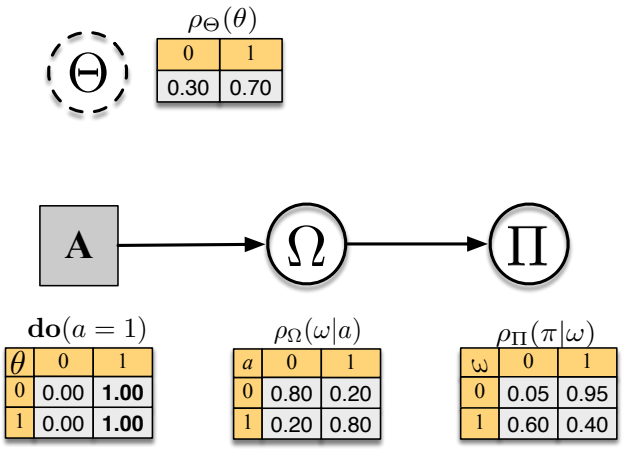

Figure 3: The effect of the intervention $\mathbf{do}(a = 1)$.

Π? To answer this question, we need to dig deeper into the example to understand precisely what is going when an intervention happens and to compare this to the features of the unperturbed system.

When the manager directs the employee to "do $a = 1$," her intervention actually changes the causal system as shown in Figure 3. The forced intervention removes the influence of Θ on A – the employee must now do $a = 1$ for certain, regardless of the state of Θ. (In the jargon of Bayesian network theory, the intervention "mutilates" the status quo causal system.) Thus, the intervention generates a new, system-level JPD, which we indicate as $P_{\mathbf{do}(a=1)}$ This as shown in Table 2, Part (a), with the implied MDK shown in Part (b). From (b), the actual effect of the intervention can be computed (which we denote as shown). Comparing this to Part (c) of Table 1, we see that $P_{\mathbf{do}(a=1)}(\pi = 1) = P(\pi = 1|a = 1)$. Working through the arithmetic, it can be shown that this equivalence holds for the intervention $\mathbf{do}(a = 0)$ as well.

To see exactly why this equivalence arises, let us return to the status quo system. From the system-level JPD, the first thing to note is that Π is statistically dependent on Θ. For example, $P(\pi = 1|\theta = 1) = 0.9985$ but $P(\pi = 1|\theta = 0) = 0.8335$. This dependence arises indirectly through the causal chain $\Theta \rightarrow A \rightarrow \Omega \rightarrow \Pi$. The second thing to note is that,

(a) JPD for $\mathbf{do}(a = 1)$

Θ	A	Ω	Π	$P_{\mathbf{do}(a=1)}$
0	0	0	0	0.000
0	0	0	1	0.000
0	0	1	0	0.000
0	0	1	1	0.000
0	1	0	0	0.000
0	1	0	1	0.000
0	1	1	0	0.000
0	1	1	1	0.000
1	0	0	0	0.003
1	0	0	1	0.007
1	0	1	0	0.057
1	0	1	1	0.133
1	1	0	0	0.144
1	1	0	1	0.336
1	1	1	0	0.096
1	1	1	1	0.224

(b) Resulting MDK

A	Ω	Π	$P_{\mathbf{do}(a=1)}$
0	0	0	0.000
0	0	1	0.000
0	1	0	0.000
0	1	1	0.000
1	0	0	0.010
1	0	1	0.190
1	1	0	0.480
1	1	1	0.320

(c) $P_{\mathbf{do}(a=1)}(\pi = 1) = .5100$

Table 2: Part (a) shows the JPD on the system as a whole generated by the intervention shown in Figure 3. Part (b) shows the resulting marginal distribution on observables. Part (c) shows the resulting probability that $\pi = 1$.

according to P, Π is statistically *independent* of Θ *given* A. For example,

$$P(\pi = 1|\theta = 0, a = 1) = P(\pi = 1|\theta = 1, a = 1) = P(\pi = 1|a = 1) = 0.5100.$$

Moreover, this is true for all values of A and Θ. One way of putting this is that P is such that, with no other knowledge, being told the state of Θ refines one's assessment about the behavior of Π; however, knowledge about the state of Θ adds *no refinement whatsoever* to one who already knows the state of A. Thus, even though the manager does not know the state of Θ (indeed, she may not even know of its influence on the employee or even of its existence), her use of the MDK to compute $P(\pi|a)$ gives her a precise assessment of the effect of A on Π. This feature of P is crucial in our example because it mirrors what happens when the manager intervenes to set A: she forcibly severs the effect of Θ on Π.

In this example, then, we say that the effect of A on Π is *identifiable* from the information available to the manager. Although the example for Figure 1 is limited to binary variables

with a specific set parameters for their conditional probability tables, it turns out that this result generalizes to variables with any finite number of states and any set of (positive) stochastic laws with which one might associate with them. The critical fact is that, for any such variables and corresponding parameterization, Π is independent of Θ given A. Therefore, computing the probability of Π given A from the MDK provides an accurate assessment of the pure effect of A on Π. This is consistent with the effect of an intervention on A. Of course, actually estimating these probabilities may not be a trivial exercise but, at least in theory, the manager can determine what action is desired of employee in this case. The following proposition summarizes this claim (a proof appears in the Appendix).

Proposition 1. Given a collection variables with arbitrary numbers of states parameterized by any positive stochastic laws consistent with Figure 1, the effect of an intervention $\mathbf{do}(a)$ on organizational performance is identifiable for all $a \in A$.

Confounding variables

Although Proposition 1 seems like good news for the manager – and indeed, it is – the positive finding depends on a very specific pattern of influence relationships between the known and unknown variables: in that case, the unknown variables only influence performance through the actions of the employee. Let us next consider a case in which both the beliefs of the employee and the intermediate consequences of the employee's actions depend on a common set of environmental conditions (Θ).

Figure 4 depicts a system with just such a structure. For example, in this case, the employee could be a salesperson whose private information is the facial expression of a customer to whom he is pitching a sale. The expression might reflect something about the customer's state of mind, which, in turn, influences which of two alternative pitch approaches is likely to be most effective. Alternatively, the employee may be a purchasing agent who has private information about the reliability and timeliness of a particular supplier. Based on that information, the employee might form beliefs about the appropriate levels of inventory to carry or about whether to maintain relationships with multiple suppliers. The reliability and timeliness of the supplier, in turn, influences the costs and benefits associated with maintain-

ing a particular inventory level or with spreading purchasing across more than one supplier. More generally, this structure corresponds to the setting used in many economic analyses of Principal-Agent problems – the employee is the agent who receives private information about the efficacy of his actions on intermediate factors, such as product quality, cost, or quantity which, in turn, affect profitability.

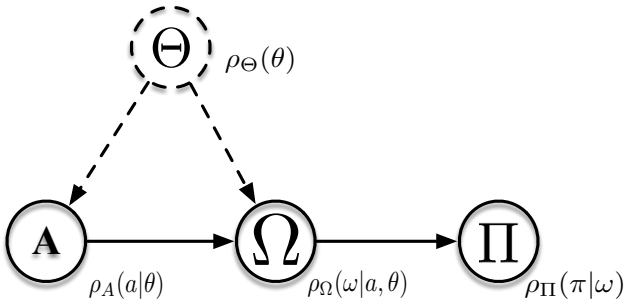

Figure 4: An organization with a confounding environmental variable.

Although the DAG in Figure 4 only differs from the previous example by the addition of a single link, the favorable result of Proposition 1 now fails. In this case, Θ is a *confounding* variable – it is not possible for the manager to use the MDK to make an accurate assessment of the effect of the employee's actions on performance. Once again, let us illustrate with the specific numerical example shown in Figure 5. Here, the computation of P is given by:

$$P(\theta, a, \omega, \pi) \equiv \rho_\Pi(\pi|\omega)\rho_\Omega(\omega|a,\theta)\rho_A(a|\theta)\rho_\Theta(\theta). \tag{3}$$

Although this equation appears quite similar to that of (1), it does differ in one important respect: θ now appears as an argument in ρ_Ω.

Table 3 reports the system-level JPD (the first and second boxes) and the MDK (the third box). Note that, while the JPD generated by this causal system is different than the one shown in Table 1, the MDK is the same. Thus, a manager faced with knowledge of this MDK would be unable to distinguish between the generating structures in Figure 2 and Figure 5. Here, as in the previous example, then, $P(\pi = 1|a = 1) = 0.51$ and $P(\pi = 1|a = 0) = 0.84$. Unfortunately, these assessments no longer indicate the isolated effect of A on Π.

The effect of A on Π is now confounded by the fact that Θ is a direct cause of A and Θ

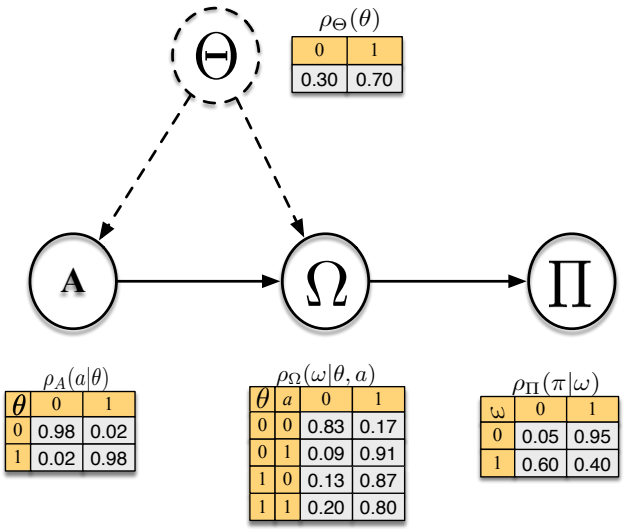

Figure 5: A numerical example of the organization in Figure 4.

System-Level JPD										MDK			
Θ	A	Ω	Π	P	Θ	A	Ω	Π	P	A	Ω	Π	P
0	0	0	0	0.0120	1	0	0	0	0.0010	0	0	0	0.0123
0	0	0	1	0.2323	1	0	0	1	0.0017	0	0	1	0.2340
0	0	1	0	0.0297	1	0	1	0	0.0072	0	1	0	0.0370
0	0	1	1	0.0198	1	0	1	1	0.0048	0	1	1	0.0246
0	1	0	0	0.0003	1	1	0	0	0.0069	1	0	0	0.0069
0	1	0	1	0.0005	1	1	0	1	0.1310	1	0	1	0.1315
0	1	1	0	0.0032	1	1	1	0	0.3290	1	1	0	0.3322
0	1	1	1	0.0022	1	1	1	1	0.2193	1	1	1	0.2215

Table 3: The JPD and MDK implied by Figure 5

and A are both direct causes of Ω. This causal triangle makes it impossible to disentangle the isolated effect of A on Ω using only the information in the MDK. Because Ω then influences Π, it is therefore impossible to isolate the effect of A on Π. This is intuitive, but difficult to demonstrate for an arbitrary set of parameters (see the proof of Proposition 2). However, working through the actual interventions in the same manner as before, it can be shown that in this case $P_{\mathbf{do}(a=1)}(\pi = 1) = 0.4927$ and $P_{\mathbf{do}(a=0)}(\pi = 1) = 0.5871$. Recalling that, left alone, the probability that $\pi = 1$ is 0.6116 ($P(\pi = 1) = 0.6116$ as in the previous example), the manager would do well to simply leave the employee alone!

To further illustrate the perniciousness of the problem here, consider the alternative parameters depicted in Figure 6. The resulting JPD and MDK are elaborated in Table

4). This system generates exactly the same MDK as the one in Figure 5. As before, $P(\pi = 1) = 0.6116$. Now, however, it can be shown that $P_{\mathbf{do}(a=1)}(\pi = 1) = 0.5134$ and $P_{\mathbf{do}(a=0)}(\pi = 1) = 0.8466$. In other words, even though the causal relations and MDK are identical between the causal systems shown in Figures 5 and 6, the intervention implications are quite different. Since the manager has no ability to discern which of these alternatives might represent the true model, given the information available to her (the known variables), she cannot determine how best to direct the employee.

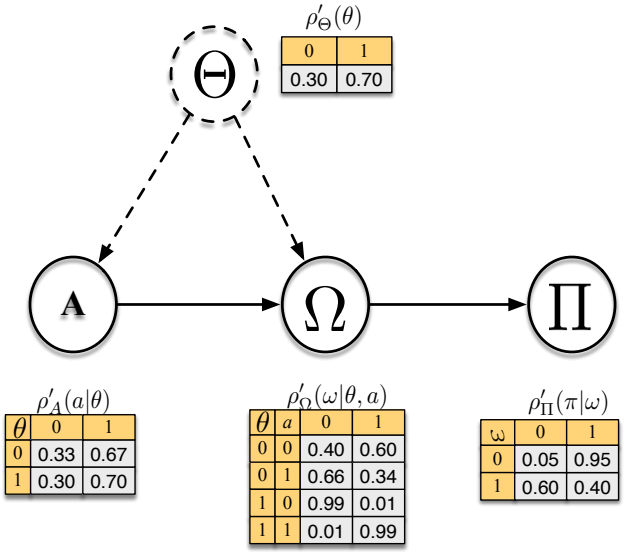

Figure 6: A second numerical example of the organization in Figure 4.

System-Level JPD										MDK			
Θ	A	Ω	Π	P	Θ	A	Ω	Π	P	A	Ω	Π	P
0	0	0	0	0.0020	1	0	0	0	0.0104	0	0	0	0.0123
0	0	0	1	0.0372	1	0	0	1	0.1967	0	0	1	0.2340
0	0	1	0	0.0358	1	0	1	0	0.0013	0	1	0	0.0370
0	0	1	1	0.0239	1	0	1	1	0.0008	0	1	1	0.0246
0	1	0	0	0.0069	1	1	0	0	0.0020	1	0	0	0.0069
0	1	0	1	0.1268	1	1	0	1	0.0047	1	0	1	0.1315
0	1	1	0	0.0406	1	1	1	0	0.2916	1	1	0	0.3322
0	1	1	1	0.0271	1	1	1	1	0.1944	1	1	1	0.2215

Table 4: The JPD and MDK implied by Figure 6

This result also turns out to hold for any organizational system with the structure shown in Figure 4. Consider an arbitrary model leading to a joint probability distribution P as

defined by (3) along with an arbitrary intervention $\mathbf{do}(a)$. As in the examples in Figures 5 and 6, construct a second set of parameters to generate a new joint probability distribution, $P' \neq P$), such that the MDKs are the same. Note that the arrow connecting Θ to Ω in Figure 4 implies that $P(\omega|a,\theta) \neq P(\omega|a,\theta')$ for at least one pair of states θ and θ' (by our definition of an influence relationship). As a result, for some $\pi \in \Pi$, $P_{\mathbf{do}(a)}(\pi) \neq P'_{\mathbf{do}(a)}(\pi)$. The double effect of Θ on Ω (directly and indirectly through A) prevents the manager from being able to isolate the effect of the employee's action on performance. The effect of the employee's action on performance is not identifiable. We summarize this result in our second proposition (once again the full proof appears in the Appendix).

Proposition 2. Given a collection variables with arbitrary numbers of states parameterized by any positive stochastic laws consistent with Figure 4, the effect of an intervention $\mathbf{do}(a)$ on organizational performance is not identifiable for any $a \in A$.

Identification under a hidden common cause

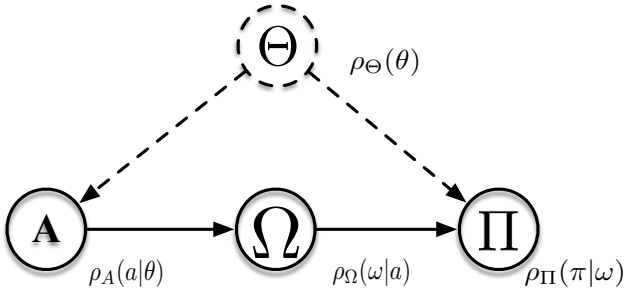

Figure 7: Organizational model with hidden common cause on A and Π.

Given the issues outlined above, one might expect that hidden common causes – i.e., those which influence both the employee's actions and organizational performance – always preclude causal identification by the manager, regardless of the specific structure of the underlying patterns of influence in the organizational system. This intuition turns out to be wrong. Consider the system depicted in Figure 7. Here, the hidden factor Θ is a common cause of A and Π. It turns out that the manager *can* deduce the effects of the actions of the employee from any MDK generated by this system. This is due to a couple of important differences between this system and those of the preceding, problematic cases. First,

Ω is independent of Θ given A; i.e., $P(\omega|a,\theta) = P(\omega|a)$. Second, the structure of influence relationships also implies that the probability of Π is independent of A given Ω; i.e., $P(\pi|a,\omega) = P(\pi|\omega)$. Together, these facts imply that the conditional probability of Π given A accurately summarizes the effects of interventions on A: from the MDK, we know the effect of A on Ω, independent of Θ, and of Ω on Π, independent of A. Therefore, we know the effect of intervening to fix the state of A on Ω and we can trace this through to the independent effect of Ω on Π.

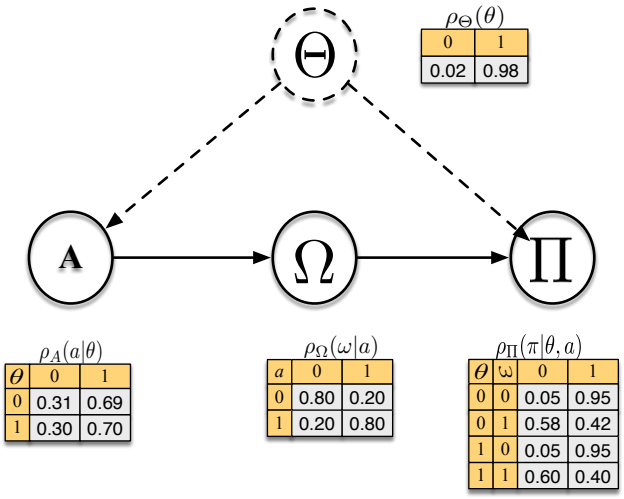

Figure 8: A numerical example of the organization in Figure 7.

System-Level JPD										MDK			
Θ	A	Ω	Π	P	Θ	A	Ω	Π	P	A	Ω	Π	P
0	0	0	0	0.0020	1	0	0	0	0.0121	0	0	0	0.0123
0	0	0	1	0.0041	1	0	0	1	0.2299	0	0	1	0.2340
0	0	1	0	0.0006	1	0	1	0	0.0363	0	1	0	0.0370
0	0	1	1	0.0005	1	0	1	1	0.0242	0	1	1	0.0246
0	1	0	0	0.0001	1	1	0	0	0.0068	1	0	0	0.0069
0	1	0	1	0.0022	1	1	0	1	0.1293	1	0	1	0.1315
0	1	1	0	0.0054	1	1	1	0	0.3268	1	1	0	0.3322
0	1	1	1	0.0040	1	1	1	1	0.2175	1	1	1	0.2215

Table 5: The joint probability distribution implied by Figure 8

To demonstrate, consider the numerical example illustrated by Figure 8. The JPD is

constructed from these parameters according to:

$$P(\theta, a, \omega, \pi) \equiv \rho_{\Pi}(\pi|\omega, \theta)\rho_{\Omega}(\omega|a)\rho_{A}(a|\theta)\rho_{\Theta}(\theta). \tag{4}$$

Table 5 reports the resulting JPD and MDK, the latter again identical to the one detailed in Tables 1 and 4). As before, from the MDK, $P(\pi = 1) = 0.6116$, $P(\pi = 1|a = 1) = 0.5100$ and $P(\pi = 1|a = 0) = 0.8400$. Computing the effects of the interventions as before, it can be shown that $P_{\mathbf{do}(a=1)}(\pi = 1) = 0.5400$ and $P_{\mathbf{do}(a=0)}(\pi = 1) = 0.8400$. The intervention is identified for these parameters. Indeed, as demonstrated by the following proposition, it is identified for any set of (positive) stochastic laws.

Proposition 3. Given a collection variables with arbitrary numbers of states parameterized by any positive stochastic laws consistent with Figure 7, the effect of an intervention $\mathbf{do}(a)$ on organizational performance is identifiable for all $a \in A$.

Although these specific examples help to provide some intuition behind the identification problem facing managers, they represent but a fraction of the possible patterns of relationships that might exist between variables that form organizational systems. The next section therefore develops a far more general set of results.

Identifiable interventions in general

The preceding sections revealed that the presence of confounding factors, unknown common causes influencing both the actions of employees and the direct effects of employee actions, prevent managers from being able to accurately assess the effects of those actions on performance. They also demonstrated that the mere presence of hidden influences on employee actions do not always prevent the manager from being able to make accurate assessments. Indeed, whether an unknown common cause actually prevents an intervention from being identifiable depends on the specific structure of the relationships between the unknown variables and the known variables they influence. Here, we provide a general characterization of when these common causes necessarily preclude causal identification.

The risk of presenting simple setups as those in the previous sections is that they encourage approaching the general case with an overly narrow interpretation of the math. For example, the previous cases might be dismissed as unrealistic since, after all, if the lone employee reports directly to the manager, then it is highly unlikely she sits passively on the side collecting data on the correlation between the employee's actions and performance for some extended period before suddenly jumping in and barking out orders. Typically, the employee's immediate supervisor is always engaging the employee, knows his moods and work habits, sees the effects of her management style, and so on. Indeed, she may have been promoted from that very role and, thus, know the hidden factors driving his actions.

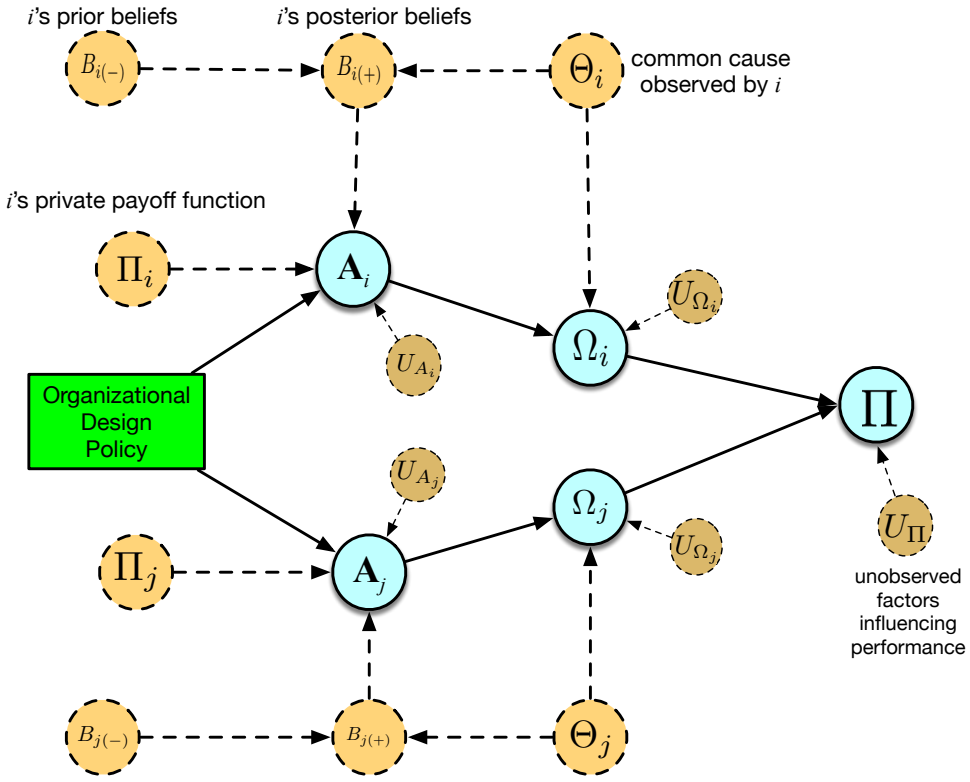

Figure 9: A slightly expanded model.

The situation we have in mind is more in keeping with, e.g., the CEO of a large corporation. Someone in an executive position receives the bulk of her information in form of marketing, financial, technology and other reports – both verbally, in written form and, more recently, in the form of data dashboards accessed on personal computers. The senior manager is several steps removed from those engaged in hands-on production, product de-

sign, distribution, sales, operations, marketing and so on. Figure 9 illustrates a slightly more fleshed-out version of the previous models that may provide a better sense of our intended scenario.

Here, we see two employees, i and j, whose actions influence their intermediate work products which, in turn, jointly affect performance. Included are variables representing each employee's prior beliefs (e.g., $B_{i(-)}$), their posterior beliefs following observation of their private information (e.g., $B_{i(+)}$), and their private payoff functions (e.g., Π_i), all of which influence their behavior (directly or indirectly). In addition, all the known variables are influenced by unknown factors (denoted by the $U_\cdot$'s). The square box indicates a policy variable, say the details of an organizational incentive system that affects both employees. The policy is the decision variable for the senior manager. Policies are typically costly to change and, hence, infrequently altered. In this context, it seems reasonable to imagine the senior manager observing the system and getting a handle on the important correlations in-between policy changes. The question remains: is perfect probabilistic knowledge of the known variables (the best possible case in such a situation) sufficient to make accurate assessments about the effects of employee actions on performance?

Variables: Let $\mathcal{V}$, indexed by $N \equiv \{1, \ldots, n\}$ with typical element V_j, $j \in N$, represent the set of variables in the objective model of an organization. We refer to an element $v_j \in V_j$ as a *state* of V_j. Assume that all variables have a finite number of states. When V_j is in a specific state v_j, we abuse notation and write $V_j = v_j$, referring to v_j as an *instantiation* of V_j; as above, we write an *instantiation of* $\mathcal{V}$ as $v = (v_1, \ldots, v_n) \in \mathcal{V} \equiv \prod_{j=1}^{n} V_j$.

We partition this set of variables into five subsets, based on the type of information that they contain. An organization consists of two or more *members* or employees ($2 \leq m < n$), indexed by $I \equiv \{1, \ldots, m\}$. Each member $i \in I$ may have (i) a set of *beliefs* (B_i) and (ii) a set of *actions* (A_i). As in the examples above, the beliefs capture ways in which information available to the individual might influence that person's choice of action. We denote the classes of action and belief variables as $\mathcal{A} \equiv \{A_1, \ldots, A_m\}$ and $\mathcal{B} \equiv \{B_1, \ldots, B_m\}$, respectively.

The objective model of the organization can also include two types of environmental vari-

ables: *known environmental* factors, denoted $\mathcal{K}$, with typical element K_l, which represent states of the world not directly chosen by any member of the organization but which the manager nevertheless can observe; and *unknown environmental* variables, denoted $\mathcal{U}$, with typical element U_g. In our examples above, the intermediate outcomes would fall in the set of known environmental factors but this set could also include observable factors external to the organization, such as the prices charged by competitors. As in the unknown environmental variables in the simple examples above, the manager cannot observe the unknown environmental variables, though other members may have awareness of them and though they may influence the behavior of the other organizational members.

The set of variables also include a variable, Π (with real number states), that represents *organizational performance*. As in the examples above, we divide these classes of variables into two sets. The set of *known variables* includes the actions of members, the known environmental factors, and organizational performance ($\mathcal{O} \equiv \mathcal{A} \cup \mathcal{K} \cup \Pi$). The set of *unknown variables*, or hidden factors, comprises both the beliefs of employees and the unknown environmental variables ($\mathcal{H} \equiv \mathcal{U} \cup \mathcal{B}$).

Causal structure: The instantiation of each variable depends on a *stochastic process*, a function ρ_j, which includes one or more antecedents, causes of V_j. As with the influence relationships in the examples, these antecedents represent direct causes, hence we label the sets of these variables as $\mathcal{D}_j$.[7] $D_j \equiv \prod_{V_k \in \mathcal{D}_j} V_k$ denotes the set of all instantiations of a subset of variables $\mathcal{D}_j \subset \mathcal{V}$, and $\rho_j(v_j|d_j)$ refers to the probability that $V_j = v_j$ given that $D_j = d_j$.

An *organizational model* consists of the set of variables combined with the set of stochastic processes governing them: $M \equiv (\mathcal{V}, \rho_1, \ldots, \rho_n)$. A model then has a corresponding graph, $G \subset N^2$, where each edge in the graph represents an influence relationship from j to k (i.e. $(j,k) \in G$ if and only if $V_j \in \mathcal{D}_k$).

We place a few restrictions on the influence paths allowed. First, to prevent consequences from becoming their own causes, we do not allow either direct or indirect recursive paths.

[7] The process is "local" in the sense that the stochastic behavior of each V_j depends only on the variables in $\mathcal{D}_j$ and "modular" in the sense that changes in one process, such as through an intervention, do not affect other processes.

Influence relationships can only operate in one direction: if $(j,k) \in G$, then $(k,j) \notin G$. We also assume that G does not include any cycles: if G has a directed path $V_j \to \cdots \to V_k$, then it cannot also include any directed path in which $V_k \to \cdots \to V_j$.

Second, we assume that the actions of any particular member depend only on his or her own beliefs and that a member's beliefs can only influence other variables in the system through the actions of the member holding those beliefs. Members can influence each other's actions but any such influence would need to flow through the effects that their actions, or the consequences of those actions, had on the beliefs of other members.

Finally, we assume that the unknown environmental variables have no antecedents—that is, they do not depend on other variables in the model. To a large extent, this assumption represents a simplification. One could imagine, for example, chains of unknown events or that unknown environmental variables might sit between other sorts of variables. We concatenate such chains. In other words, $V_j \leftarrow U_g \leftarrow \cdots \leftarrow U_h \to \cdots \to U_k \to V_i$ would become $V_j \leftarrow U_h \to V_i$ and $V_j \to U_h \to V_i$ simplifies to $V_j \to V_i$ without loss of generality. In cases where the instantiation of a variable does not depend on any other variables, we write $\rho_j(v_j)$. To avoid division-by-zero complications, we assume that all ρ_j are strictly positive.

Stochastic implications: Every organizational model $M \equiv (\mathcal{V}, \rho_1, \ldots, \rho_n)$ implies a joint probability distribution. We write $P_M(v)$ to indicate the probability of $v \in \mathcal{V}$ implied by M:

$$P_M(v) = \prod_{j=1}^{n} \rho_j(v_j|d_j). \tag{5}$$

One can also factor this probability as:

$$P_M(v) = \prod_{j=1}^{n} P_M(v_j|d_j). \tag{6}$$

When the context clearly implies M, we omit the subscript. Given our question of interest, we frequently focus on the joint distributions of the known variables. Two models M and M' are *observationally equivalent* if $G = G'$ and $P_M(\mathcal{O}) = P(\mathcal{O})_{M'}$—in other words, if they have the same structure of influence relationships and the same joint probability distributions on

the known variables.

Interventions: We assume that the manager can direct any organizational member to do any action and that the member will follow that directive. For each member i, let $A_i^+ \equiv A_i \cup \{\text{idle}\}$ denote the set of interventions available to the manager ("idle" means that the manager gives i no specific direction.

Given M, $v \in V$, and $\hat{a}_i \in A_i^+$, for each ρ_j in M, define the *intervened* process $\rho_{j|\hat{a}_i}$ as:

$$\rho_{j|\hat{a}_i}(v_j|d_j) \equiv \begin{cases} 1 & \text{if } v_j = \hat{a}_i \text{ (implies } \hat{a}_i \neq \text{idle)} \\ \rho_j(v_j|d_j) & \text{if } v_j \in A_i^+ \text{ and } \hat{a}_i = \text{idle} \\ 0 & \text{if } v_j \in A_i^+ \text{ and neither } v_j = \hat{a}_i \text{ nor } \hat{a}_i = \text{idle} \\ \rho_j(v_j|d_{j|\hat{a}_i}) & \text{if } v_j \notin A_i^+ \end{cases}, \tag{7}$$

where $d_{j|\hat{a}_i}$ represents d_j with any component corresponding to A_i replaced by $\hat{a}_i$. This definition simply indicates that an intervention alters the stochastic process so that $\hat{a}_i \neq$ "idle" means that $a_i = \hat{a}_i$ occurs with probability one. Actions left idle continue to behave as specified by M. We write the distribution induced by an intervention as:

$$P_{M|\hat{a}_i}(v) = \prod_{j=1}^{n} \rho_{j|\hat{a}_i}(v_j|d_j). \tag{8}$$

When the context clearly implies M, we write $P_{\hat{a}_i}$ and the joint distribution over the known variables as $P_{\hat{a}_i}(\mathcal{O})$.

Causal identification: The manager can obtain information on the distributions of the known variables, $P_M(\mathcal{O})$, but cannot observe the influence relationships themselves nor any of the unknown variables. Does $P_M(\mathcal{O})$ provides sufficient information for the manager to assess whether a particular intervention would improve organizational performance? The following definition is adapted from Pearl (2009).

Definition 1 (Identifiability)**.** Given an organizational model M and some $i \in I$, the intervention $\hat{a}_i \in A_i^+$ is *identifiable* if $P_{M|\hat{a}_i}(\mathcal{O}) = P_{M'|\hat{a}_i}(\mathcal{O})$ for all alternative models M'

that satisfy: (i) M and M' imply the same graph of influence relationships G; and (ii) $P_M(\mathcal{O}) = P_{M'}(\mathcal{O})$.

In other words, given a true description of reality M, an intervention is identifiable if any alternative parameterization of the qualitative pattern of influence relationships, G, that happens to result in the same marginal distribution on known variables, $P_M(\mathcal{O})$, also results in the same distribution on known variables post-intervention, $P_{M|\hat{a}_i}(\mathcal{O})$. Thus, the focus is upon what patterns of influence relationships, the G part, are such that equivalence in the MDK implies equivalence in the intervention effect. When this is true, the MDK provides sufficient information to compute the effect of the intervention.

Given this definition, we can prove the following result by removing the effects of the unknown beliefs through the law of total probability and by then applying Tian and Pearl (2002) to our context (see the appendix for the proof). A *path* exists between two variables when a set of influence relationships connects them (ignoring the directionality of those relationships), potentially through multiple intervening variables. Let us define a *bi-directed path* as a path which includes an unknown environmental variable between two other variables (e.g., $B_i \leftarrow U_h \rightarrow K_l$). Note that because unknown environmental variables are themselves roots, the arrows coming out of the unknown environmental variables in these paths will necessarily point in opposite directions.

Proposition 4. Given an organizational model M, an intervention $\hat{a}_i$ is identifiable if and only if there is no bi-directed path connecting A_i to any V_j such that $A_i \rightarrow V_j$.

When applying this proposition, one must consider all of the potential paths between two variables. For example, in Figure 4, two paths exist between the actions of the employee and the intermediate outcomes. One is the direct path: $A \rightarrow \Omega$. The other is an indirect path: $A \leftarrow B \leftarrow \Theta \rightarrow \Omega$. Because an unknown environmental variable (Θ) resides on this indirect path, it is bi-directed. Consistent with the proposition and as demonstrated above, the effect of an intervention in this example is therefore not identifiable.

Given the wide range of possible organizational models, this result appears somewhat surprising on two dimensions. First, identifiability depends only upon the structure of influence relations, as summarized by G. The specific parametric quantification of the model does

not matter. Second, identification only fails when a confounding common cause exists with respect to a employee's action and one of its immediate consequences. Any other pattern of influence relationships allows identification.

Unfortunately, one can easily imagine examples that would become problematic. Any situation in which an employee receives a signal – about a buyer or a supplier or about some other external factor – and that signal both influences the beliefs of the employee and the consequences of the employee's actions would create a bi-directed path. In any real-world setting, these cases seem rampant.

CULTURE AS A SOLUTION

Our notion of culture here refers to a group having a set of shared myths, rituals, routines, artefacts, and vocabulary, and perhaps a set of common assumptions. Many of those studying culture at a societal level have been skeptical of whether such shared elements shape action, instead seeing them more as a means of organizing experiences and justifying behaviors (Swidler, 1986; Patterson, 2014). But organizational culture has usually been seen as functional—as enhancing performance by allowing for more effective coordination among organization members (e.g., Gordon and DiTomaso, 1992b; Kotter and Heskett, 1992; Burt et al., 1994; Sorensen, 2002).

Two mechanisms, in particular, have received substantial attention as the probable sources of these performance advantages: shared values and more efficient communication. Organizations with strong cultures – meaning those with higher levels of concensus on these shared elements – have been seen as better able to coordinate because they develop shared categorization schema and specialized vocabularies (Srivastava et al., forthcoming). Natural languages often prove cumbersome, even misleading, within specific contexts. Words have multiple meanings and errors in the transmission of a message can easily occur when the sender and receiver have different meanings in mind. Although jargon has a pejorative connotation, within a particular community, specialized language and meanings can enable more efficient communication. Weber (2003)**?**, for example, in an experimental setting, found

that small groups naturally develop specialized vocabularies, which allowed these groups to increase their productivity by 500%, on average, over the rounds of the experiment.

But shared language cannot solve the managerial inference problem. Shared language can prevent problems of miscommunication and by doing so facilitate coordination. But it does not change the structure of the organizational system. From the perspective of our model, shared language simply eliminates frictions that might exist in conveying information or in the implementation of a particular managerial directive. In the best-case scenarios examined above, we assumed away these issues and yet inference could still prove problematic.

Organizations with strong cultures have also been seen as having members with relatively homogeneous goals and values (Kotter and Heskett, 1992; O'Reilly and Chatman, 1996), whether through the selective recruitment or the selective retention of members more closely aligned with the values of the organization or through members adopting these beliefs through a process of social influence. Having values more closely aligned with those of the other members of the organization may improve performance by increasing the commitment of members to the organization (O'Reilly and Chatman, 1996). Shared values may also allow for more efficient coordination if peer pressure rather than managerial oversight can ensure that employees follow organizational rules and routines (Gordon and DiTomaso, 1992a; Kreps, 1990).

But again shared values cannot solve the inference problem. Values remain fundamentally unobservable. How can one determine what another truly believes? Moreover, even if one assumed that individuals could and would accurately relay their beliefs to others, values by themselves rarely imply a specific course of action (Swidler, 1986). Individuals may even hold multiple values pointing them in disparate directions, meaning that they must choose between them. In the context of our model, these values fall in the category of employee beliefs, factors that influence employee actions but which the manager cannot observe. Rather than solving causal inference problems, they are the source of them.

How then might culture solve the managerial inference problem? Note that the problematic cases have a very specific form: they stem from situations in which an unknown environmental variable influences both the beliefs of an employee and the direct consequences

of that employee's actions. The key to solving this problem resides in replacing these hidden beliefs as antecedents of employee behavior with observable factors. Interestingly, two dimensions of corporate culture do just that: i) joint commitment to collective beliefs, and ii) organizational routines.

Joint commitment

Our notion of joint commitment to collective beliefs draws on recent work in the philosophy of social ontology (in particular, Gilbert, 2014)[8]. As noted above, the problem with beliefs stems from the fact that they remain fundamentally unobservable. How then can organization members commit effectively to any particular belief or set of beliefs? This recent stream of philosophy argues that groups do so by committing jointly to *emulate* a single entity, a representative agent, who operates according to the shared beliefs.

Note that members of the group need not actually hold the shared beliefs – or even the same beliefs – according to this perspective. That eliminates the potential difficulties associated with expecting people to articulate their values or to adopt a set of beliefs. They need not even have awareness of the actual values held by the representative agent, the entity believed to operate according to the shared beliefs. But they do need to agree on the representative agent and to sublimate their own beliefs and interests to following the actions of the representative agent.

But who is the representative agent? Obviously, few groups of people discuss explicitly the idea that they should follow the actions of a specific member. But many groups come to see one member as embodying the group goals, through the person's actions and through the opinions voiced. That person might not act as the formal leader for the group but they may become its moral compass.

Over time, the representative agents may become ritualized. Many organizations, particularly those that have been seen as exemplars of strong culture, appear to have someone perceived as holding the "true" values of the organization. These representative agents often

[8]This area of philosophy has been developing rapidly. Recent book-length treatments on social ontology and group agency would include: List and Pettit (2011); Miller (2010); Gilbert (2014); Tuomela (2013); Bratman (2014); Pratten (2014); Tollefsen (2015)

appear in the form of a charismatic founder or early leader. Consider Sam Walton in the case of Wal-Mart, General Johnson in Johnson & Johnson, or Steve Jobs in Apple. Those affiliated with the organization see these early leaders as embodying the values of their organizations. Parables of how they responded in particular situations become seen as guides for future action.

Gilbert (2014) also argues that joint commitment requires each member to express their willingness to commit to this emulation and to do so publicly. Here, the philosophy of social ontology appears to have reached a similar conclusion to cultural sociologists. (Patterson, 2014, p. 8) notable states that "if it is not shared and public it is not cultural." Everyone in the group must understand that everyone else has committed to the shared belief (Gilbert, 2014; Patterson, 2014).

How might organization members express their assent to the joint commitment? Such expression need not occur explicitly, in the sense that organization members declare their intention to join the collective. But it must occur through observable acts. Here, the shared rituals and shared language of strong culture organizations may act as a signal of assention to the joint commitment. Outsiders often feel as though these rituals give organizations with strong cultures a cult-like feel, rightly so. Much as cults rely on these small acts to signal commitment to the group, strong culture organizations. Strong culture organizations. Those who do not modify their own behaviors and language toward those of the group do not fit and eventually leave, either voluntarity or involuntarily (Srivastava et al., forthcoming).

Although shared language and shared values cannot solve the inference problem, joint commitment to a shared belief can. It does so by replacing the effects of unknown individual-level beliefs with those of beliefs known by all, including the manager.

To capture this intuition in our formal model, we introduce a new variable into the model, B^*, the *collective belief* of the organization. This variable has its own influence on other variables in the organizational sytem, represented by ρ_{B^*}. If an employee commits to the collective belief, then it supplants the influence of their own individual beliefs (formally, $B^* \in \mathcal{D}_k$ and $B_k \notin \mathcal{D}_k$). Note that if the organization includes more than one member, then all members must assent to the collective belief (i.e. for all $i \in I$, $B^* \rightarrow A_i$ and $B_i \nrightarrow A_i$). If

the organizational model includes such a collective belief, we say that the collective belief is *operative*.

Although all members of the organization must commit to the collective belief and this belief itself remains constant across all members, the implications of the belief can depend on a wide variety of factors. Returning to the example above, the collective belief could include a belief about how a salesperson should respond to a customer with a particular facial expression. But such a collective belief would presumably arise from seeing how the representative agent – the individual presumed to hold the beliefs of the organization – responded in a similar circumstance. That fact, however, does impose some limitations on the sorts of beliefs that could become collective. Whereas individual employees might have individual-level beliefs about specific customers or suppliers, collective beliefs will usually only pertain to classes of objects, such as a type of customer, and of situations. Otherwise, the representative agent would need to do everything and to interact with everyone.

Crucially, the need for public commitment implies that every member of the organization must have awareness of the collective belief, *including the manager*. Figures 10 and 11 provide some intuition as to how the introduction of a collective beliefs alters the nature of the influence relationships. The first figure illustrates a number of possible relationship structures. Note that panels (b), (c), and (d) all include bi-directed paths between the actions of the employee and an outcome of those actions (e.g., in panel (b), $A_i \leftarrow B_i \leftarrow U_k \rightarrow K_g$). By Proposition 4 these cases therefore are not identifiable, the manager cannot infer the consequences of an intervention.

But consider how these examples change with the introduction of a collective belief in Figure 11. In all of the cases, the actions of the employees now depend not on the individual beliefs of those employees but on the collective belief of the organization. How does this solve the inference problem? Consider panel (b). Although the collective belief itself depends on an environmental variable that, by definition, the manager cannot observe. The states of those variables influence the collective belief, which we assume that the manager holds (just as every other member of the organization).

Proposition 5. If M is an organizational model in which a collective belief is operative, all

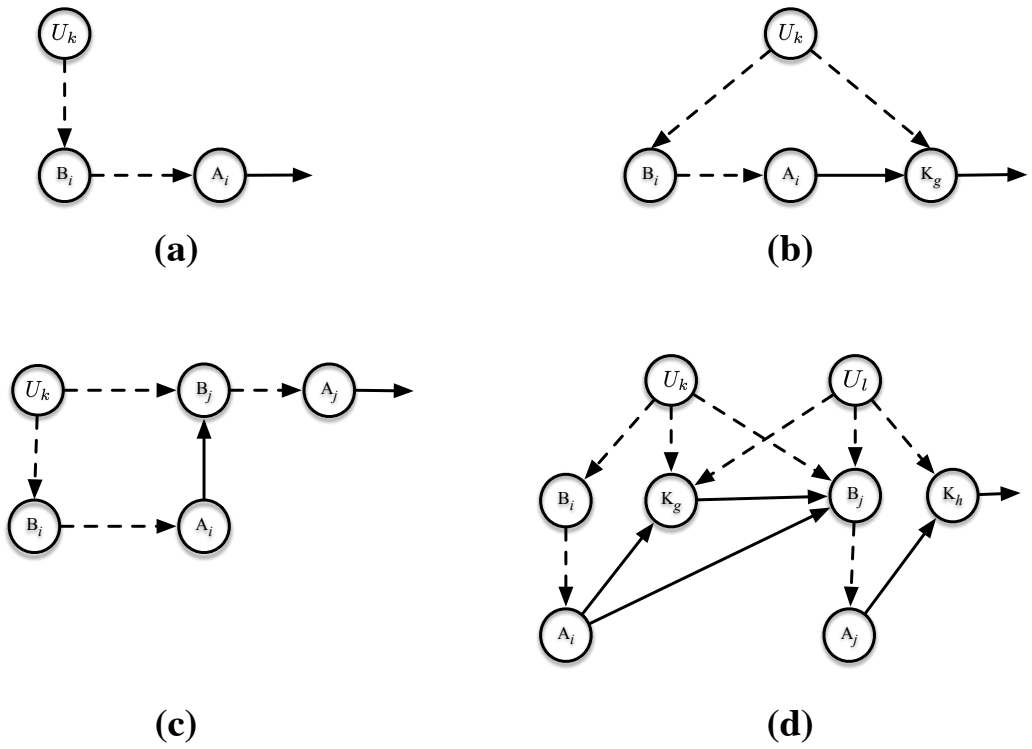

Figure 10: Some possible influence structures.

interventions are identifiable.

We would note, however, that collective beliefs also impose a constraint. Although managers can infer the consequences of potential interventions, they may not feel free to enact them. Unless the manager happens to serve as the representative agent, an intervention might run counter to the actions implied by the collective belief. Indeed, the representation agent might even oppose the intervention on the grounds that it does not fit with the values of the organization. Such conflict could result in the end of the collective belief—by imposing another course of action, the manager has signaled that she no longer agrees to the joint commitment. Or, it might lead to the exit of the manager (or leader).

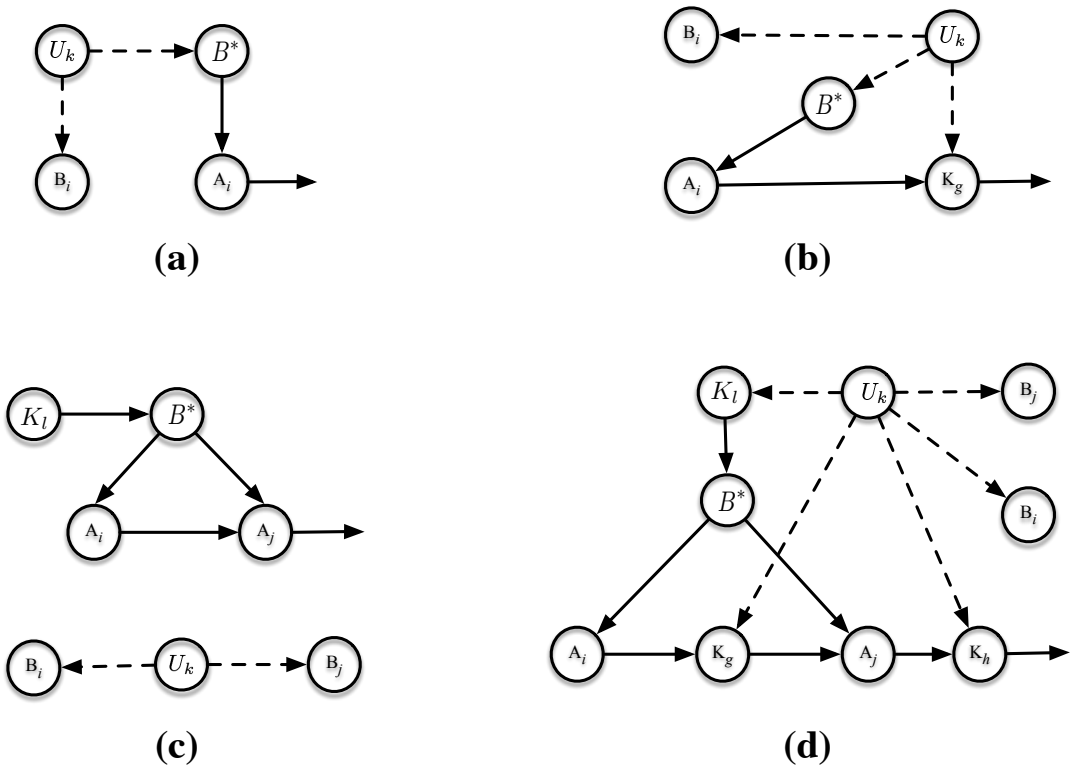

Figure 11: Influence structures modified by the introduction of a collective belief.

Organizational routines

Organizational routines represent another common feature of organizations. Although routines have been defined in many ways, we have in mind a set of "if-then" rules that systematization the learning and tacit knowledge accumulated by the organization (March et al., 1958). They need not be written down anywhere, though many organizations do have handbooks and other documents that record some of this information. We nevertheless expect that even in the most highly-codified organizations, that the majority of the day-to-day routines that govern operations and interactions get passed from person to person either through a verbal exchange or through watching and mimicry.

Routines exist even in organizations that most would not consider to have strong cultures. As Grant (1996, p. 379) puts it:

> Observation of any work team, whether it is a surgical team in a hospital operating room or a team of mechanics at a grand prix motor race, reveals closely-

> coordinated working arrangements where each team member applies his or her specialist knowledge, but where the patterns of interaction appear automatic. This coordination relies heavily upon informal procedures in the form of commonly-understood roles and interactions established through training and constant repetition, supported by a series of explicit and implicit signals.

But organizations with stronger cultures probably both have a more extensive set of routines and, importantly, a more homogenous understanding of these routines among the various members (Schein, 2010).

Although they have a similar effect in terms of replacing private individual-level beliefs with observable factors, with respect to our model, routines differ in two important respects from collective beliefs: (i) routines may be local – that is, they can exist and operate at the level of an individual or of a subunit of an organization (e.g., teams or departments); and (ii) they do not require fulfillment of the joint commitment condition associated with collective intentions. Consider, for example, how routines would operate in Figure 1. The if-then rules, understood by all, effectively replaces the influence of individual-level beliefs on employee actions. Figure 12 illustrates how the introduction of this set of procedures changes the organizational model.

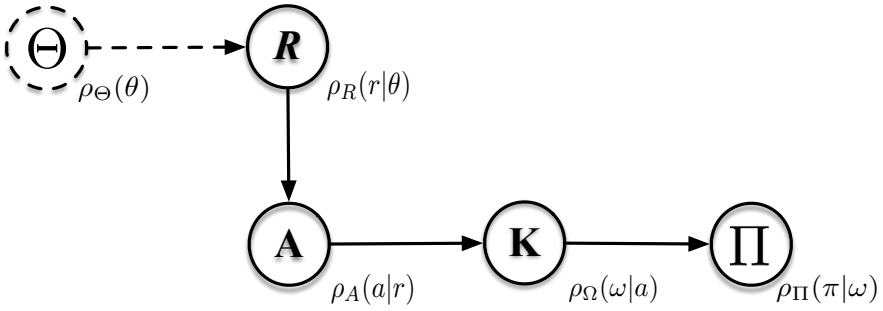

Figure 12: Routine replaces the tacit knowledge of Figure 1

Translating this intuition into the formal model involves replacing $B_i \dashrightarrow A_i$ in some model M with $R_i \rightarrow A_i$ in M'. When such a replacement has been made, we that *the tacit knowledge of i is codified in the model M'*. Importantly, the manager can observe which routine has been activated even when she cannot observe the (potentially hidden) factors that triggered it.

As a relevant aside, suppose $\mathcal{D}_k$ is the set of direct causes of B_i in M. Then, a case of *perfect* codification is one in which: (i) the variables and graphs associated with M and M' are identical with the exception of R_i being substituted for B_i in M'; and (ii) the stochastic processes ρ'_{R_i} and ρ'_{A_i} are such that $P(a_i|d_k) = P'(a_i|d_k)$. In other words, the stochastic effect of hidden information on A_i is the same under both models. We do *not* require perfect codification for identifiability, as demonstrated in the following proposition.

Proposition 6. Suppose that $\hat{a}_i$ is not identifiable for some $i \in I$ in organizational model M. If M' is an alternative model in which the tacit knowledge of i is codified, then $\hat{a}_i$ is identifiable.

Discussion

The preceding analysis examines causal identification of the effects of employee behavior in the presence of unknown influences. In an organization in which a member's behavior is driven by his beliefs regarding the consequences of his actions which, in turn, are shaped by private information about hidden environmental variables, the effects of managerial interventions may not be identifiable. When this is the case, the effects managerial interventions on organizational behavior cannot be estimated. The manager's "unknown unknowns" confound her ability to assess the implications of organizational policies. We provide a general characterization of the patterns of influence relations between variables, both hidden and known, that lead to a failure of identifiability.

We also discover that an interesting, and previously overlooked, feature of two components of organizational culture – collective beliefs and routines – is the ability to resolve the identification problem. Thus, culture, may provide the manager with the capacity to understand the implications of her available interventions. Because routines may operate at the local, individual level, they may permit more refined behavior than collective beliefs at the global, organizational level. On the other hand, collective beliefs (in the sense examined here) may be more amenable to managerial manipulation. Deeper examination of these comparisons and trade-offs strikes us as a fruitful avenue of future inquiry.

Interesting though the culture angel may be, it is worth imagining other ways to identify the effects of a employee's actions on organizational behavior. As suggested in the Introduction, one possibility is to run a randomized experiment on feasible interventions in order to quantify the true effects of those interventions. Typically, however, the costs of such endeavors – both direct and opportunity-wise – are prohibitive. Alternatively, promoting managers from within the organization might solve the problem since those promoted have the experience to know the effects of the actions of someone in their previous role (at least, on the immediate consequences of those actions if not on organizational variables further downstream). This is surely one of the key reasons to promote from within. That said, organizational members sometimes quit and, occasionally, managers are hired from outside. In these cases, identifiability remains a problem. Therefore, even in organizations dominated by internally developed of managerial talent, one cannot take the premise of identification for granted.

An obvious question is, if we are comfortable assuming that employees are willing to operate in good faith according to the collective belief, why not simply *ask* them about the consequences of their actions directly? In many cases, this may be reasonable – provided the employee has a sufficient grasp of his local situation to provide a precise description of it. However, there is nothing in our setup that requires this level of self-awareness or communication skill: employees may well be influenced by hidden factors in ways that are not obvious or apparent to them. The beauty of identification, when it is possible, is that the motivations, knowledge, expressiveness, honesty, etc., of the employee are not relevant to the problem because they play no role in isolating the causal consequences of their actions. Once the identification step is complete, however, the implementation of mechanisms designed to elicit the desired behavior almost certainly requires a sense of these issues.

If, through experience in his job, individual i knows not only that people in that role observe U_k, but also its distribution and its relationship to W_g through U_l, then appending that knowledge to what the manager knows (e.g., by promoting i into management), would serve to "reveal" the hidden process. This is shown in Figure **??**, in which previously hidden variables are transformed into known, non-action variables. Again, it is not critical that the

manager observe these variables per se – what matters is that she is aware of them, their structure and their stochastic behavior, all of which could be added to her of knowledge via promotion of i into management.

Summing up, this paper makes two novel contributions to organizational theory. The first is a full characterization (an "if and only if" proposition) of the structure of organizational influence relations required to identify the effects of employee behavior on performance in the presence of hidden variables. Our main finding relies upon an adaptation of Tian and Pearl (2002), which we extend to accommodate the special nature of beliefs.[9] To the best of our knowledge, we are the first to apply this form of identification analysis in the context of organizational management.

The second is our discovery that certain features of organizational culture have the potential to ameliorate problematic influence structures within organizations. This finding identifies a previously unappreciated aspect of organizational culture: that, in addition to its traditional conception as a determinant of organizational behavior (and, hence, performance), culture may also serve as a vehicle by which the consequences of managerial interventions come to be understood. Our hope is that this novel theoretical claim is sufficiently intriguing so as to motivate additional research.

[9]Specifically, beliefs are represented as unknown variables that may have known parents in the causal graph. Tian and Pearl (2002) assume all hidden variables are root nodes. Such models are called *semi-Markovian*. Even though our model relaxes this restriction, the structure of beliefs are such that the extension is straightforward.

APPENDICES

In the following proofs, we adopt the notational convention of using $\hat{a}$ to represent the intervention "$\mathbf{do}(a = \hat{a})$."

A Proof of Proposition 1

Intuitively, one can factor P in accordance with (1), as:

$$P(\theta, a, \omega, \pi) = P(\pi|\omega)P(\omega|a)P(a|\theta)P(\theta). \tag{9}$$

Equation (9) implies that certain conditional independencies exist in the joint distribution implied under any parameterization of the primitive stochastic processes that interact to generate it. For example, by the Chain Rule, one can factor *any* joint probability distribution P on $\mathcal{V}$ as:

$$P(\theta, a, \omega, \pi) = P(\pi|\omega, a, \theta)P(\omega|a, \theta)P(a|\theta)P(\theta). \tag{10}$$

Together equations (1) and (10) imply that $P(\omega|a, \theta) = P(\omega|a)$; in other words, they imply that, conditional on a, ω does not depend in any way on either θ or b. This conclusion is essential for our first finding. Its generalization, the well-known "d-separation" result of Verma and Pearl (1988), will appear throughout our analysis.

Define the indicator function $1_{\hat{a}} : A \to \{0, 1\}$ such that $1_{\hat{a}}(a) = 1$ if $\hat{a} = a$ and zero otherwise. An intervention $\hat{a}$ results in a new joint probability distribution on $\mathcal{V}$, denoted $P_{\hat{a}}$, and constructed as in (1) but with $1_{\hat{a}}$ substituted for $\rho_A(a|\theta)$:

$$P_{\hat{a}}(\theta, a, \omega, \pi) \equiv \rho_\Pi(\pi|\omega)\rho_\Omega(\omega|a)1_{\hat{a}}(a)\rho_\Theta(\theta), \tag{11}$$

which is nonzero only if $a = \hat{a}$. This implies that (11) can be factored as

$$P_{\hat{a}}(\theta, a, \omega, \pi) = P_{\hat{a}}(\pi|\omega)P_{\hat{a}}(\omega|a)P_{\hat{a}}(a)P_{\hat{a}}(\theta). \tag{12}$$

Suppose the manager makes some intervention $\hat{a} \in A$. From the equivalencies between (1) and (12), it follows that

$$P_{\hat{a}}(\theta, a, \omega, \pi) = \begin{cases} P(\pi|\omega)P(\omega|\hat{a})P(\theta) & \text{if } a = \hat{a}, \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, by the law of total probability (see, e.g., Pearl, 2009, Ch. 1, loc. 468, Kindle version),

$$\begin{aligned} P_{\hat{a}}(\omega, \pi) &= \sum_{\theta \in \Theta} P(\pi|\omega)P(\omega|\hat{a})P(\theta), \\ &= P(\omega|\hat{a})P(\pi|\omega) \sum_{\theta \in \Theta} P(\theta), \\ &= P(\omega|\hat{a})P(\pi|\omega). \end{aligned} \tag{13}$$

Since one can compute the quantities in (13) from $P(\mathcal{A}, ,)$, the effect of the invention is identifiable. Given that the distributions and the intervention had been chosen artibitrarily, moreover, this conclusion will hold for any potential intervention and for any pair P and P' such that $P(\mathcal{A}, ,) = P'(\mathcal{A}, ,)$.

We can say more. By the Chain Rule, $P(\mathcal{A}, ,)$ can always be factored as

$$P(a, \omega, \pi) = P(a)P(\omega|a)P(\pi|\omega, a). \tag{14}$$

By Bayes' Rule,

$$P(\omega, \pi|a) = \frac{P(a, \omega, \pi)}{P(a)}, \tag{15}$$

where, $P(a) > 0$ by our assumption of strictly positive stochastic processes. Combining (14) and (15),

$$P(\omega, \pi|a) = P(\omega|a)P(\pi|\omega, a).$$

As noted earlier, π is conditionally independent of a given ω. Therefore,

$$P(\omega, \pi|a) = P(\omega|a)P(\pi|\omega). \tag{16}$$

Thus, $P_{\hat{a}}(\omega, \pi) = P(\omega, \pi|\hat{a})$. That is, the effect of an intervention $\hat{a}$ is simply the conditional probability of (ω, π) given $\hat{a}$ calculated from $P(A, \Omega, \Pi)$.

B Proof of Proposition 2

To begin, we now have

$$\begin{aligned} P(a, \omega, \pi) &= \sum_{\theta \epsilon \Theta} \rho_\Pi(\pi|\omega)\rho_\Omega(\omega|a, \theta)\rho_A(a\theta)\rho_\Theta(\theta) \\ &= \rho_\Pi(\pi|\omega) \sum_{\theta \epsilon \Theta} \rho_\Omega(\omega|a, \theta)\rho_A(a|\theta)\rho_\Theta(\theta). \end{aligned} \tag{17}$$

Proceeding similarly for the intervention $\hat{a}$,

$$P_{\hat{a}}(\theta, a, \omega, \pi) = \begin{cases} \rho_\Pi(\pi|\omega)\rho_\Omega(\omega|\hat{a}, \theta)\rho_\Theta(\theta) & \text{if } a = \hat{a}, \\ 0 & \text{otherwise} \end{cases}.$$

This implies

$$\begin{aligned} P_{\hat{a}}(\omega, \pi) &= \sum_{\theta \epsilon \Theta} \rho_\Pi(\pi|\omega)\rho_\Omega(\omega|\hat{a}, \theta)\rho_\Theta(\theta), \\ &= \rho_\Pi(\pi|\omega) \sum_{\theta \epsilon \Theta} \rho_\Omega(\omega|\hat{a}, \theta)\rho_\Theta(\theta), \\ &= P(\pi|\omega)P_{\hat{a}}(\omega). \end{aligned} \tag{18}$$

(THE FOLLOWING MUST BE SIMPLIFIED FOR THE REMOVAL OF B FROM THE EXAMPLE.) Now consider an alternative parameterization of the autonomous processes, distinguished with a prime (i.e., ρ vs. ρ'). Set $\rho'_\Pi = \rho_\Pi$ and $\rho'_\Theta = \rho_\Theta$. Next, pick an arbitrary distribution $\chi \in \Delta^+(B)$ such that, for all $b \in B$, $\chi(b) \neq P(b)$. Then, construct a process ρ'_B

such that

$$\forall b \in B, \theta \in \Theta, \chi(b) = \sum_{\theta \in \Theta} \rho'_B(b|\theta)\rho_\Theta(\theta), \text{ and} \tag{19}$$

$$\forall b' \neq b, \chi(b) \neq \rho'_B(b'|\theta). \tag{20}$$

Equation (19) has $|\Theta| \geq 2$ unknowns and, hence, is underdetermined. This guarantees existence of a solution satisfying (20). By construction, the left-hand side of (19) is equal to $P'(b)$ with the implication that, for all $b \in B$, $P'(b) \neq P(b)$.

Choose a $\theta \in \Theta$ for which $P'(b) \neq \rho'_B(b|\theta)$.[10] Next, select any one from the infinite number of $\chi'(\cdot|\theta) \in \Delta^+(A)$ satisfying, for all $a \in A$, $P(a) \neq \chi'(a|\theta) \neq P(a|\theta)$. Finally, solve for ρ'_A such that

$$\forall a \in A, \sum_{b \in B} \rho'_A(a|b)P'(b) = P(a), \text{ and} \tag{21}$$

$$\sum_{b \in B} \rho'_A(a|b)\rho'_B(b|\theta) = \chi'(a|\theta). \tag{22}$$

This is a system of two equations with $|B| \geq 2$ unknowns. Noting that by (20), for all $b, b' \in B$, $P'(b') \neq \rho'_B(b|\theta)$, this is a consistent linear system and, hence, has at least one solution. The left-hand side of (22) equals $P'(a|\theta)$, which ensures that $P'(a|\theta) \neq P(a|\theta)$ for all $a \in A$ and at least one $\theta \in \Theta$ (there may be more). By (**??**), we also have: (i) for each $a \in A$, the left-hand side of (21) is equal to $P'(a)$, which implies $P'(a) = P(a)$; and (ii) since, for all $b \in B$, $P'(b) \neq P(b)$, for each $a \in A$ there exists a $b \in B$ such that, $\rho'_A(a|b) \neq \rho_A(a|b)$.

Next, choose ρ'_Ω such that

$$\forall a \in A, \omega \in \Omega, \sum_{\theta \in \Theta} \sum_{b \in B} \rho'_\Omega(\omega|a, \theta)\rho'_A(a|b)\rho'_B(b|\theta)\rho_\Theta(\theta) = P(\omega, a), \text{ subject to} \tag{23}$$

$$\sum_{\theta \in \Theta} \rho'_\Omega(\omega|a, \theta)\rho_\Theta(\theta) = \chi(\omega|a) + \sum_{\theta \in \Theta} \rho_\Omega(\omega|a, \theta)\rho_\Theta(\theta). \tag{24}$$

[10] There is at least one such θ by $\Theta \to B$ (influence relations are nontrivial).

where $\chi(\cdot|a) \in \Delta^+(\Omega)$. Note that (23) reduces to

$$\sum_{\theta \in \Theta} \rho'_\Omega(\omega|a,\theta) P'(a|\theta) P(\theta) = P(\omega, a). \tag{25}$$

For each $a \in A$, this is a consistent system of $2 \times |\Omega|$ equations in $(|\Theta| + 1) \times |\Omega|$ unknowns. From (25) and the fact that $P'(a) = P(a)$, we conclude $P'(\omega|a) = P(\omega|a)$.

With (23), all the stochastic processes are defined for the alternative model, thereby completing P' via (**??**). Since $\rho'_\Pi(\pi|\omega) = \rho_\Pi(\pi|\omega)$, $P'(\pi|\omega) = P(\pi|\omega)$, for all $(a, \omega, \pi) \in O$,

$$\begin{aligned} P(a,\omega,\pi) &= P(\pi|\omega)P(\omega|a)P(a), \\ &= P'(\pi|\omega)P'(\omega|a)P'(a), \\ &= P'(a,\omega,\pi). \end{aligned}$$

Thus, the requisite equality of the marginal distributions on organizational variables is satisfied.

Identifiability of $\hat{a}$ fails if, for some $\omega \in \Omega, \pi \in \Pi$, $P_{\hat{a}}(\pi,\omega) \neq P'_{\hat{a}}(\pi,\omega)$. The construction in (18) implies $P_{\hat{a}}(\pi,\omega) = P(\pi|\omega)P_{\hat{a}}(\omega)$, where

$$P_{\hat{a}}(\omega) = \sum_{\theta \in \Theta} \rho_\Omega(\omega|\hat{a},\theta)\rho_\Theta(\theta) \tag{26}$$

Then, by condition (24), for all $a \in A$, $\omega \in \Omega$, $P_{\hat{a}}(\omega) - P_{\hat{a}}(\omega) = \chi(\omega|\hat{a}) > 0$. Therefore, for all $\pi \in \Pi$,

$$P'_{\hat{a}}(\pi,\omega) - P_{\hat{a}}(\pi,\omega) = \frac{\chi(\omega|\hat{a})}{P(\pi|\omega)} > 0. \tag{27}$$

As a result, $\hat{a}$ is not identifiable.

C Proof of Proposition 3

In this system, $P(\mathcal{V})$ becomes:

$$P(\theta, a, \omega, \pi) = \rho_{\Pi}(\pi|\omega, \theta)\rho_{\Omega}(\omega|a)\rho_{A}(a|\theta)\rho_{\Theta}(\theta). \tag{28}$$

Given an intervention $\hat{a}$,

$$P_{\hat{a}}(\theta, a, \omega, \pi) = \begin{cases} \rho_{\Pi}(\pi|\omega, \theta)\rho_{\Omega}(\omega|\hat{a})\rho_{\Theta}(\theta) & \text{if } a = \hat{a}, \\ 0 & \text{otherwise} \end{cases}. \tag{29}$$

Consider an arbitrary parameterization of the organizational model. From (29),

$$\begin{aligned} P_{\hat{a}}(\omega, \pi) &= \sum_{\theta \in \Theta} P(\pi|\omega, \theta)P(\omega|\hat{a})P(\theta), \\ &= P(\omega|\hat{a}) \sum_{\theta \in \Theta} P(\pi|\omega, \theta)P(\theta), \\ &= P(\omega|\hat{a}) \sum_{a \in A} \sum_{\theta \in \Theta} P(\pi|\omega, \theta)P(\theta|a)P(a). \end{aligned} \tag{30}$$

From the conditional independencies inherent in this organizational structure (see Appendix), the following equivalencies are true:

$$P(\theta|a) = P(\theta|a, \omega), \text{ and} \tag{31}$$

$$P(\pi|\omega, \theta) = P(\pi|\theta, \omega, a). \tag{32}$$

Substituting (31) and (32) into (30) yeilds

$$\begin{aligned} P_{\hat{a}}(\omega, \pi) &= P(\omega|\hat{a}) \sum_{a \in A} P(a) \sum_{\theta \in \Theta} P(\pi|\theta, \omega, a)P(\theta|a, \omega), \\ &= P(\omega|\hat{a}) \sum_{a \in A} P(\pi|\omega, a)P(a). \end{aligned} \tag{33}$$

Noting that the quantities on the right hand side of (33) are all computable from $P(\mathcal{O})$ leads to the following proposition.

D Proof of Proposition 4

D.1 Lemma

Begin with an organizational model $M \equiv (\mathcal{V}, \rho_1, \ldots, \rho_n)$. Consider an arbitrary agent $i \in I$ and construct a model $M' \equiv (\mathcal{V}', \rho'_1, \ldots, \rho'_{n-1})$ which is identical to M with the following exceptions. First, remove the belief variables:

$$\mathcal{V}' = \mathcal{V} \backslash \bigcup_{i \in I} B_i.$$

Next, for all $i \in I$, set $\mathcal{D}'_{a_i} \equiv \mathcal{D}_{b_i}$ and define

$$\rho'_{A_i}(a_i | d_{a_i}) = \sum_{b_i \in B_i} \rho_{A_i}(a_i | b_i) \rho_{B_i}(b_i | d_{B_i}). \tag{34}$$

Then, $P_{M|\hat{a}}(\mathcal{O}) = P_{M'|\hat{a}}(\mathcal{O})$.

Proof. Recall, our notational convention that $b \in \mathcal{B}$ is used to mean $b \in \mathsf{X}_{i \in I} B_i$, etc., so that we write $P(v) = P(\pi, k, a, b, u)$, $P(o) = P(\pi, k, a)$, and so on.

First,

$$\begin{aligned}
P_M(\pi, k, a, u) &= \sum_{b \in \mathcal{B}} P_M(\pi, k, a, b, u) \\
&= \sum_{b \in \mathcal{B}} \left[P_M(\pi | d_\pi) P_M(u) \prod_{i \in I} P_M(a_i | b_i) P_M(b_i | d_{B_i}) \prod_{K_j \in \mathcal{K}} P_M(k_j | d_{K_j}) \right], \\
&= P_M(\pi | d_\pi) P_M(u) \prod_{K_j \in \mathcal{K}} P_M(k_j | d_{K_j}) \sum_{b \in \mathcal{B}} \prod_{i \in I} P_M(a_i | b_i) P_M(b_i | d_{B_i}) \\
&= P_M(\pi | d_\pi) P_M(u) \prod_{K_j \in \mathcal{K}} P_M(k_j | d_{K_j}) \prod_{i \in I} \sum_{b_i \in B_i} P_M(a_i | b_i) P_M(b_i | d_{B_i}) \qquad (35)
\end{aligned}$$

The last two steps are because beliefs only influence own actions. Turning to M',

$$
\begin{aligned}
P_{M'}(\pi, k, a, u) &= P_{M'}(\pi|d_\pi) P_{M'}(u) \prod_{K_j \in \mathcal{K}} P_{M'}(k_j|d_{K_j}) \prod_{i \in I} P_{M'}(a_i|d_{A_i}) \\
&= P_M(\pi|d_\pi) P_M(u) \prod_{K_j \in \mathcal{K}} P_M(k_j|d_{K_j}) \prod_{i \in I} \sum_{b_i \in B_i} P_M(a_i|b_i) P_M(b_i|d_{B_i}) \qquad (36) \\
&= P_M(\pi, k, a, u). \qquad (37)
\end{aligned}
$$

Step (36) results from definition (34) and step (37) from (35). Let a_{-i} denote the action profile $a \in \mathcal{A}$ with a_i removed. From the definition of an intervention (7) and the preceding result,

$$
\begin{aligned}
P_{M'|\hat{a_i}}(\mathcal{O}) &= \sum_{u \in \mathcal{U}} P_{M'}(\pi|d_\pi) \prod_{K_j \in \mathcal{K}} P_{M'}(k_j|d_{K_j}) \prod_{j \in I \setminus \{i\}} P_{M'}(a_j|d_{A_j}) P_{M'}(u) \\
&= \sum_{u \in \mathcal{U}} P_M(\pi|d_\pi) \prod_{K_j \in \mathcal{K}} P_M(k_j|d_{K_j}) \prod_{j \in I \setminus \{i\}} \sum_{b_j \in B_j} P_M(a_j|b_j) P_M(b_j|d_{B_j}) P_M(u) \\
&= P_{M|\hat{a_i}}(\mathcal{O}).
\end{aligned}
$$

□

D.2 Proof

The model M' meets all the criteria for Tian and Pearl (2002), Theorem 3, which also provides the formula for $P_{M'|\hat{a_i}}$ when the intervention is identifiable.

References

Alchian, Armen A. and Harold Demsetz. 1972. "Production, information costs, and economic organization." *American Economic Review* 62:777–795.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bratman, ME. 2014. *Shared agency: A planning theory of acting together*.

Burt, Ronald S., S. M. Gabbay, G. Holt, and Peter Moran. 1994. "Contingent organization as a network theory: The cutlure performance contingency function." *Acta Sociologica* 37:345–370.

Coase, Ronald H. 1937. "The nature of the firm." *Economica* 16:386–405.

Elwert, Felix. 2013. *Handbook of Causal Analysis for Social Research*, chapter Graphical causal models, pp. 245–273. Dordrecht: Springer Science+Business Media.

Elwert, Felix and Nicholas Christakis. 2006. "Widowhood and race." *American Sociological Review* 71:16–41.

Elwert, Felix and Christopher Winship. 2014. "Endogenous selection bias: The problem of conditioning of a collider variable." *Annual Review of Sociology* 40:31–53.

Gilbert, M. 2014. *Joint commitment: How we make the social world*. New York: Oxford University Press.

Gordon, GG and N DiTomaso. 1992a. "Predicting corporate performance from organizational culture*." *Journal of management studies* .

Gordon, G. G. and N. DiTomaso. 1992b. "Predicting corporate performance from organizational culture." *Journal of Management Studies* 29:783–799.

Grant, RM. 1996. "Prospering in dynamically-competitive environments: Organizational capability as knowledge integration." *Organization science* 7:375–87.

Grossman, Sanford J. and Oliver D. Hart. 1986. "The costs and benefits of ownership: A theory of vertical and lateral integration." *Journal of Political Economy* 94:691–719.

Harrison, JR and G Carroll. 2006. *Culture and demography in organizations*. Princeton: Princeton University Press.

Harrison, J. Richard and Glenn Carroll. 1991. "Keeping the faith: A model of cultural transmission in formal organizations." *Administrative Science Quarterly* 36:552–582.

Jensen, Michael C. and William H. Meckling. 1976. "Theory of the firm: Managerial behavior, agency costs, and ownership structure." *Journal of Financial Economics* 3:305–360.

Kotter, J. P. and J. L. Heskett. 1992. *Corporate Culture and Performance*. New York: Free Press.

Kreps, D. 1990. "Corporate culture and economic theory." In *Perspectives on Positive Political Economiy*, edited by J. E. Alt and K. Shepsle, pp. 90–143. Cambridge: Cambridge University Press.

List, Christian and Philip Pettit. 2011. *Group agency: the possibility, design, and status of corporate agents*. Oxford: Oxford University Press.

March, James G. 1962. "The business firm as a political coalition." *Journal of Politics* 24:662–678.

March, James G. 1978. "Bounded rationality, ambiguity, and the engineering of choice." *The Bell Journal of Economics* 9:587–608.

March, James G. 1991. "Exploration and exploitation in organizational learning." *Organization Science* 2:71–87.

March, James G. and Herbert A. Simon. 1958. *Organizations*. New York: John Wiley.

March, J G, H A Simon, and H S Guetzkow. 1958. *Organizations*. John Wiley & Sons Inc.

Milgrom, Paul R. and John Donald Roberts. 1992. *Economics, organization, and management*. Englewood Cliffs, NJ: Prentice-Hall.

Miller, S. 2010. *The moral foundations of social institutions: A philosophical study*. Cambridge: Cambridge University Press.

Morgan, Stephen and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

O'Reilly, Charles A. 1989. "Corporations, culture and commitment: Motivation and social control in organizations." *California Management Review* 31:9–25.

O'Reilly, Charles A. and Jennifer A. Chatman. 1996. "Culture as social control: Corporations, culture and commitment." *Research in Organizational Behavior* 18:157–2000.

Ouchi, William G. 1981. *Theory Z*. Reading, MA: Addison-Wesley.

Parsons, Talcott. 1960. *Structure and Process in Modern Societies*. Glencoe, IL: Free Press.

Patterson, Orlando. 2014. "Making sense of culture." *Annual Review of Sociology* 40:1–30.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 2 edition.

Pearl, J. 2009. *Causality: Models, reasoning, and inference*. Cambridge University Press, second edition.

Pfeffer, Jeffrey and Gerald R. Salancik. 1974. "Organizational decision making as a political process: The case of a university budget." *Administrative Science Quarterly* 19:135–151.

Pratten, S. 2014. *Social ontology and modern economics*. London: Routledge.

Ross, Stephen A. 1973. "The economic theory of agency: The principal's problem." *American Economic Review* 63:134–139.

Schein, E. H. 2010. *Organizational culture and leadership*. San Francisco: John Wiley & Sons Inc, fourth edition.

Simon, Herbert A. 1945. *Administrative Behavior*. New York: Macmillan.

Simon, Herbert A. 1962. "The architecture of complexity." *Proceedings of the American Philosophical Society* 106:467–482.

Sorensen, JB. 2002. "The strength of corporate culture and the reliability of firm performance." *Administrative science quarterly* 47:70–91.

Srivastava, Sameer B., Amir Goldberg, V. Govind Manian, and Christopher Potts. forthcoming. "Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations." *Management Science* .

Swidler, Ann. 1986. "Culture in action: Symbols and strategies." *American Sociological Review* 51:273–286.

Taylor, Frederick W. 1911. *The Principles of Scientific Management*. New York: Harper.

Tian, Jin and Judea Pearl. 2002. "A General Identification Condition for Causal Effects." *Aaai/Iaai* pp. 567–573.

Tollefsen, D. 2015. *Groups as agents*. Cambridge: Polity Press.

Tuomela, Raimo. 2013. *Social Ontology: Collective Intentionality and Group Agents*. Number April 2014. Oxford: Oxford University Press.

Verma, Tom and Judea Pearl. 1988. "Causal Networks: Semantics and Expressiveness." In *Proceedings of the Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*, pp. 352–59, Corvallis, Oregon.

Weber, Max. 1968. *Economy and Society*.