Abstractions in Causal Models and Game Structures

Sylvia S. Kerkhove S.S.KERKHOVE@UU.NL
Utrecht University, Utrecht, Netherlands

Natasha Alechina N.A.ALECHINA@OU.NL
Open University, Heerlen, Netherlands, and *Utrecht University, Utrecht, Netherlands*

Mehdi Dastani M.M.DASTANI@UU.NL
Utrecht University, Utrecht, Netherlands

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

We investigate abstractions in causal and strategic models of multi-agent systems, exploiting the relationship between these models. The paper contains two main results. The first one demonstrates that abstraction in causal models is a faithful correspondent of abstraction in strategic models, i.e., for a given causal model, if we generate a corresponding strategic model and abstract this model, we will obtain the same model as when we abstract the causal model first and then generate its corresponding strategic model. The second result is that a causal dependency in an abstract (high-level) model entails a causal dependency in the original (low-level) model. This allows us to reason about causes in a simpler abstract model and derive conclusions about causality in the much larger low-level model. These results set the stage for studying and using abstractions of causal models in multi-agent settings.

Keywords: causality, abstraction, multi-agent systems, concurrent game structures

1. Introduction

Causality plays an important role in multi-agent settings. In such settings, agents' decisions may depend on each other and those decisions may initiate a causal chain of effects in their environment. Multi-agent settings are complex, they consist of a large number of agents, agent decisions and environment attributes. This makes reasoning in these settings complicated, especially if causal relations are taken into account. Besides, while causal models could depict multi-agent settings, they are not designed to reason about strategic abilities (Kerkhove et al. (2025)). There are hence two challenges in such causally driven multi-agent settings. First, how to deal with the complexity; and second, how to deal with strategic reasoning.

In recent years, there has been some work studying this first issue by developing abstraction techniques for causal models. For example, Chalupka et al. (2017) propose an approach to deal with the complexity by automatically learning an abstraction from low-level data, and Zennaro (2022) provides an overview of other approaches that deal with the complexity issue. However, in this paper we will focus on the work by Beckers and Halpern (2019) that builds on the work by Rubenstein et al. (2017). In their work, abstraction is defined as a surjective map from low-level variable assignments to assignments of high-level variables. This approach seems most promising to generate abstractions that can be used to compare reasoning about causality in both higher- and lower-level models.

The second issue, how to deal with strategic reasoning, is addressed in Kerkhove et al. (2025). They propose a systematic way to generate a concurrent game structure (CGS) from a given causal

model describing a multi-agent system. In the resulting causal CGS, agent actions correspond to interventions on variables in the causal model.

In this paper, we deal with these two challenges in one unified framework, where we start with a structural causal model with its associated causally informed multi-agent setting for strategic reasoning of involved agents (depicted by arrow (2) in Figure 1). This paper contains two main results. The first one demonstrates that abstraction in causal models is a faithful correspondent of abstraction in concurrent game structures. More precisely, we show that the existing abstracting technique introduced by Beckers and Halpern (2019) (arrow (1) in Figure 1) is equivalent to the abstraction technique for CGS that we propose in this paper (arrow (3) in Figure 1), in the sense that the abstracted causal CGS is identical to the causal CGS that is associated with the causal model obtained from abstracting the original causal model using the technique from Beckers and Halpern (2019). This equivalence is illustrated in Figure 1 by the two paths starting from a low-level causal model $\mathcal{M}_L$ and ending at the high-level concurrent game structure CGS_H (the paths $1-4$ and $2-3$). In other words, Figure 1 shows that if we start with a causal model $\mathcal{M}_L$, it does not matter whether we first abstract (arrow (1)) and then generate a causal CGS (arrow (4)), or if we first generate a causal CGS (arrow (2)) and then abstract the CGS (arrow (3)), we will always end up with the same causal CGS, CGS_H.

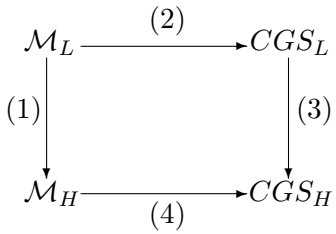

Figure 1: A visual representation of the equivalence between causal CGS that are first generated and then abstracted, and causal CGS that are generated from an abstracted causal model.

The second result shows that a causal relationship in a high-level model entails a causal relationship in the low-level model. More precisely, we prove that if in a high-level abstract model, some high-level variables having certain values is an actual cause (as defined in Halpern (2016)) of some formula φ being true, then in the low-level causal model the corresponding low-level variables and their values are the cause of a low-level translation of φ. This makes it possible to reason about causes in a smaller abstract model and derive conclusions about causality in much larger low-level model. Actual causality may be used to determine responsibility and explanations of events in multi-agent settings (Halpern (2016)).

In the following section, we will first discuss the necessary definitions of causal models and abstraction that we will be using. In Section 3, we will prove the equivalence of the abstraction technique from Beckers and Halpern (2019) to the one that we introduce for causal CGS. In Section 4, we will show how actual causality is preserved in abstractions. Finally, we will discuss limitations and future work in Section 5.

2. Background

In this section, we will define the causal models and abstraction techniques that we will be using in the rest of this paper. We will also give the definition of the causal concurrent game structure from Kerkhove et al. (2025).

2.1. Causal Models

We consider deterministic structural causal models as used in Halpern (2016). In such models, the world is modelled through a set of variables, and the causal dependencies are depicted by (structural) equations relating the variables.

Definition 1 (Causal Model, Causal Setting (Halpern (2016))) *A* causal model $\mathcal{M}$ *is a pair* $(\mathcal{S}, \mathcal{F})$*, where* $\mathcal{S}$ *is a signature and* $\mathcal{F}$ *defines a set of structural equations, relating the values of the variables. A* signature $\mathcal{S}$ *is a tuple* $(\mathcal{U}, \mathcal{V}, \mathcal{R})$*, where* $\mathcal{U}$ *is a set of exogenous variables,* $\mathcal{V}$ *is a set of endogenous variables and* $\mathcal{R}$ *associates with every variable* $X \in \mathcal{U} \cup \mathcal{V}$ *a nonempty set* $\mathcal{R}(X)$ *of possible values for* X*.*

A causal setting *is a tuple* $(\mathcal{M}, \mathbf{u})$*, where* $\mathcal{M}$ *is a causal model and* $\mathbf{u}$ *a setting for the exogenous variables in* $\mathcal{U}$*, called the* context*.*

Exogenous variables are variables that are not of interest to the modeller, i.e. the way they get assigned their value is not modelled. The values of endogenous variables are determined by the model, these are the variables that the modeller is explicitly interested in. A formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$, is called a *primitive event* (Halpern (2016)).

If there are no cyclic dependencies in the causal model, we say that it is recursive. In such models, the value of the context determines the values of all endogenous variables. This works in a recursive manner. First there are endogenous variables that purely depend on exogenous variables, these are called level-1 variables and get assigned a value first. Then there are level-2 variables that depend on exogenous variables and level-1 variables. This continues until all variables have been assigned a value (Halpern (2016)).

In a causal model it is possible to reason about what would happen if a variable Y would be set to a value y. This is an intervention and it is written as $Y \leftarrow y$. If variable X has value x after the intervention $Y \leftarrow y$, this is written as $(\mathcal{M}, \boldsymbol{u}) \models [Y \leftarrow y]X = x$, or equivalently $(\mathcal{M}^{Y \leftarrow y}, \boldsymbol{u}) \models X = x$. It is also possible to do interventions on sets of variables, written $\boldsymbol{Y} \leftarrow \boldsymbol{y}$. In general, interventions are possible on every variable. It is however possible to define a set of allowed interventions, $\mathcal{I}$, which restricts the variables on which we are allowed to intervene, as was done in Beckers and Halpern (2019). It might after all not be physically possible to directly intervene on all variable sets.

Let us now look at an example of a causal model.

Example 1 *Consider a grid world with two agents who want to bake a cake. In order to do that, agent* 1 *needs to pick up eggs and agent* 2 *needs to pick up flour before they come to the kitchen with the oven to bake the cake. This is schematically shown in Figure 2. For simplicity, agents can only move north or east. We can describe this with the following causal model:* $\mathcal{M}_L = (\mathcal{U}_L, \mathcal{V}_L, \mathcal{R}_L, \mathcal{F}_L)$*, where* $\mathcal{U}_L = \{U_{M_t^k}\}_{k,t\in\{1,2\}}$*,* $\mathcal{V}_L = \{M_t^k, P_0^k, P_t^k, Eggs^1, Flour^2, Cake\}_{k,t\in\{1,2\}}$*, because agents can make multiple moves,* M_t^k *is the move made by agent* k *at time* t*, we set* $t_{max} = 2$*, so agents get to take* 2 *steps.* P_t^k *is the position of agent* k *at time* t*,* $Flour^2$ *is whether agent* 2 *has*

picked up the flour, and $Eggs^1$ *is whether agent* 1 *has picked up the eggs,* $Cake$ *is whether the agents succeed in baking the cake.* $\mathcal{R}_L(M_t^k) = \{North, East\}$, $\mathcal{R}_L(P_t^k) = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,0\rangle, \langle 1,1\rangle\}$, *and* $\mathcal{R}_L(Flour^k) = \mathcal{R}_L(Eggs^k) = \mathcal{R}_L(Cake) = \{\top, \bot\}$. *The structural equations are:*

- $M_t^k := U_{M_t^k}$; $P_0^k := \langle 0,0\rangle$;
- $P_t^k := \begin{cases} P_{t-1}^k + \langle 1,0\rangle, & \text{if } M_t^k = East \text{ and } \quad P_{t-1}^k = \langle 0, y\rangle \\ P_{t-1}^k + \langle 0,1\rangle, & \text{if } M_t^k = North \text{ and } \quad P_{t-1}^k = \langle x, 0\rangle \\ P_{t-1}^k, & \text{else;} \end{cases}$
- $Eggs^1 := \exists t : P_t^1 = \langle 0,1\rangle$; $Flour^2 := \exists t : P_t^2 = \langle 1,0\rangle$; $Cake := (\forall k : P_{t_{max}}^k = \langle 1,1\rangle) \wedge (Eggs^1 \wedge Flour^2)$.

The set of allowed interventions $\mathcal{I}_L$ *is given by the logical closure of* $\{(\bigwedge_{t=1}^{t_{max}} M_t^k \leftarrow m), Cake \leftarrow c\}_{k\in\{1,2\}}$, *for some values* $m \in \mathcal{R}_L(M_t^k), c \in \mathcal{R}_L(Cake)$.

Following notation in Beckers and Halpern (2019), we might write $\mathcal{M}(\boldsymbol{u})$ to denote vector $\boldsymbol{v} \in \mathcal{R}(\mathcal{V})$ such that $(\mathcal{M}, \boldsymbol{u}) \models \mathcal{V} = \boldsymbol{v}$. This is naturally extended to models that are intervened on: $\mathcal{M}(\boldsymbol{u}, \boldsymbol{Y} \leftarrow \boldsymbol{y})$ denotes $\boldsymbol{v} \in \mathcal{V}$ such that $(\mathcal{M}, \boldsymbol{u}) \models [\boldsymbol{Y} \leftarrow \boldsymbol{y}]\mathcal{V} = \boldsymbol{v}$.

Given a causal model we can reason about actual causes. An actual cause is explaining a specific event (Halpern (2016)). For example, in the specific event that the light turns on, the flicking of the switch was the actual cause. There are contending definitions for actual causes (see Halpern (2016); Beckers and Vennekens (2018); Gladyshev et al. (2023) for some). In this paper we will consider the modified Halpern-Pearl definition.

Definition 2 (modified HP Definition (Halpern (2016))) *Let* φ *be a Boolean combination of primitive events.* $\boldsymbol{X} = \boldsymbol{x}$ *is an* actual cause *of* φ *in the causal setting* $(\mathcal{M}, \mathbf{u})$ *if the following 3 conditions hold:*

AC1. $(\mathcal{M}, \mathbf{u}) \models \boldsymbol{X} = \boldsymbol{x}$ *and* $(\mathcal{M}, \mathbf{u}) \models \varphi$;

AC2. *There is a set* $\boldsymbol{W}$ *of variables in* $\mathcal{V}$ *and a setting* $\boldsymbol{x}'$ *of variables in* $\boldsymbol{X}$ *s.t. if* $(\mathcal{M}, \mathbf{u}) \models \boldsymbol{W} = \mathbf{w}^*$, *then* $(\mathcal{M}, \mathbf{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \mathbf{w}^*]\neg\varphi$.

AC3. X *is minimal; there is no strict subset* $\boldsymbol{X}'$ *of* $\boldsymbol{X}$ *s.t.* $\boldsymbol{X}' = \boldsymbol{x}'$ *satisfies* AC1 *and* AC2, *where* $\boldsymbol{x}'$ *is the restriction of* $\mathbf{x}$ *to the variables in* $\boldsymbol{X}'$.

If $\boldsymbol{W} = \emptyset$, we call $\boldsymbol{X} = \boldsymbol{x}$ a *but-for cause* of φ.

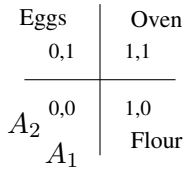

Figure 2: An example of the 2 by 2 baking grid world with 2 agents.

Example 2 *In the baking grid-world from Example 1, in the causal setting where agent* 1 *moves north and east, and agent* 2 *moves east and north, we have that* $M_1^1 = North$ *is an actual cause (a but-for cause to be precise) of the cake being baked. After all, if the agent would have moved east at step one instead, they would not have grabbed the eggs, and they would not have had all the ingredients to bake the cake.*

2.2. Causal Abstraction

In this section we will discuss the abstraction definitions introduced in Beckers and Halpern (2019). The base of an abstraction is a surjective map between variable settings.

Definition 3 *If* $\mathcal{M}_L$ *and* $\mathcal{M}_H$ *are causal models, define* $\tau : \mathcal{R}(\mathcal{V}_L) \rightarrow \mathcal{R}(\mathcal{V}_H)$ *to be a surjective map, mapping vectors of values of all endogenous variables of* $\mathcal{M}_L$ *to vectors of values of all endogenous variables of* $\mathcal{M}_H$*.*

Define $\tau_{\mathcal{U}} : \mathcal{R}(\mathcal{U}_L) \rightarrow \mathcal{R}(\mathcal{U}_H)$ *to be a surjective map, mapping vectors values of all exogenous variables of* $\mathcal{M}_L$ *to vectors of values of all exogenous variables of* $\mathcal{M}_H$*.*

Next to mapping vectors of variable values between different levels of abstraction, Beckers and Halpern (2019) also define a way to map interventions from a lower to a higher level. Given a set $\mathcal{V}$ of endogenous variables, $\boldsymbol{X} \subseteq \mathcal{V}$, and $\boldsymbol{x} \in \mathcal{R}(\boldsymbol{X})$, let $Rst(\mathcal{V}, \boldsymbol{x}) = \{\boldsymbol{v} \in \mathcal{R}(\mathcal{V}) : \boldsymbol{x}$ is the restriction of $\boldsymbol{v}$ to $\boldsymbol{X}\}$. $Rst(\mathcal{V}, \boldsymbol{x})$ is hence the set of vectors in $\mathcal{R}(\mathcal{V})$ that, when restricted to the variables in $\boldsymbol{X}$, equal $\boldsymbol{x}$.

Definition 4 (Beckers and Halpern (2019)) *Given* $\tau : \mathcal{R}(\mathcal{V}_L) \rightarrow \mathcal{R}(\mathcal{V}_H)$*, define* $\omega_\tau(\boldsymbol{X} \leftarrow \boldsymbol{x}) = \boldsymbol{Y} \leftarrow \boldsymbol{y}$ *if there exists* $\boldsymbol{Y} \subseteq \mathcal{V}_H$ *and* $\boldsymbol{y} \in \mathcal{R}(\boldsymbol{Y})$ *such that* $\tau(Rst(\mathcal{V}_L, \boldsymbol{x})) = Rst(\mathcal{V}_H, \boldsymbol{y})$ *(as usual, given* $T \subseteq \mathcal{R}(\mathcal{V}_L)$*, we define* $\tau(T) = \{\tau(\boldsymbol{v}_L) : \boldsymbol{v}_L \in T\}$*). If there does not exist such a* $\boldsymbol{Y}$ *and* $\boldsymbol{y}$*, we take* $\omega_\tau(\boldsymbol{X} \leftarrow \boldsymbol{x})$ *to be undefined. Let* $\mathcal{I}_H^\tau$ *be the set of interventions for which* ω_τ *is defined, and let* $\mathcal{I}_H^\tau = \omega_\tau(\mathcal{I}_L^\tau)$*.*

For any $\boldsymbol{X}$ and $\boldsymbol{x}$, there can be at most one $\boldsymbol{Y}$ and $\boldsymbol{y}$ such that $\omega_\tau(\boldsymbol{X} \leftarrow \boldsymbol{x}) = \boldsymbol{Y} \leftarrow \boldsymbol{y}$. Suppose there could also be a $\boldsymbol{Z}$ and $\boldsymbol{z}$ such that $\omega_\tau(\boldsymbol{X} \leftarrow \boldsymbol{x}) = \boldsymbol{Z} \leftarrow \boldsymbol{z}$, we would have that $Rst(\mathcal{V}_H, \boldsymbol{y}) = Rst(\mathcal{V}_H, \boldsymbol{z})$. Let $\boldsymbol{v} \in Rst(\mathcal{V}_H, \boldsymbol{y})$, then we can without loss of generality write $\boldsymbol{v} = (y_1, y_2, .., y_k, v_{k+1}, ..., v_n)$, where $(y_1, ..., y_i) = \boldsymbol{y}$. If $\boldsymbol{Z} \neq \boldsymbol{Y}$, restricting $\boldsymbol{v}$ to $\boldsymbol{z}$ gives that there must be a $v_l \in (v_{k+1}, ..., v_n)$ that is also in $\boldsymbol{z}$, or that $\boldsymbol{Z} \subseteq \boldsymbol{Y}$. In this latter case, we have that $Rst(\mathcal{V}_H, \boldsymbol{z}) \neq Rst(\mathcal{V}_H, \boldsymbol{y})$, because $Rst(\mathcal{V}_H, \boldsymbol{z})$ would be larger, given that the possibilities for the v_i's would be bigger, so we cannot have that $\boldsymbol{Z}$ is contained in $\boldsymbol{Y}$. Returning to the first option, that a $v_l \in (v_{k+1}, ..., v_n)$ is also in $\boldsymbol{z}$. We also have that $\boldsymbol{v}' = (y_1, .., y_k, v_{k+1}, ..., v'_l, ..., v_n) \in Rst(\mathcal{V}_H, \boldsymbol{y})$, by the definition of Rst. However v'_l is not in $\boldsymbol{z}$, so $\boldsymbol{v}' \notin Rst(\mathcal{V}_H, \boldsymbol{z})$, and hence $Rst(\mathcal{V}_H, \boldsymbol{y}) \neq Rst(\mathcal{V}_H, \boldsymbol{z})$. Finally, if $\boldsymbol{Y} = \boldsymbol{Z}$, but $\boldsymbol{y} \neq \boldsymbol{z}$, $Rst(\mathcal{V}_H, \boldsymbol{y}) \neq Rst(\mathcal{V}_H, \boldsymbol{z})$, because there are no vectors that when restricted to $\boldsymbol{Y}$ equal two different vectors.

Definition 5 (Beckers and Halpern (2019)) $\tau_{\mathcal{U}} : \mathcal{R}(\mathcal{U}_L) \rightarrow \mathcal{R}(\mathcal{U}_H)$ *is* compatible *with* $\tau : \mathcal{R}(\mathcal{V}_L) \rightarrow \mathcal{R}(\mathcal{V}_H)$ *if, for all* $\boldsymbol{Y} \leftarrow \boldsymbol{y} \in \mathcal{I}_L$ *and* $\boldsymbol{u}_L \in \mathcal{R}(\mathcal{U}_L)$*,*

$$\tau(M_L(\boldsymbol{u}_L, \boldsymbol{Y} \leftarrow \boldsymbol{y})) = M_H(\tau_{\mathcal{U}}(\boldsymbol{u}_L), \omega_\tau(\boldsymbol{Y} \leftarrow \boldsymbol{y})).$$

Definition 5 says that no matter whether you abstract using τ, or using $\tau_{\mathcal{U}}$ and ω_τ, you will end up with the same abstraction.

Now, we can put everything together to define what it means for a causal model to be an abstraction of another. We denote the largest possible set of allowed interventions with $\mathcal{I}_H^*$. Hence $\mathcal{I}_H^*$ is the set of the interventions we could take in a model where we would not have specified the allowed interventions at all. The following is a rewritten version of the definition in Beckers and Halpern (2019).

Definition 6 *$(M_H, \mathcal{I}_H)$, with $\mathcal{V}_H = \{Y_1, ..., Y_n\}$, is a* constructive τ-abstraction *of $(M_L, \mathcal{I}_L)$ if the following conditions hold:*

- *τ and $\tau_{\mathcal{U}}$ exist, with $\tau_{\mathcal{U}}$ being compatible with τ;*
- *$\mathcal{I}_H^* = \omega_\tau(\mathcal{I}_L)$;*
- *There exists a partition $P = \{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_{n+1}\}$ of $\mathcal{V}_L$, where $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ are non-empty, and mappings $\tau_{Y_i} : \mathcal{R}(\boldsymbol{Z}_i) \to \mathcal{R}(Y_i)$ for $i = 1, ..., n$ such that $\tau(\boldsymbol{v}_L) = \tau_{Y_1}(\boldsymbol{z}_1) \cdot ... \cdot \tau_{Y_n}(\boldsymbol{z}_n)$, where $\boldsymbol{z}_i$ is the projection of $\boldsymbol{v}_L$ onto the variables in $\boldsymbol{Z}_i$, and $\cdot$ is the concatenation operator on sequences.*

M_H is a constructive abstraction *of M_L if it is a constructive τ-abstraction M_L for some τ.*

If we have a constructive abstraction, each higher level variable has an associated set of lower level variables that determine the value of the higher level variable.

Example 3 *We define the causal model $\mathcal{M}_H = (\mathcal{U}_H, \mathcal{V}_H, \mathcal{R}_H, \mathcal{F}_H, \mathcal{I}_H^*)$, with $\mathcal{U}_H = \{U_{Flour}, U_{Eggs}\}$, $\mathcal{V} = \{Flour, Eggs, Cake\}$, where $Flour$ denotes whether the flour gets to the oven, $Eggs$ whether the eggs get to the oven, and $Cake$ whether the cake gets baked. The range of $\mathcal{R}_H(Flour) = \mathcal{R}(Eggs) = \mathcal{R}(Cake) = \{\top, \bot\}$. The structural equations are:*

- *$Flour := U_{Flour}$; $Eggs := U_{Eggs}$;*
- *$Cake := Eggs \wedge Flour$.*

This is a constructive abstraction of the earlier baking grid-world causal model described in Example 1. To see this, define $\tau : \mathcal{R}_L(\mathcal{V}_L) \to \mathcal{R}_H(\mathcal{V}_H)$, as $\tau(m_1^1, m_2^1, m_1^2, m_2^2, p_1^1, p_2^1, p_1^2, p_2^2, e^1, f^2, c) = \langle e, f, c\rangle$, where e is given by $e = (m_1^1 = North) \wedge (m_2^1 = East)$, and $f = (m_1^2 = East) \wedge (m_2^2 = North)$. This is a surjective map, that can be partitioned into variable sets.

Also define $\tau_{\mathcal{U}} : \mathcal{R}_L(\mathcal{U}_L) \to \mathcal{R}_H(\mathcal{U}_H)$, as $\tau_{\mathcal{U}}(u_{M_1^1}, u_{M_2^1}, u_{M_1^2}, u_{M_2^2}) = \langle u_{Eggs}, u_{Flour}\rangle$, where $u_{Eggs} = (u_{M_1^1} = North) \wedge (u_{M_2^1} = East)$, and $u_{Flour} = (u_{M_1^2} = East) \wedge (u_{M_2^2} = North)$. $\tau_{\mathcal{U}}$ is also surjective.

$\tau_{\mathcal{U}}$ is compatible with τ. For example, let $\boldsymbol{u}_L = \langle U_{M_1^1} = North, U_{M_2^1} = East, U_{M_1^2} = East, U_{M_2^2} = North\rangle$, and consider the intervention $(M_1^1 \leftarrow East) \wedge (M_2^1 \leftarrow North)$. We have that $\mathcal{M}_L(\boldsymbol{u}_L, ((M_1^1 \leftarrow East) \wedge (M_2^1 \leftarrow North))) = \langle E, N, E, N, \langle 1, 0\rangle, \langle 1, 1\rangle, \langle 1, 0\rangle, \langle 1, 1\rangle, \bot, \top, \bot\rangle$. So both agents take the same path and pick up flour and no eggs, so there will not be a cake. Now, $\tau(E, N, E, N, \langle 1, 0\rangle, \langle 1, 1\rangle, \langle 1, 0\rangle, \langle 1, 1\rangle, \bot, \top, \bot) = (\neg Eggs, Flour, \neg Cake)$. This is the same as $\mathcal{M}_H(\tau_{\mathcal{U}}(\boldsymbol{u}_L)\ \omega_\tau((M_1^1 \leftarrow East) \wedge (M_2^1 \leftarrow North)) = \mathcal{M}_H((\top, \top), (Eggs \leftarrow \bot)) =$

$(\neg Eggs, Flour, \neg Cake)$. *We can see that this equality also holds for other pairs of settings and interventions.*

Now, we just need to show that $\omega_\tau(\mathcal{I}_L) = \mathcal{I}_H^*$. *To see that, consider that* $\omega_\tau(M_1^1 \leftarrow North, M_2^1 \leftarrow East) = Eggs \leftarrow \top$, *as* $\tau(Rst(\mathcal{V}_L, (M_1^1 = North, M_2^1 = East))) = Rst(\mathcal{V}_H, (Eggs = \top))$, *as* $Eggs$ *can only be true if agent* 1 *moves north and east. Similarly,* $\omega_\tau(M_1^1 \leftarrow North, M_2^1 \leftarrow North) = Eggs \leftarrow \bot$. *We can do something similar for* $Flour$, *and* $\omega_\tau(Cake \leftarrow c) = Cake \leftarrow c$ *for any value* $c \in \mathcal{R}_H(Cake)$.

$\mathcal{M}_H$ *is hence a constructive* τ*-abstraction of* $\mathcal{M}_L$.

2.3. Causal CGS

A causal concurrent game structure is a type of concurrent game structure (CGS) that is generated from a causal model in such a way that certain properties of the causal model are carried over. They were introduced by Kerkhove et al. (2025). To construct a causal CGS, first divide the set of endogenous variables, $\mathcal{V}$, into a set of agent variables, V_a, whose values are controlled by the agents of the model, and into a set of environment variables, V_e, whose values follow from the agent variables and the context of the causal setting. We then order these variables, as defined below.

Definition 7 (Agent Rank (Kerkhove et al. (2025))) *An* agent ranking function *of a causal model* $\mathcal{M}$ *is a function* $\rho : \mathcal{V} \to \{0, ..., n\}$, *where* n *is the number of distinct variable levels for agent variables in* $\mathcal{M}$, *such that for all* $A, B \in V_a$, $\rho(A) > \rho(B) > 0$ *if and only if the variable level of* A *is higher than the variable level of* B, *and* $\rho(A) = \rho(B)$ *if and only if* A *and* B *have the same variable level. For all* $X \in V_e$, $\rho(X) = \rho(A) - 1$ *if* $\exists A \in V_a$ *such that the variable level of* X *is lower or equal to the variable level of* A, *and there is no* $B \in V_a$ *that has a variable level between* X *and* A. *If such an* A *does not exist, i.e. if the variable level of* X *is higher than the variable level of all* $A \in V_a$, *then* $\rho(X) = n$. *The* agent rank *of a variable* $A \in V_a$ *is* $\rho(A)$.

The variables are ordered in such a way that variables that depend on others are later in the CGS. The set of actions for an agent in the CGS is the set of different values for a corresponding agent variable in the causal model. Performing an action in the CGS corresponds to an intervention on that agent variable in the causal model. We formalise this in the definition below.

Definition 8 (Causal Concurrent Game Structure) *Let* $(\mathcal{M}, \boldsymbol{u})$ *be a causal setting, with* $n = \max_{Y \in V_a} \rho(Y)$. *A causal concurrent game structure generated by this causal setting is a tuple* $\langle N, Q, d, \delta, \Pi, \pi \rangle$ *such that:*

- $N = |\mathcal{V}_a|$, *is the number of agents;*
- $Q = \{q_{0,0}\} \cup \{q_{i,j} \mid 1 \leq i \leq n, 0 \leq j < m_i\}$, *where* $m_i = \prod_{Y \in \mathcal{V}_a, \rho(Y) \leq i} |\mathcal{R}(Y)|$, *is the set of states.;*
- *For* $k \in \{1, ..., N\}$, *and* $q_{i,j} \in Q$, $d_k(q_{i,j}) = \mathcal{R}(V_k)$, *are the moves available for agent* k *at state* $q_{i,j}$, *if the agent variable* V_k *for agent* k *has agent rank* $\rho(V_k) = i + 1$, *else* $d_k(q_{i,j}) = \{0\}$;
- δ *is the transition function. For* $i < n$, *for any* $\mathbf{a}_{i,j} \in \mathbf{A}_i$, $\delta(q_{i,j}, \mathbf{a}_{i,j}) = q_{i+1,j'}$, *where* $|\mathbf{A}_i| \cdot j \leq j' \leq |\mathbf{A}_i| \cdot (j+1) - 1$, *under the condition that if* $\mathbf{a}_{i,j} \neq \mathbf{a}'_{i,j}$, *then* $\delta(q_{i,j}, \mathbf{a}_{i,j}) \neq \delta(q_{i,j}, \mathbf{a}'_{i,j})$. *If* $i = n$, $\delta(q_{i,j}, \mathbf{a}_{i,j}) = q_{i,j}$;

- $\Pi = \{X = x | X \in \mathcal{V}, x \in \mathcal{R}(X)\}$, *is the set of propositions;*
- *The labelling function* π *is defined recursively, with* $\pi(q_{0,0}) = \mathcal{M}(\boldsymbol{u})$, *and* $\pi(\delta(q_{i,j}, \boldsymbol{a}_{i,j})) = \mathcal{M}(\boldsymbol{u}, (\boldsymbol{X}_{i,j} \leftarrow \boldsymbol{x}_{i,j}, \boldsymbol{A}_{i,j} \leftarrow \boldsymbol{a}_{i,j}))$. *Here* $\boldsymbol{a}_{i,j}$ *is an action profile for state* $q_{i,j}$, $\boldsymbol{A}_{i,j} \leftarrow \boldsymbol{a}_{i,j} := \{A_k \leftarrow a_k \mid A_k \in V_a, \rho(A_k) = i + 1 \text{ and } a_k \in \boldsymbol{a}_{i,j}\}$ *is an intervention constructed based on action profile* $\boldsymbol{a}_{i,j}$, *and* $\boldsymbol{X}_{i,j} \leftarrow \boldsymbol{x}_{i,j}$ *is recursively defined by:* $\boldsymbol{X}_{i+1,j'} \leftarrow \boldsymbol{x}_{i+1,j'} := \boldsymbol{X}_{i,j} \leftarrow \boldsymbol{x}_{i,j} \cup \boldsymbol{A}_{i,j} \leftarrow \boldsymbol{a}_{i,j}$, *if* $\delta(q_{i,j}, \boldsymbol{a}_{i,j}) = q_{i+1,j'}$ *with* $\boldsymbol{X}_{0,0} \leftarrow \boldsymbol{x}_{0,0} = \emptyset$.

Example 4 *Consider the causal model for the baking grid-world as described in Example 1. We consider the move variables* M_t^k *to be the agent variables and all other variables are the environment variables. The move variables all have an agent rank of* 1, *and they all have* 2 *possible values. We hence get that the set of states of the corresponding causal CGS will be* $Q = \{q_{0,0}, q_{1,0}, ..., q_{1,15}\}$, *as there are* 4 *move variables,* 2 *for each agent, so we get the starting state plus* 2^4 *states in the first level. All agent variables get to take an action in state* $q_{0,0}$ *and there they can choose to move either north or east. Hence, despite there only being* 2 *agents in the system, there will be* 4 *agent actions in the causal CGS, as each agent has* 2 *corresponding agent actions. Some of the transitions and valuations are shown in Figure 3.*

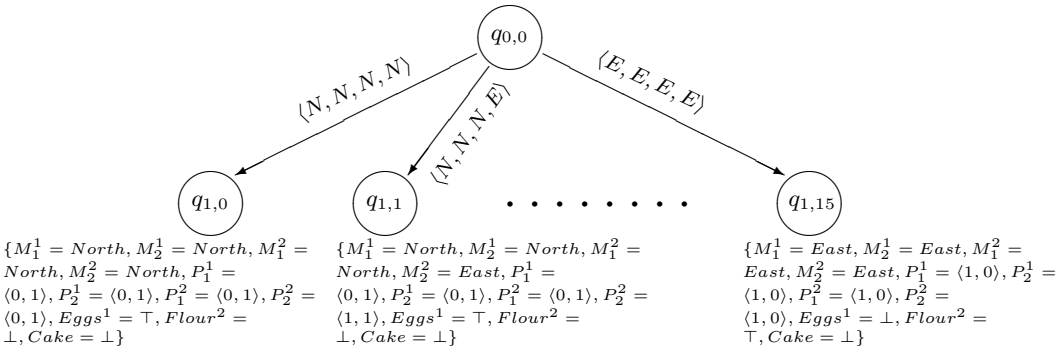

Figure 3: A visual representation of the causal CGS of Example 4, we omitted some states and transitions for space reasons. N denotes the action $North$ and E denotes the action $East$.

CGS are useful to reason about agent strategies. Strategies describe the actions that agents would do in certain specified situations.

Definition 9 (Strategy in Concurrent Game Structures (Alur et al. (2002))) *Given a concurrent game structure* $S = \langle N, Q, d, \delta, \Pi, \pi \rangle$, *a* strategy *for agent* $a \in \{1, ..., N\}$ *is a function* f_a, *that maps any (non-empty) finite sequence* λ *of states in* Q *to an action the agent can take at the last state of the sequence. I.e. if* q *is the last state of* λ, *then* $f_a(\lambda) \in d_a(q)$. *We write* $F_A = \{f_a \mid a \in A\}$ *for a set of strategies of the agents in* $A \subseteq \{1, ..., N\}$.

Following notation in Kerkhove et al. (2025), we will use the notation $F_{X_k=x}$ for the agent strategy where agent variable X_k gets assigned value x. We write $F_{\boldsymbol{X}=\boldsymbol{x}}$ to denote the set of strategies $\{F_{X_k=x} \mid X_k \in \boldsymbol{X}, x \in \boldsymbol{x}\}$.

A causal CGS also has an associated causal strategy profile which captures the behaviour of the agents according to the underlying causal model.

Definition 10 (Causal Strategy Profile Kerkhove et al. (2025)) *Given a causal setting* $(\mathcal{M}, \mathbf{u})$ *and the causal CGS generated by this setting. Define the* causal strategy profile $F_{\mathcal{M}}$ *as* $F_{\mathcal{M}} = \{F_{X_k} \mid X_k \in V_a\}$, *where* $F_{X_k}(q_{i,j}) = 0$ *if* $\rho(X_k) \neq i+1$, *and* $F_{X_k}(q_{i,j}) = x_k$ *otherwise, where* x_k *is such that* $(\mathcal{M}, \mathbf{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]X_k = x_k$, *with* $\boldsymbol{X} = \{X_{k'} \mid \rho(X_{k'}) < \rho(X_k)\}$ *and* $\boldsymbol{x} = \{x_{k'} \mid x_{k'} \in \alpha[q_{i,j}]\}$. [1]

We write $F_k \circ F_{\mathcal{M}}$ to denote the strategy where agent k follows F_k and the rest of the agents follow the causal strategy profile.

2.4. CGS Abstraction

We propose an abstraction technique for causal CGS, adapted from one used in Varricchione et al. (2026) that abstracts multi-agent environments to epistemic CGS.

Definition 11 *A causal CGS* $H = \langle N_H, Q_H, d_H, \delta_H, \Pi_H, \pi_H \rangle$ *abstracts a causal CGS* $L = \langle N_L, Q_L, d_L, \delta_L, \Pi_L, \pi_L \rangle$ *if there is a surjective map* t *from* Π_L *to* Π_H *such that for any transition* $s \xrightarrow{\sigma} s'$ *in* H, *there is a sequence of joint actions* α *in* L *s.t. for all states* q *of* L, *such that* $\pi_H(s) = t(\pi_L(q))$, *we have that* $q \xrightarrow{\alpha} q'$ *in* L, *where* $\pi_H(s') = t(\pi_L(q'))$. *We say that a state* s *in* H *abstracts a state* q *in* L *if* $\pi_H(s) = t(\pi_L(q))$.

This definition says that H is an abstraction of L, if there is a surjective map relating the propositions in both models, and whenever a state s in H is an abstraction of one in L, any transition from s corresponds to a (sequence of) transitions in L.

Example 5 *Consider the causal CGS,* CGS_H, *depicted in Figure 4. This CGS uses the variables* $Eggs$, *to indicate whether eggs got brought to the oven,* $Flour$, *to indicate whether flour got brought to the oven, and* $Cake$, *to indicate whether the cake was baked. This causal CGS is an abstraction of the causal CGS,* CGS_L *from Example 4. To see this, consider that* $\Pi_L = \{M_t^k = North, M_t^k = East \mid 0 \leq t, k \leq 1\} \cup \{P_t^k = \langle x, y \rangle \mid 0 \leq x, y, t, k \leq 1\} \cup \{Eggs^1 = \bot, Eggs^1 = \top, Flour^2 = \bot, Flour^2 = \top, Cake = \bot, Cake = \top\}$, *and* $\Pi_H = \{Eggs = \bot, Eggs = \top, Flour = \bot, Flour = \top, Cake = \bot, Cake = \top\}$.

We can define the map $t : \Pi_L \to \Pi_H$ *as* $t(M_1^1 = m_1^1, M_2^1 = m_2^1, M_1^2 = m_1^2, M_2^2 = m_2^2, P_1^1 = p_1^1, P_2^1 = p_2^1, P_1^2 = p_1^2, P_2^2 = p_2^2, Eggs^1 = e^1, Flour^2 = f^2, Cake = c) = \langle Eggs = e, Flour = f, Cake = c \rangle$, *where* e *is given by* $e = (M_1^1 = North) \wedge (M_2^1 = East)$, *and* $f = (M_1^2 = East) \wedge (M_2^2 = North)$. *Now, to see that this map adheres to Definition 11 consider, for example, the transition* $s_{0,0} \to s_{1,1}$ *in* CGS_H. *There is only one state in* CGS_L *that is abstracted by* $s_{0,0}$, *namely* $q_{0,0}$, $t(\pi_L(q_{0,0})) = \pi_H(s_{0,0})$. *So for the sequence of joint actions, we can simply define any action that leads to a state in* CGS_L *that is abstracted by* $q_{1,1}$. *So take the joint action* $\langle Eggs = \bot, Flour = \bot \rangle$. *This gives the transition* $s_{0,0} \xrightarrow{\langle \bot, \bot \rangle} s_{1,0}$, *and indeed* $s_{1,0}$ *abstracts* $q_{1,1}$, *as* $t(\pi_L(q_{1,1})) = \pi_H(s_{1,0})$. *This can also be checked for the other possible transitions, showing that* CGS_H *is indeed an abstraction of* CGS_L.

1. Here $\alpha[q_{i,j}]$ is the *action path* of of the state $q_{i,j}$, defined formally as: for $0 \leq k \leq N$, an action a_k is in $\alpha[q_{i,j}]$ if and only if $\rho(A_k) \leq i$ and there exists an action profile $\mathbf{a}_{i',j'}$, containing a_k, such that $q_{i',j'} \in \lambda[q_{i,j}, i]$ (the history of $q_{i,j}$) and $\delta(q_{i',j'}, \mathbf{a}_{i',j'}) \in \lambda[q_{i,j}, i]$.

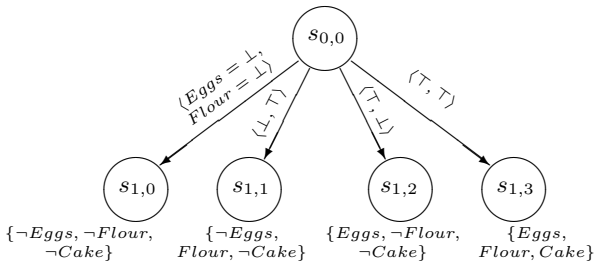

Figure 4: A visual representation of the abstracted causal CGS of Example 5.

3. Connecting Abstractions

In this section, we are going to prove that the constructive abstraction as defined in Beckers and Halpern (2019), in combination with the translation to causal CGS, gives the same result as the CGS abstraction that we defined in Definition 11. See Figure 1 for an illustration.

In order to prove this, we first need a couple of lemmas. The first lemma states that for constructive τ-abstractions (of causal models), it holds that if an intervention from the lower level gets mapped to an intervention in the higher level, and we then add another disjoint intervention in the lower level, these interventions can be concatenated in the higher level.

Lemma 12 *Let $\mathcal{M}_H$ be a constructive τ-abstraction of $\mathcal{M}_L$, let $\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L \in \mathcal{I}_L$ and $\boldsymbol{X}_H \leftarrow \boldsymbol{x}_H, \boldsymbol{Y}_H \leftarrow \boldsymbol{y}_H \in \mathcal{I}_H$, with $\boldsymbol{X}_L \cap \boldsymbol{Y}_L = \emptyset$ and $\boldsymbol{X}_H \cap \boldsymbol{Y}_H = \emptyset$. If $\omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L) = \boldsymbol{X}_H \leftarrow \boldsymbol{x}_H$, then $\omega_\tau(\boldsymbol{X}_L \cup \boldsymbol{Y}_L \leftarrow \boldsymbol{x}_L \cup \boldsymbol{y}_L) = \boldsymbol{X}_H \cup \boldsymbol{Y}_H \leftarrow \boldsymbol{x}_H \cup \boldsymbol{y}_H$.*

Sketch of Proof This follows because the map τ is a concatenation of surjective maps for all endogenous variables in $\mathcal{V}_H$. For the full proof see Appendix A. ■

Now, we will also need a result that is very similar to a lemma stated in Kerkhove et al. (2025), which says that when following a certain global strategy in the causal CGS, will lead to a leaf-state that corresponds to a certain intervention. The following lemma slightly modifies that to be applicable when starting from an arbitrary immediate state. The proof is also very similar to the one in Kerkhove et al. (2025) and can be found in Appendix B.

Lemma 13 *Let GS be a causal CGS based on a causal setting $(\mathcal{M}, \boldsymbol{u})$. If $q_{n,m}$ is the leaf-state of GS that results from the strategy profile $F_{\boldsymbol{Y}=\boldsymbol{y}} \circ F_{\mathcal{M}}$, with all agent variables in $\boldsymbol{Y}$ having an agent rank higher than k, when executed starting from any state $q_{k,l}$ corresponding to $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x}}, \mathbf{u})$ then $q_{n,m}$ corresponds to $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \mathbf{u})$.*

Now, we get to the main result of this section. We show that the constructive abstraction as defined in Beckers and Halpern (2019) (Definition 6), in combination with the translation to causal CGS, gives the same high-level causal CGS as when we would first translate the causal model to a causal CGS, and then abstract the causal CGS using Definition 11 (see Figure 1 for an illustration).

Theorem 14 *Given a causal model $\mathcal{M}_L$ and the causal CGS generated by it, CGS_L, if a causal model $\mathcal{M}_H$ is a constructive τ-abstraction of $\mathcal{M}_L$, then the causal CGS, CGS_H, generated by $\mathcal{M}_H$ is an abstraction of CGS_L.*

Sketch of Proof We need to show that CGS_H is an abstraction of CGS_L, so we must show that there is a surjective map t_τ from Π_L to Π_H that satisfies Definition 11. The main idea of this proof is that we can define t_τ to be basically the same as the map τ that abstracts the causal model. We note that $\Pi_L = \{X = x \mid X \in \mathcal{V}_L, x \in \mathcal{R}(\mathcal{V}_L)\}$ and that $\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H)$, so we can define $t_\tau : \Pi_L \to \Pi_H$ to be s.t. $t_\tau(X_1 = x_1, ..., X_n = x_n) = (Y_1 = y_1, ..., Y_m = y_m)$, where $(y_1, ..., y_m) = \tau(x_1, ..., x_n)$. We use lemmas 12 and 13 to show that this t_τ satisfies Definition 11. We show that for any transition $s \xrightarrow{\mathbf{Y} \leftarrow \mathbf{y}} s'$ in CGS_H, with q any state of CGS_L abstracted by s (i.e. $\pi_H(s) = t_\tau(\pi_L(q))$), there is a sequence of joint actions α in CGS_L such that $q \xrightarrow{\alpha} q'$, where q' is abstracted by s'. The full proof can be found in Appendix C. ■

When we go back to the examples of Section 2, we see that the causal CGS, CGS_H, of Example 5 is indeed the causal CGS for the abstracted causal model of Example 3.

This result helps to motivate the use of constructive abstractions. After all, the abstraction for causal CGS has been based on standard approaches in the literature and the constructive abstraction gives a comparable result. In the next section, we will further motivate its use by showing that actual causality is also carried over in the abstraction, in the sense that actual causes in the higher level causal model are abstractions of actual causes in the lower level model.

4. Causality Preservation in Abstractions

In this section, we show that actual causality is preserved in constructive abstractions, in the sense that if a set of variables having certain values is a cause of an event in the higher level model, then there is a corresponding set of variables having corresponding values, that is a cause of a corresponding lower-level event. In order to show this, we first need the following definition to be formally able to define what we mean with this correspondence. The definition inductively defines the pre-image of a Boolean combination of events φ.

Definition 15 *Given a Boolean combination of events φ, define τ_φ^{-1} inductively as:*

- $\tau_\varphi^{-1}(Y = y) := \tau_Y^{-1}(Y = y)$;
- $\tau_\varphi^{-1}(\psi \wedge \phi) := \tau_\varphi^{-1}(\psi) \wedge \tau_\varphi^{-1}(\phi)$; $\quad \tau_\varphi^{-1}(\psi \vee \phi) := \tau_\varphi^{-1}(\psi) \vee \tau_\varphi^{-1}(\phi)$;
- $\tau_\varphi^{-1}(\neg\psi) := \neg\tau_\varphi^{-1}(\psi)$.

We do not define the implication separately as it can be written as a combination of disjunction and negation.[2] Specifically note that the event $\boldsymbol{Y} = \boldsymbol{y}$ is in fact a conjunction of the events $Y_1 = y_1, ..., Y_n = y_n$. So $\tau_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} = \boldsymbol{y}) = \tau_{Y_1}(Y_1 = y_1) \wedge ... \wedge \tau_{Y_n}(Y_n = y_n)$.

Now, as the main result of this section, we show that if $\boldsymbol{X} = \boldsymbol{x}$ is a modified HP cause of φ, then there is a subset in the pre-image of $\boldsymbol{X} = \boldsymbol{x}$ that is a cause of the pre-image of φ.

Theorem 16 *Let $(\mathcal{M}_H, \mathcal{I}_H)$ be a constructive τ-abstraction of $(\mathcal{M}_L, \mathcal{I}_L)$. Let $\boldsymbol{X} = \boldsymbol{x}$ be a modified HP cause of boolean combination of events φ in $(\mathcal{M}_H, \mathbf{u}_H)$. Then, for all $\mathbf{u}_L$ such that $\tau_{\mathcal{U}}(\mathbf{u_L}) = \mathbf{u}_H$, there is a $\boldsymbol{Z} = \boldsymbol{z}$, a subset of an element in $\tau_{\boldsymbol{X}}^{-1}(\boldsymbol{X} = \boldsymbol{x})$, that is a modified HP cause of $\tau_\varphi^{-1}(\varphi)$ in $(\mathcal{M}_L, \mathcal{I}_L, \tau_{\mathcal{U}}(\mathbf{u_L}))$.*

2. Technically we also do not need to define both disjunction and conjunction, as they can be written in terms of each other. However, formulas are more readable if we can use both disjunction and conjunction.

Proof Take an arbitrary $\boldsymbol{u}_L \in \mathcal{R}_L(\mathcal{U}_L)$ such that $\tau_{\mathcal{U}}(\mathbf{u_L}) = \mathbf{u}_H$. Here we just show the case $\varphi = (Y = y)$, the proof for a general Boolean combination φ can be found in Appendix D.

By assumption, $\boldsymbol{X} = \boldsymbol{x}$ is a modified HP cause of $Y = y$, so $(\mathcal{M}_H, \mathcal{I}, \mathbf{u}_H) \models \boldsymbol{X} = \boldsymbol{x} \wedge Y = y$, we know that $\tau(\mathcal{M}_L(\mathbf{u}_L)) = \mathcal{M}_H(\tau_{\mathcal{U}}(\mathbf{u}_H)) = \mathcal{M}_H(\mathbf{u}_H)$. Hence, since $\boldsymbol{X} = \boldsymbol{x} \subseteq \mathcal{M}_H(\mathbf{u}_H)$, $\boldsymbol{X} = \boldsymbol{x} \subseteq \tau(\mathcal{M}_L(\mathbf{u}_L))$. Since τ is constructive, $\boldsymbol{x} = \tau_{\boldsymbol{X}}(\boldsymbol{Z} = \boldsymbol{z})$, where $\boldsymbol{Z} = \boldsymbol{z} \subseteq \mathcal{M}_L(\mathbf{u}_L)$. Similarly, $Y = y \subseteq \tau(\mathcal{M}_L(\mathbf{u}_L))$, so there exists a $\boldsymbol{Y}_L \in \mathcal{V}_L$ such that the τ_Y mapping from $\mathcal{V}_L$ to Y has $\tau_Y(\boldsymbol{y}_L) = y$ so $\boldsymbol{Y}_L = \boldsymbol{y}_L \subseteq \mathcal{M}_L(\mathbf{u}_L)$. So now we have that $(\mathcal{M}_L, \mathbf{u}_L) \models \boldsymbol{Z} = \boldsymbol{z} \wedge \boldsymbol{Y}_L = \boldsymbol{y}_L$.

Because $\boldsymbol{X} = \boldsymbol{x}$ is a modified HP cause, we have that there exists a value $\boldsymbol{x}'$ of $\boldsymbol{X}$ and a $\boldsymbol{W} \subseteq \mathcal{V}_H$ with $(\mathcal{M}_H, \mathcal{I}_H, \mathbf{u}_H) \models \boldsymbol{W} = \boldsymbol{w}^*$ such that $(\mathcal{M}_H, \mathbf{u}_H) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}^*]Y \neq y$, i.e. $Y = y \notin \mathcal{M}_H((\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}^*), \mathbf{u}_H)$.

Since $\boldsymbol{W} = \boldsymbol{w}^* \subseteq \tau(\mathcal{M}_L(\mathbf{u}_L))$, there must be a $\boldsymbol{W}_L = \boldsymbol{w}^*_L \subseteq \mathcal{M}_L(\mathbf{u}_L)$ such that $\tau_{\boldsymbol{W}}(\boldsymbol{w}^*_L) = \boldsymbol{w}^*$. Since by the definition of constructive abstraction, $\tau_{\mathcal{U}}$ needs to be compatible with τ (Definition 5), we can write $\mathcal{M}_H((\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}^*), \mathbf{u}_H)$ as $\mathcal{M}_H(\omega_\tau(\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \tau_{\mathcal{U}}(\mathbf{u}_L)) = \tau(\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \mathbf{u}_L))$, where $\boldsymbol{z}'$ is defined as $\boldsymbol{z}$ was before, but for $\boldsymbol{x}'$ instead of $\boldsymbol{x}$. $\omega_\tau(\boldsymbol{Z} \leftarrow \boldsymbol{z}') = \boldsymbol{X} \leftarrow \boldsymbol{x}'$ as $\tau(Rst(\mathcal{V}_L, \boldsymbol{z}')) = Rst(\mathcal{V}_H, \boldsymbol{x}')$, since the abstraction is constructive (this is a similar argument to Lemma 12). We know that $\boldsymbol{z}' \neq \boldsymbol{z}$, because $Y = y$ is in $\mathcal{M}_H(\omega_\tau(\boldsymbol{Z} \leftarrow \boldsymbol{z}), \tau_{\mathcal{U}}(\mathbf{u}_L))$, but not in $\mathcal{M}_H(\omega_\tau(\boldsymbol{Z} \leftarrow \boldsymbol{z}'), \tau_{\mathcal{U}}(\mathbf{u}_L))$, so $\boldsymbol{z}'$ must be different from $\boldsymbol{z}$.

By AC2 of Definition 2, $Y = y \notin \tau(\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \mathbf{u}_L))$, so $\boldsymbol{Y}_L = \boldsymbol{y}_L$ cannot be in $\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \mathbf{u}_L)$. Hence $(\mathcal{M}_L, \mathbf{u}_L) \models [\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L]\boldsymbol{Y}_L \neq \boldsymbol{y}_L$, which satisfies AC2. Now we have that either $\boldsymbol{Z} = \boldsymbol{z} \in \tau^{-1}_{\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x})$ is a modified HP cause of $\boldsymbol{Y}_L = \boldsymbol{y}_L$, or a subset of $\boldsymbol{Z} = \boldsymbol{z}$ is. ■

Above, we see that the surjective mapping τ between the variable values of two causal models is not enough to preserve causal relations. In order for actual causality to be preserved, we also need that τ is compatible with $\tau_{\mathcal{U}}$ (Definition 5). This allows us to move from reasoning at the higher-level to the lower-level and back.

Looking back at the baking grid-world example from Section 2, we see that in the abstracted model from Example 3, in the causal setting where the agents bring the flour to the oven, but not the eggs, $\neg Eggs$ is a but-for cause of $\neg C$. We have that $\tau^{-1}_E(Eggs = \bot) = \{\langle M^1_1 = North, M^1_2 = North\rangle, \langle M^1_1 = East, M^1_2 = North\rangle, \langle M^1_1 = East, M^1_2 = East\rangle\}$, and that $\tau^{-1}_C(C = \bot) = \{C = \bot\}$. $\boldsymbol{u}_L = \langle U_{M^1_1} = East, U_{M^1_2} = North, U_{M^2_1} = East, U_{M^2_2} = North\rangle$ is such that $\tau_{\mathcal{U}}(\boldsymbol{u}_L) = \boldsymbol{u}_H$. In this setting, we have that $\langle M^1_1 = East, M^1_2 = North\rangle$ is indeed in the pre-image of $Eggs = \bot$.

This result shows that constructive abstractions of causal models make sense, as it is guaranteed that any cause (according to the modified HP definition) in the higher-level corresponds to a cause in the lower level. There are no new causes created in the higher-level.

5. Conclusion and Discussion

In this work, we provided two arguments for the use of constructive abstractions for causal models, as introduced in Beckers and Halpern (2019). For the first argument, we used the translation to causal CGS proposed in Kerkhove et al. (2025), by showing that constructive abstractions of causal models give equivalent results to the abstraction of causal CGS that we propose based on earlier ab-

stractions of multi-agent systems in the literature, when applying the translation to causal CGS. For the second argument, we showed that actual causality is preserved through the abstraction process.

Some limits of this work are that we only consider the HP-definition of causality. As we said, there are other definitions of actual causality present in the literature. However, none are as well-known as the HP-definition. It is also important to note that Theorem 16 still holds if we instead consider the more restrictive but-for cause, because in that case the set $\boldsymbol{W}$ is simply empty. While most people agree that but-for is not enough for causality, there is consensus that but-for is necessary for actual causality (Halpern (2016)).

Another limitation is that we only consider deterministic models. We made this decision because the causal CGS has so far only been defined for deterministic causal models. There are also not yet any widely supported definitions of actual causality in probabilistic models. Nevertheless, it would be useful to see whether constructive abstraction could also be used to abstract probabilistic models, especially given that many practical causal models are probabilistic, as they have to account for noise in the measurements. It might be impossible to give exact abstractions in that case, but we could study whether certain abstraction errors as proposed in Beckers et al. (2020) still preserve actual causality.

Another direction for future work would be to investigate what else could be preserved in causal abstractions, or to compare the causal models and causal CGS even more. Galimullin et al. (2025) proposes a logic that allows updates of multi-agent models to do a form of counterfactual reasoning. We could use this to compare the effect of interventions in both causal models and the causal CGS.

We believe that the results in this work will support the use of constructive abstractions, thus allowing for causal models to become less complex and more suitable to human reasoning. One can imagine a causal model of all the trains on a country's rail system. Such a model could contain many variables, for when a train arrives at a platform, when it leaves, when it passes a certain point, etc. Because there are large causal chains in the model, it is complex to figure out the actual cause of delays on a certain portion of the tracks. An abstraction of this causal model could for example not look at individual trains, but look at delays at stations, grouping multiple trains together. This will still be informative, because a disruption at an earlier station can cause delays at a later station, but this abstracted model will be smaller and hence easier to reason with. This would also make the model more explainable, in the sense that one can explain the original complex and detailed model by abstracting over irrelevant features.

Acknowledgments

This publication is part of the CAUSES project (KIVI.2019.004) of the research programme Responsible Use of Artificial Intelligence which is financed by the Dutch Research Council (NWO) and ProRail.

References

Rajeev Alur, Thomas A Henzinger, and Orna Kupferman. Alternating-time temporal logic. *Journal of the ACM (JACM)*, 49(5):672–713, 2002.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685, July 2019. doi: 10.1609/aaai.v33i01.33012678.

Sander Beckers and Joost Vennekens. A principled approach to defining actual causation. *Synthese*, 195(2):835–862, February 2018. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-016-1247-1.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pages 606–615. PMLR, 2020.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

Rustam Galimullin, Maksim Gladyshev, Munyque Mittelmann, and Nima Motamed. Changing the rules of the game: Reasoning about dynamic phenomena in multi-agent systems. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 829–838, 2025.

Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, Dragan Doder, and Brian Logan. Dynamic causality. In *Proceedings of the 26th European Conference on Artificial Intelligence*, pages 867–874, 2023. doi: 10.3233/FAIA230355. URL https://ebooks.iospress.nl/volumearticle/64287.

Joseph Y Halpern. *Actual causality*. MIT Press, 2016.

Sylvia S. Kerkhove, Natasha Alechina, and Mehdi Dastani. Causes and strategies in multiagent systems. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1098–1106, 2025.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In Gal Elidan and Kristian Kersting, editors, *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI-17)*. AUAI Press, August 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf.

Giovanni Varricchione, Natasha Alechina, Mehdi Dastani, and Brian Logan. Synthesising reward machines for cooperative multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 85, 2026. To be published.

Fabio Massimo Zennaro. Abstraction between structural causal models: A review of definitions and properties. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

Appendix A. Proof of Lemma 12

Proof We use the notation $\boldsymbol{X} \cdot \boldsymbol{Y}$ to denote the concatenation of sequences. So $\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L = \boldsymbol{X}_L \cdot \boldsymbol{Y}_L \leftarrow \boldsymbol{x}_L \cdot \boldsymbol{y}_L$, as $\boldsymbol{X}_L$ and $\boldsymbol{Y}_L$ are disjunct. We have that $\omega_\tau(\boldsymbol{X}_L \cdot \boldsymbol{Y}_L \leftarrow \boldsymbol{x}_L \cdot \boldsymbol{y}_L) = \boldsymbol{Z} \leftarrow \boldsymbol{z}$ if there exist $\boldsymbol{Z}, \boldsymbol{z}$ such that $\tau(Rst(\mathcal{V}_L, \boldsymbol{x}_L \cdot \boldsymbol{y}_L)) = Rst(\mathcal{V}_H, \boldsymbol{z})$. It holds that $\tau(Rst(\mathcal{V}_L, \boldsymbol{x}_L)) = Rst(\mathcal{V}_H, \boldsymbol{x}_H)$, since we have $\omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L) = \boldsymbol{X}_H \leftarrow \boldsymbol{x}_H$. To see that $\boldsymbol{z}$ must contain $\boldsymbol{x}_H$, let $\boldsymbol{v}_H$ be any element of $Rst(\mathcal{V}_H, \boldsymbol{z})$. Without loss of generality, as the order of the variables does not matter, we can write $\boldsymbol{v}_H$ as $\boldsymbol{z} \cdot \boldsymbol{z}_H$. Since $\boldsymbol{v}_H \in Rst(\mathcal{V}_H, \boldsymbol{z})$ there is some $\boldsymbol{v}_L \in Rst(\mathcal{V}_L, \boldsymbol{x}_L \cdot \boldsymbol{y}_L)$ such that $\tau(\boldsymbol{v}_L) = \boldsymbol{v}_H$, which, as the abstraction is constructive, can be

written as $\tau_1(\boldsymbol{v}_1) \cdot \tau_2(\boldsymbol{v}_2) \cdot ... \cdot \tau(\boldsymbol{v}_n)$, where $\tau_1(\boldsymbol{v}_1) \cdot ... \tau_m(\boldsymbol{v}_m) = \boldsymbol{z}$. Since $\boldsymbol{v}_L \in Rst(\mathcal{V}_L, \boldsymbol{x}_L \cdot \boldsymbol{y}_L)$, we can write $\boldsymbol{v}_L$ as $\boldsymbol{x}_L \cdot \boldsymbol{y}_L \cdot \boldsymbol{r}_L$. We have that $\boldsymbol{v}_1 \cdot ... \boldsymbol{v}_m = \boldsymbol{x}_L \cdot \boldsymbol{y}_L$, because no matter $\boldsymbol{r}_L$, $\tau(\boldsymbol{x}_L \cdot \boldsymbol{y}_L \cdot \boldsymbol{r}_L) = \boldsymbol{z} \cdot \boldsymbol{z}_H$ (for some arbitrary $\boldsymbol{z}_H$), so $\boldsymbol{z}$ must be determined by $\boldsymbol{x}_L \cdot \boldsymbol{y}_L$. So, $\boldsymbol{z} = \tau_1(\boldsymbol{x}_1) \cdot ... \cdot \tau_k(\boldsymbol{x}_k) \cdot \tau_{k+1}(\boldsymbol{y}_1) \cdot ... \cdot \tau_m(\boldsymbol{y}_{m-k})$. We have that $\tau(Rst(\mathcal{V}_L, \boldsymbol{x}_L)) = \boldsymbol{x}_H$, so by the same argument as above, $\tau_1(\boldsymbol{x}_1) \cdot ... \cdot \tau_k(\boldsymbol{x}_k) = \boldsymbol{x}_H$. Hence $\boldsymbol{z} = \boldsymbol{x}_H \cdot \boldsymbol{y}_H$, which means that $\omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L) = \boldsymbol{X}_H \leftarrow \boldsymbol{x}_H, \boldsymbol{Y}_H \leftarrow \boldsymbol{y}_H$.

■

Appendix B. Proof of Lemma 13

In order to prove Lemma 13, we need a lemma stated in Kerkhove et al. (2025), which states that the value of a variable does not change in states whose first index is higher than the agent rank of the variable.

Lemma 17 *Let GS be a causal CGS generated by the causal model $\mathcal{M}$. For any endogenous causal variable $X \in \mathcal{V}$ of $\mathcal{M}$, with $\rho(X) = i$, it holds that $(X = x) \in \pi(q_{i,j})$ for some state $q_{i,j}$ of GS, if and only if $(X = x) \in \pi(q_{i',j'})$ for all states $q_{i',j'}$ that are descendants of $q_{i,j}$.*

Now, using this, we can prove Lemma 13:

Proof [Lemma 13] We are going to prove the correspondence, i.e. $(X = x) \in \pi(q_{n,m}) \Leftrightarrow (\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \mathbf{u}) \models X = x$ by induction on the agent rank of X.

Base Step: If $\rho(X) = 0$, $X \in V_e$ and does not depend on any other endogenous variables, so if $(X = x) \in \pi(q_{n,m})$, Lemma 17 implies that $(\mathcal{M}, \mathbf{u}) \models X = x$. Because X does not depend on any agent variables, it will keep the same value when intervening on agent variables, so $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \mathbf{u}) \models X = x$ as well. For the other way around, if $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$, by the same reasoning we have that $(\mathcal{M}, \boldsymbol{u}) \models X = x$ and hence $(X = x) \in \pi(q_{0,0})$, again by Lemma 17 $(X = x) \in \pi(q_{n,m})$ as well.

Induction Hypothesis: Suppose that for all $X \in \mathcal{V}$ s.t. $\rho(X) \leq i$, $(X = x) \in \pi(q_{n,m})$ if and only if $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$.

Inductive Step: Let X be such that $\rho(X) = i + 1$. First suppose that $X \in V_a$.

- If $X \in \boldsymbol{X}$, it got its value $x \in \boldsymbol{x}$ from the intervention $\boldsymbol{X} \leftarrow \boldsymbol{x}$. So $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$. Moreover $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}}, \boldsymbol{u}) \models X = x$ as well, and hence by the premise of the lemma, $(X = x) \in \pi(q_{k,l})$. When we now use Lemma 17 again, we get that $(X = x) \in \pi(q_{n,m})$ as well. Hence $(X = x) \in \pi(q_{n,m})$ if and only if $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$.
- If $X \in \boldsymbol{Y}$, we have that it got its value $x \in \mathbf{y}$ in $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$, from the intervention $\boldsymbol{Y} \leftarrow \boldsymbol{y}$. $(X = x)$ will also be in $\pi(q_{n,m})$, because following the strategy $F_{\boldsymbol{Y} = \boldsymbol{y}}$ will lead to X choosing an action leading to the value $x \in \boldsymbol{y}$ to X, in a state $q_{k',l'}$, on the path between $q_{k,l}$ and $q_{n,m}$, where $k < k' \leq n$. Once again by Lemma 17, we have that $(X = x) \in \pi(q_{n,m})$ as well.
- If $X \notin \boldsymbol{X} \cup \boldsymbol{Y}$, let $(X = x) \in \pi(q_{n,m})$, the value x was determined by $F_{\mathcal{M}}$, in particular, $(\mathcal{M}, \boldsymbol{u}) \models [\boldsymbol{Z} \leftarrow \boldsymbol{z}]X = x$, where $\boldsymbol{Z} = \{Z \mid Z \in V_a, \rho(Z) < \rho(X)\}$ and $\boldsymbol{z} = \{z \mid z \in \alpha[q_{i,j}]\}$, where $q_{i,j}$ is the state on the path to $q_{n,m}$ where X got to take an action. By the inductive hypothesis we know that $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models \boldsymbol{Z} = \boldsymbol{z}$, and hence $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$ as well, because all variables X depends on have the same values in $\pi(q_{n,m})$ as in $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u})$. On the other hand, if $(\mathcal{M}^{\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models X = x$, we know that x is determined only by variables with a lower agent rank, by the inductive hypothesis all those, with their values, are in $\pi(q_{n,m})$. The

value of X in $\pi(q_{n,m})$ is determined by $F_{\mathcal{M}}$, so $(\mathcal{M}, \boldsymbol{u}) \models [\boldsymbol{Z} \leftarrow \boldsymbol{z}]X = x'$. All variable-value pairs Z, z are the variable-value pairs from $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \boldsymbol{u})$, so x' must be x, as all variables of lower agent rank have the same value.
Now suppose $X \in V_e$, and let $(X = x) \in \pi(q_{n,m})$. X depends only on variables of a lower level, specifically all agent variables of agent rank less than or equal to $i+1$. By the above and the inductive hypothesis, we know that all those variables have the same value in $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \boldsymbol{u})$, as in $\pi(q_{n,m})$, hence $X = x$ must also be induced by $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \boldsymbol{u})$, since there are no other interventions done after the agent variables of rank $i+1$ got their final value. Now suppose $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \boldsymbol{u}) \models X = x$, X depends only on variables of lower levels, all of those agent variables have the same value in $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \boldsymbol{u}) \models X = x$ as in $\pi(q_{n,m})$ by the inductive hypothesis. Hence $(X = x) \in \pi(q_{n,m})$ as well, since the environment variables follow the causal model in both $(\mathcal{M}^{\boldsymbol{X}\leftarrow\boldsymbol{x},\boldsymbol{Y}\leftarrow\boldsymbol{y}}, \boldsymbol{u})$ as in $\pi(q_{n,m})$. ■

Appendix C. Proof of Theorem 14

Proof We need to show that CGS_H is an abstraction of CGS_L, so we must show that there is a surjective map t_τ from Π_L to Π_H that satisfies Definition 11. Now, note that $\Pi_L = \{X = x \mid X \in \mathcal{V}_L, x \in \mathcal{R}(\mathcal{V}_L)\}$ and that $\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H)$, so we can define $t_\tau : \Pi_L \to \Pi_H$ to be s.t. $t_\tau(X_1 = x_1, ..., X_n = x_n) = (Y_1 = y_1, ..., Y_m = y_m)$, where $(y_1, ..., y_m) = \tau(x_1, ..., x_n)$.

We will now show that this t_τ satisfies Definition 11. Let $s \xrightarrow{\boldsymbol{Y}\leftarrow\boldsymbol{y}} s'$ be any transition in CGS_H, and let q be any state of CGS_L abstracted by s (i.e. $\pi_H(s) = t_\tau(\pi_L(q))$).[3] We now need to show that there is a sequence of joint actions α in CGS_L such that $q \xrightarrow{\alpha} q'$, where q' is abstracted by s'.[4]

Now, because CGS_H is a causal CGS, we have that state s corresponds to a causal setting $\mathcal{M}_H(\boldsymbol{u}, (\boldsymbol{X} \leftarrow \boldsymbol{x}))$ for some intervention $\boldsymbol{X} \leftarrow \boldsymbol{x}$, and that state s' corresponds to $\mathcal{M}_H(\boldsymbol{u}, (\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{Y} \leftarrow \boldsymbol{y}))$. We know that $\mathcal{M}_H$ is a constructive τ-abstraction of $\mathcal{M}_L$ (Definition 6), so there has to be a surjective $\tau_{\mathcal{U}}$ compatible with τ. We can hence write $\mathcal{M}_H(\boldsymbol{u}, (\boldsymbol{X} \leftarrow \boldsymbol{x}))$ as $\mathcal{M}_H(\tau_{\mathcal{U}}(\boldsymbol{u}_L), \omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L)) = \tau(\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}_L \leftarrow \boldsymbol{x}_L))$, where $\tau_{\mathcal{U}}(\boldsymbol{u}_L) = \boldsymbol{u}$ and $\omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L) = \boldsymbol{X} \leftarrow \boldsymbol{x}$ (this last statement is possible because for a constructive abstraction, we have that $\mathcal{I}_H = \omega_\tau(\mathcal{I}_L)$, so every $\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L$ has at least one corresponding low-level intervention).

So, to recap, s corresponds to $\tau(\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}_L \leftarrow \boldsymbol{x}_L))$. Note that $(Y = y) \in \pi_H(s)$ iff $\tau(\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}_L \leftarrow \boldsymbol{x}_L)) \models Y = y$. We also have that as s abstracts q, $\pi_H(s) = t_\tau(\pi_L(q))$. Because q is a state in a causal CGS, CGS_L, it must correspond to some causal setting $\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L)$ for some intervention $\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L$. Since $t_\tau(\pi_L(q)) = \pi_H(s)$, we have by the definition of t_τ that $\tau(\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L)) = \tau(\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}_L \leftarrow \boldsymbol{x}_L))$, because s corresponds to this latter causal setting, and s abstracts q.

By an argument similar to the one above, we have that s' corresponds to $\mathcal{M}_H(\tau_{\mathcal{U}}(\boldsymbol{u}_L), \omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L))$. Define α to be the sequence of joint actions that follows from following the strategy profile $F_{\boldsymbol{Y}_L=\boldsymbol{y}_L} \circ F_{\mathcal{M}_L}$ starting at the lowest-level state in CGS_L corresponding to $\mathcal{M}_L(\boldsymbol{u}_L, \boldsymbol{X}_L \leftarrow \boldsymbol{x}_L)$. Since we start in the lowest-level state corresponding to this causal setting, α will contain the actions that perform the interventions $\boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L$ and the actions $\boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L$, with $\boldsymbol{Z}_L \cap \boldsymbol{Y}_L = \emptyset$, where the variables in $\boldsymbol{Z}_L$ are agent variables with an agent rank higher

3. We denote a transition in the causal CGS CGS_H as $s \xrightarrow{\boldsymbol{Y}\leftarrow\boldsymbol{y}} s'$, as transitions in a causal CGS correspond to interventions, we make this explicit with this notation.

4. Note that here we use the notation $q \xrightarrow{\alpha} q'$ to denote a sequence of transitions, rather than just one.

than $\boldsymbol{X}_L$ and at most the rank of the highest ranked variable in $\boldsymbol{Y}_L$, and where $\boldsymbol{z}_L$ is such that $(\mathcal{M}_L, \boldsymbol{u}_L) \models [\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L]\boldsymbol{Z}_L = \boldsymbol{z}_L$. $\boldsymbol{Z}_L$ are hence those variables whose values are determined according to the causal strategy profile. This will execute all actions in $\boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L$ while following the causal model $\mathcal{M}_L$ for any other actions that need to be taken.

From Lemma 13 it follows that applying this action profile from an arbitrary state q abstracted by s will lead to a state q' corresponding to $\mathcal{M}_L(\boldsymbol{u}_L, (\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L))$. Because the abstraction is strong, we get that $\tau(\mathcal{M}_L(\boldsymbol{u}_L, (\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L)))$ equals $\mathcal{M}_H(\tau_{\mathcal{U}}(\boldsymbol{u}_L), \omega_\tau(\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L))$. By Lemma 12, and because s abstracts q (so $\omega_\tau(\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L) = \omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L)$ as we saw above), we have that $\omega_\tau(\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L) = \omega_\tau(\boldsymbol{X}'_L \leftarrow \boldsymbol{x}'_L) \wedge \omega_\tau(\boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L) \wedge \omega_\tau(\boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L) = \omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L) \wedge \omega_\tau(\boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L) \wedge \omega_\tau(\boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L) = \omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L))$.

Therefore, a state abstracting q' should correspond to $\tau(\mathcal{M}_L(\boldsymbol{u}_L, (\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L)))$. However as we recall, $\boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L$ is exactly the intervention on $\boldsymbol{Z}_L$ that does not change the variable values from the values that the variables got from the interventions on the other variables. Hence $\mathcal{M}_L(\boldsymbol{u}_L, (\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L, \boldsymbol{Z}_L \leftarrow \boldsymbol{z}_L)) = \mathcal{M}_L(\boldsymbol{u}_L, (\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L))$. Therefore, the state abstracting q' should correspond to $\tau(\mathcal{M}_L(\boldsymbol{u}_L, (\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L))) = \mathcal{M}_H(\tau_{\mathcal{U}}(\boldsymbol{u}_L), \omega_\tau(\boldsymbol{X}_L \leftarrow \boldsymbol{x}_L, \boldsymbol{Y}_L \leftarrow \boldsymbol{y}_L))$. This is exactly what s' corresponds to. Hence, q' is abstracted by s', so we have shown that a sequence α of joint actions exists for any transition $s \xrightarrow{\sigma} s'$ such that for any state q abstracted by s, $q \xrightarrow{\alpha} q'$, where q' is abstracted by s'. Hence CGS_H is an abstraction of CGS_L. ■

Appendix D. Full Proof of Theorem 16

Proof In the main text, we have shown that if $\boldsymbol{X} = \boldsymbol{x}$ is a modified HP cause of $Y = y$ in $(\mathcal{M}_H, \mathbf{u}_H)$, then, for all $\boldsymbol{u}_L$ such that $\tau_{\mathcal{U}}(\mathbf{u_L}) = \mathbf{u}_H$, there is a $\boldsymbol{Z} = \boldsymbol{z}$, a subset of an element in $\tau^{-1}_{\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x})$, that is a modified HP cause of $\tau_Y^{-1}(y)$ in $(\mathcal{M}_L, \mathcal{I}_L, \tau_{\mathcal{U}}(\mathbf{u_L}))$. Now, we are going to show that it also holds for a general boolean combination of events φ instead of just $Y = y$. In order to do that, we are going to show it for conjunction, disjunction and negation.

First conjunction, assume that it holds for the formulas φ and ψ in $(\mathcal{M}_H, \mathbf{u}_H)$. Suppose that $\boldsymbol{X} = \boldsymbol{x}$ is a cause of $\varphi \wedge \psi$ in $(\mathcal{M}_H, \mathbf{u}_H)$. We follow the same argument as in the main text, with the difference that now instead of $\boldsymbol{Y}_L = \boldsymbol{y}_L$ we get $\varphi_L \wedge \psi_L = \tau^{-1}_{\varphi \wedge \psi}(\varphi \wedge \psi)$. We have that not both φ and ψ are entailed by $\tau(\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \boldsymbol{u}_L)$. Hence, $(\mathcal{M}_L, \mathcal{I}_L, \mathbf{u}_L) \models [\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L]\neg(\varphi_L \wedge \psi_L)$. This satisfies AC2 of Definition 2, so just like with the proof for $Y = y$, we have that either $\boldsymbol{Z} = \boldsymbol{z}$ is a cause, or a subset is.

Now for disjunction, we again assume that it holds for the formulas φ and ψ in $(\mathcal{M}_H, \mathbf{u}_H)$. The difference with the conjunction case is that neither φ nor ψ are entailed by $\tau(\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \boldsymbol{u}_L)$. So neither of their preimages can be entailed by $\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \boldsymbol{u}_L))$. Therefore $(\mathcal{M}_L, \mathcal{I}_L, \mathbf{u}_L) \models [\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L]\neg(\varphi_L \vee \psi_L)$.

Finally the negation case, say that $\boldsymbol{X} = \boldsymbol{x}$ is a cause of $\neg\varphi$. Now we get that φ is entailed by $\tau(\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \boldsymbol{u}_L))$. Therefore, the preimage of $\neg\varphi$ must be entailed by $\mathcal{M}_L((\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L), \boldsymbol{u}_L)$. Hence, $(\mathcal{M}_L, \mathcal{I}_L, \mathbf{u}_L) \models [\boldsymbol{Z} \leftarrow \boldsymbol{z}', \boldsymbol{W}_L \leftarrow \boldsymbol{w}^*_L]\, \varphi_L$.

Since any Boolean combination of events φ can be written as a combination of these connectives, we have shown that it holds for any Boolean combination of events φ. ■