The dynamic logic of causality:

from counterfactual dependence to causal intervention

Alexandru Baltag
(ILLC, University of Amsterdam)

Causality, Interventions and Counterfactuals

Pearl's **causal models** have become the standard/dominant approach to *representing and reasoning about causality*.

The setting is based on the *'static' notion* of **causal graphs**, but it also makes an essential use of the *'dynamic'* notion of **causal interventions**.

In particular, Halpern and Pearl used this setting to define and investigate various notions of **actual causality**.

As noted by many, causal interventions have an **obvious 'counterfactual' flavor**.

But... their **relationship with the counterfactual conditionals** (*à la* Lewis-Stalnaker) has remained murky! A lot of confusion surrounds this topic, affecting even the crucial contributions of (Galles & Pearl 1998), (Halpern 2000, 2019), (Briggs 2012), (Halpern 2013), (Zhang 2013) etc.

This Talk

The purpose of this talk is threefold:

1. understand interventions as **dynamic modalities** (*rather than conditionals*);
2. elucidate the **relationship between dynamic intervention modalities and counterfactual conditionals**;
3. formalize and completely axiomatize a **Causal Intervention Calculus** (**CIC**), that is general enough to give a causal meaning to arbitrary Lewis conditionals, but also expressive enough to capture the various notions of actual causality proposed in the literature.

0. Preliminaries: Causal Signatures

Signature: a structure $\mathcal{S} = (\mathcal{V}, V = \bigcup_n V_n, F)$, where:

- $\mathcal{V} = \{X_0, X_1, \ldots, X_N\}$ is a finite set of **causal variables**, some being designated as *exogenous*, and the others as *endogenous*.
  For simplicity, I assume **only one exogenous variable**: X_0.
  Causal variables are 'empirical' variables, having *intrinsic meaning*, e.g. denoting time, velocity, temperature, etc (unlike FOL variables, which are simple placeholders).
- $V = \bigcup_n V_n$ is a set of **(multi-sorted) logical variables** (which are variables in the sense of FOL: simple placeholders); here, for each $n \geq 0$, $V_n = \{x_n, y_n, \ldots\}$ is a (finite or countably infinite) set of *logical variables of type/sort* X_n;
- F is a map associating to each endogenous variable X_n (with $n \geq 1$) some function symbol F_n of arity $N = |\mathcal{V}| - 1$.

NOTE: The standard notion of causal signature in the literature includes *only causal variables*. The set V of logical variables is my addition, allowing us to denote (and quantify over) *values*.

Causal Models

Model (over a signature $\mathcal{S} = (\mathcal{V}, V, F)$):
a structure $M = (\mathcal{R}, \mathcal{F})$, consisting of:

- a map $\mathcal{R}$ assigning to each causal variable $X_n \in \mathcal{V}$ some *non-empty range of values* $\mathcal{R}_n$;
- a structural map $\mathcal{F}$, assigning to each endogenous variable $X_n \in Var$ (with $n \geq 1$) some 'structural' causal function $\mathcal{F}_n : \mathcal{R}_0 \times \cdots \mathcal{R}_{n-1} \times R_{n+1} \times \cdots \times \mathcal{R}_N \to \mathcal{R}_n$ of arity $N = |\mathcal{V}| - 1$.

A **context** u is an assignment of values to the exogenous variable(s): in my simplified presentation, this is just a value $u \in \mathcal{R}_{X^0}$.

An **assignment** is a map α, assigning to each logical variable x_n of type X_n some value $\alpha(x_n) \in \mathcal{R}_n$.

A **state** is a pair $s = (u, \alpha)$ of a context u and an assignment α.

Local dependence (direct and indirect)

Given a context $u \in \mathcal{R}_0$, the **(local) direct dependence relation at** u is a relation $X_n \rightsquigarrow_u X_m$ (("X_n directlu affects X_m in context u ") between variables X_n and X_m, defined for $n > 0$ by:

$X_n \rightsquigarrow_u X_m$ iff $m \neq 0, n$ and $\forall i \neq 0, n, m \, \exists x_i \in \mathcal{R}_i \, \exists x_n, x'_n \in \mathcal{R}_n$ s.t.

$$\mathcal{F}_m(u, x_1, \ldots, x_n, \ldots, x_N) \neq \mathcal{F}_m(u, x_1, \ldots, x'_n, \ldots, x_N);$$

while for $n = 0$, we set

$$X_0 \rightsquigarrow_u X_m \quad \text{iff} \quad m > 0 \text{ and } \forall n > 0 (X_n \not\rightsquigarrow_u X_m).$$

The **(local) indirect dependence** relation $X \preceq_u Y$ ("X affects Y in context u ") is the *reflexive-transitive closure* of the relation $X \rightsquigarrow_u Y$:
$X \preceq_u Y$ says that there exists a finite chain of variables that start with X and end with Y, s.t. each variable directly affects the next one.

Recursive models

A causal model M is **recursive** if, for every context $u \in \mathcal{R}_0$, the *indirect dependence relation* $\preceq_u$ *is a partial order* (-or equivalently, the *direct dependence relation is acyclic*).

NOTE: If M is recursive, then each relation $\preceq$ structures the set of variables $\mathcal{V}$ into a *rooted tree*, having the exovariable X_0 as its (unique) root.

Most of the literature on causality (including all known logical axiomatizations of interventions) assume that the causal model is recursive.

So **from now I will restrict the setting to recursive models**.

1. Interventions as dynamic transformations

Given a causal model $M = (\mathcal{R}, \mathcal{F})$, an endogenous variable $X_n \in Var - \{X_0\}$ and a value $a \in \mathcal{R}_n$, the *result of the intervention* $X_n \leftarrow a$ on the model M is the model

$$M^{X_n \leftarrow a} = (\mathcal{R}, \mathcal{F}^{X_n \leftarrow a}),$$

obtained by putting:

- $\mathcal{F}_n^{X_n \leftarrow a} := a$ (the constant function that maps every value $b \in \mathcal{R}_n$ into a), and
- $\mathcal{F}_m^{X_n \leftarrow a} := \mathcal{F}_m$ in rest (for all $m \neq n$).

Multi-variable Interventions

For a tuple $\vec{X} = (X_{i_1}, \ldots, X_{i_k})$ of endogenous variables and a corresponding tuple of values $\vec{a} = (a_1, \ldots, a_k) \in \mathcal{R}_{i_1} \times \cdots \mathcal{R}_{i_k}$, we can similarly define the *result of the intervention* $\vec{X} \leftarrow \vec{a}$ on the model $M = (\mathcal{R}, \mathcal{F})$ to be the model

$$M^{\vec{X} \leftarrow \vec{a}} \;=\; (\mathcal{R}, \mathcal{F}^{\vec{X} \leftarrow \vec{a}}),$$

obtained by putting:

- $\mathcal{F}_{i_n}^{\vec{X} \leftarrow \vec{a}} := a_n$, and
- $\mathcal{F}_{m}^{\vec{X} \leftarrow \vec{a}} := \mathcal{F}_m$, for all $m \notin \{i_1, \ldots, i_k\}$.

Reduction to repeated single-variable interventions

It is easy to see that multi-variable interventions are reducible to compositions of single-variable interventions:
given $\vec{X} = (X_{i_1}, \ldots, X_{i_k})$ and $\vec{a} = (a_1, \ldots, a_k) \in \mathcal{R}_{i_1} \times \cdots \mathcal{R}_{i_k}$, we have

$$M^{\vec{X} \leftarrow \vec{a}} = ((M^{X_{i_1} \leftarrow a_1}) \ldots)^{X_{i_k} \leftarrow a_k}.$$

This is why in our logical language we will only consider as primitive the dynamic modalities $[X \leftarrow x]$ for single-variable interventions (and defined the others as abbreviations).

Intervention Operators

Judea Pearl first proposed **intervention operators**, which were then formalized in (Galles & Pearl 1998) and axiomatized in (Halpern 2000), then generalized in (Briggs 2012), (Halpern 2013), (Zhang 2013), etc.
For formulas φ (in some formal language), states $s = (u, \alpha)$ in a recursive model M, strings $\vec{X}$ of endogenous variables and matching strings $\vec{a}$ of values, we put:

$$s \models_M [\vec{X} \leftarrow \vec{a}]\varphi \text{ iff } s \models_{M^{\vec{X} \leftarrow \vec{a}}} \varphi$$

Most of these authors take this to be a kind of **causal conditional** $(\vec{X} = \vec{a}) \Box\!\!\rightarrow \varphi$ ("if the values of $\vec{X}$ were $\vec{a}$, then φ would hold") and compare it with **counterfactual conditionals** in the Lewis-Stalnaker tradition. But... not all their properties match!

However, Pearl and Halpern were aware of the analogy with dynamic modalities, as witnessed by their notation $[\vec{X} \leftarrow \vec{a}]$.

Dynamic operators

Dynamic modalities $[\alpha]\varphi$ ("after action α is performed, φ will be true) occur in formalisms such as PDL (Propositional Dynamic Logic), PAL (Public Announcement Logic), DEL (Dynamic Epistemic Logic), QDL (Quantum Dynamic Logic) etc.

Their standard semantics is the usual Kripke-relational semantics: given a set W of "possible worlds" or 'states', endowed with an accessibility relation R_α (the "transition relation"), the operator $[\alpha]$ is the Kripke modality for R_α; i.e.

$$w \models [\alpha]\varphi \text{ iff } \forall w' \, (wR_\alpha w' \Rightarrow w' \models \varphi)$$

In fact, PAL and DEL use a variation of this semantics in which the accessibility relations R_α are **between models**, or more precisely go from a world w in the current model M to a world w' in an **updated model** M'.
At an abstract level, these can still be subsumed under the general pattern of a Kripke modality, if we take as our "worlds" to be the **model-world pairs** (M, w).

Test modalities

In the particular case of 'test' actions $\alpha :=?!\psi$ (that in some sense correspond to 'testing', 'observing' or 'announcing' a proposition ψ), the corresponding relations R_ψ will be indexed by propositions.

EXAMPLES: PDL "test" $[?\psi]\varphi$ (in programming), public announcement operator $[!\psi]\varphi$, private announcement $[!\psi_a]\varphi]$ to some agent a, quantum tests $[?_q\psi]\varphi$, etc.

Interventions as dynamic modalities

To package intervention operators $[\alpha]\varphi$ as dynamic modalities, we have to think of our "worlds" as **pairs** $w = (M, s)$ consisting of a causal model M and a state $s = (u, \alpha)$ for this model.

Furthermore, we introduce accessibility relations $R_{\vec{X} \leftarrow \vec{A}}$ between "worlds", defined by:

$$(M, s) R_{\vec{X} \leftarrow \vec{a}} (M', s') \text{ iff } M' = M^{\vec{X} \leftarrow \vec{a}} \text{ and } s = s'.$$

Then the semantics of intervention operators becomes a special case of the standard Kripke semantics for dynamic modalities: in fact, a special case of the above-mentioned propositionally-indexed dynamic modalities $[?!\psi]\varphi$, with $\psi := (\vec{X} = \vec{a})$.

Nevertheless, this does **NOT** show that these are *counterfactuals* of the form $(\vec{X} = \vec{a}) \Box\!\!\rightarrow \varphi$.

The static formula $\vec{X} = \vec{a}$ is one of the **effects (results)** of the intervention $\vec{X} \leftarrow \vec{a}$; but the intervention is **not** reduceable to this effect. The "world" $(M^{\vec{X} \leftarrow \vec{a}}, s)$ is not necessarily the closest world to (M, s) that satisfies the effect $\vec{X} = \vec{a}$.

Interpretation

In general, it's best to think of interventions, not necessarily as "actual" actions (that may happen **in** a given world or model), but as counterfactual actions: *abstract manipulations of the causal structure*, that happen **to** the model (world).

However, for practical applications, it is *essential that sometimes such interventions can also be actually performed* in the world (e.g. controlled experiments).

2. Relationship with counterfactual conditionals

There is a persistent confusion in the literature between intervention modalities and counterfactual conditionals.

Pearl and others claimed that intervention modalities constitute a new, better semantics for counterfactual conditionals, one that is 'structural' or 'causal', i.e. based on the causal structure of the world, rather than on comparative similarity. Others claimed that the structural counterfactuals are a special case of Lewis counterfactuals, in which comparative similarity is based on causality.

I will argue that interventions can *be used* to define a causal-based notion of counterfactual conditionals, but that these are not exactly the same as the intervention modalities. Still, the two are interdefinable.

Moreover, the notion of counterfactual conditional obtained in this way falls well within the frame of Lewis' account!

RECALL: Stalnaker conditionals

Stalnaker's semantics for conditionals assumes given a **selection function** $f(w, \psi)$:
for every possible world $w \in W$ and every proposition (expressed by some sentence) ψ, the function f selects some world $f(w, \psi) \in W$, interpreted as "**the closest ψ-world to w**".

The selection function f is required to satisfy a number of **semantic conditions**, including CENTERING:

$$\text{if } w \models \psi, \text{ then } f(w, \psi) = w.$$

Stalnaker's semantical clause for the conditional $\psi \mathrel{\square\!\!\rightarrow} \varphi$ is

$$w \models \psi \mathrel{\square\!\!\rightarrow} \varphi \text{ iff } f(w, \psi) \models \varphi.$$

RECALL: Lewis conditionals

Lewis generalized this, by *dropping the uniqueness of the "closest world"*:
the selection functions f now associates a (possibly empty) **set of worlds** $f(w, \varphi) \subseteq W$ to every world w and proposition φ.

Once again, a number of conditions are required for Lewis conditionals. In particular, CENTERING comes now in a strong version ("STRONG CENTERING")

$$\text{if } w \models \varphi, \text{ then } f(w, \varphi) = \{w\},$$

and a weak version ("WEAK CENTERING")

$$\text{if } w \models \varphi, \text{ then } w \in f(w, \varphi).$$

Lewis' semantical clause for the conditional $\psi \mathrel{\Box\!\!\rightarrow} \varphi$ is

$$w \models \psi \mathrel{\Box\!\!\rightarrow} \varphi \text{ iff } w' \models \varphi \text{ for every } w' \in f(w, \psi).$$

Equivalent presentation: comparative similarity relations

An equivalent presentation of Lewis conditionals uses **comparative similarity** relations between worlds $w, w', w'' \in W$:

$$w' \preceq_w w''$$

means that w' **is 'at least as close to** w **as** w''**"** (i.e. w' is *at least as similar* to w as w'').

Once again, there are a number of requirements: comparative similarity is *reflexive*, *transitive*, *total* (i.e. every two worlds are comparable: either $w' \preceq_w w''$ or $w'' \preceq_w w'$), and satisfies WEAK CENTERING

$$w \leq_w w'.$$

The selection function $f(w, \psi)$ can be taken to be the set of closest ψ-worlds:

$$f(w, \psi) \ := \ \{w' \in W : w' \models \psi \,\&\, \forall w''(w'' \models \varphi \Rightarrow w' \preceq_w w'')\}$$

Every selection function f satisfying Lewis' conditions can be represented in this way.

Conditionals as 'test' modalities

Lewis conditionals $\psi \Box\!\!\rightarrow \varphi$ can be considered as a **very special case of test modalities** $[?!\psi]\varphi$:

by taking the proposition-indexed relation R_ψ, given by

$$w R_\psi w' \;\text{ iff }\; w' \in f(w, \psi),$$

and applying the above-mentioned Kripke/relational semantics for test modalities, we obtain the Lewis/Stalnaker conditional as a special case!

Going backwards: dynamic modalities as "generalized conditionals"?

In a trivial sense, the Kripke semantics of **any** test modality $[?!\psi]\varphi$ can be considered as a 'wild generalization' of conditionals: if we take as our Stalnaker-Lewis selection function to be

$$f(w, \psi) := \{w' : wR_{\psi}w'\},$$

then the relational/dynamic semantics becomes "the *same*" as the selection-function semantics.

Arbitrary dynamic modalities $[\alpha]\varphi$ can be considered as a further, even "wilder", generalization.

NOTE though that these 'selection functions' **will not necessarily satisfy any of the Stalnaker-Lewis conditions**.
So this repackaging of dynamic Kripke semantics does NOT show that all dynamic modalities are counterfactual conditionals!

Pearl on causal counterfactuals

The confusion between interventions and causal counterfactuals originates from Pearl himself (Pearl 1998, 2000) and was clearly spelled out in (Galles and Pearl 1998).
They read $[\vec{X} \leftarrow \vec{a}]\psi$ as some kind of counterfactual conditional $(\vec{X} = \vec{a}) \boxrightarrow \psi$, and claim that all Lewis' axioms hold for this interpretation.

But... *this is an artifact of their restrictions on the language*: both they and (Halpern 2000) use a very restricted syntax, consisting essentially of *propositional logic* based on atoms of the form

$$X = a$$

(where X is any causal variable and a is any value), as well as on intervention operators of the form

$$[\vec{X} \leftarrow \vec{a}]\varphi,$$

where φ *is itself a conjunction of formulas of the form* $X = a$.

Pearl's comparative similarity

Galles and Pearl simply *ignore the Lewis axioms that cannot be even expressed in their language* (since they involve disjunctions). Moreover, the Lewis axiom of Modus Ponens for Counterfactuals *only* holds for conjunctions of formulas of the form $X = a$!

To compare with Lewis' account, Galles and Pearl take the **"worlds" to be causal-variable assignments**, and they define the *distance* $d(w, w')$ as the **minimal number of local interventions needed for transforming** w **into** w'.
This gives rise to a natural notion of *comparative similarity*: world w' is at least as similar to w as a world w'' iff $d(w, w') \leq d(w, w'')$.

They seem to believe that intervention modalities can be thought of as Lewis conditionals based on this notion of similarity (but they are not clear on this, so I am not sure they actually believe this).

Pearl reiterates the same claims (more explicitly) in his book on Causality.

Briggs and Zhang

Briggs 2012 starts by noting that:

> *"A number of recent authors (Galles and Pearl 1998; Hiddleston 2005; Halpern 2000) advocate a causal modeling semantics for counterfactuals. But the precise logical significance of the causal modeling semantics remains murky. Particularly important, yet particularly under-explored, is its relationship to the similarity-based semantics for counterfactuals developed by Lewis (Counterfactuals 1973)."*

Then Briggs... proceeds to add to this murkiness, by treating intervention modalities and Lewis counterfactuals as is they are two different versions of the same thing!

Briggs' "common framework"

Specifically, Briggs reformulates the intervention semantics into a selection-based unified setting (for both Lewis' similarity semantics and causal modelling): Brigg's so-called "common framework", based on selection functions (not necessarily satisfying the Lewis-Stalnaker conditions).

That seems pretty obvious and rather trivial: as we already saw (and it was known for a long time), the Kripke-style semantics of **any** formula-indexed relational modality $[\varphi?!?]\psi$ based on some underlying relation R_φ (and in particular any dynamic modality, even the kind that involves changes of model: public announcements, upgrades, quantum test etc) can be reformulated in terms of "set-valued selection functions" $f(w, \varphi)$ (-albeit ones that do not necessarily satisfy any of the requirements that are specific to conditionals).

Further complications

To avoid having set-valued selection functions, Briggs adopts Fine's truthmaker semantics (in which sets of states are thought of as "worlds"): but in this case, this move just complicates the semantics further, for no good reasons and no clear benefits!

By this trick, Briggs can maintain Pearl's reading of $[\vec{X} \leftarrow \vec{a}]\psi$ as some kind of counterfactual conditional $(\vec{X} = \vec{a})\Box\!\!\rightarrow \psi$ (albeit of a very non-standard kind), without relying on Pearl's proposal of intervention-based comparative similarity.

But, of course, this is just *wrong*: not every formula-indexed Kripke modality gives is a "conditional" (not even speaking of a counterfactual conditional), but only those satisfying the Lewis-Stalnaker conditions (or closely related ones).

No Weak Centering: failure of Modus Ponens

All these authors have problems generalizing this semantics from the very restricted format of Galles and Pearl to the arbitrary format of Lewis conditionals $\psi \Box\!\!\rightarrow \varphi$.

Briggs has a generalization, but only when the premise ψ contains no instances of the conditional operator $\Box\!\!\rightarrow$: so no nested 'conditionals'.

In particular, Weak Centering fails in this system. Briggs derives very radical conclusions from this generalization:

> *"The extended language is unlike any logic of Lewis's: modus ponens is invalid (for counterfactuals–my note), and classical logical equivalents cannot be freely substituted in the antecedents of conditionals."*

I think this simply shows that Briggs' notion of "causal counterfactuals" is *not* the right one: **intervention modalities are NOT the same as causal counterfactuals**.

Causal counterfactuals versus intervention modalities

I will argue that the confusion arises from the identification of "worlds" with causal-variable assignments ('states'), while interventions are in fact *model-changing operations*.

So, to faithfully compare causal modelling with Lewis semantics, one needs to identify the "worlds" with the models themselves, or rather with the pairs $(model, state)$.

It is then very easy to introduce natural notions of comparative similarity relations between causal models, that satisfy all Lewis conditions (including Weak Centering, in fact even Strong Centering).

Various similarity relations

In fact, there are several different such notions, which implement different views on how "miraculous" is a given intervention.

One of these is essentially the one proposed by Galles and Pearl (counting the *number of variables* that are being intervened upon), except that it has to be lifted to the level of models.

Another natural proposal involves counting the *number of causal links* that are being disconnected by a given intervention: intervening upon a variable X that corresponds in the causal graph to a node with many parents is a "bigger" intervention that one acting only upon a variable Y with very few parents.

Finally, we might also count the *number of value-changes*: i.e. the number of variables whose values are different between the two assignments; but in this case, we need to choose a way to compare the relative size of value-changes and causal changes.

All the resulting causal conditionals will automatically satisfy all the axioms of Lewis' counterfactual logic.

The Difference: causality disruption

But recall that an intervention $\vec{X} \leftarrow \vec{a}$ changes the "world" (model) in **three ways**:

1. the *value of* $\vec{X}$ *is 'fixed'* to $\vec{a}$ (i.e. $\vec{X} = \vec{a}$ becomes true);
2. the variables in $\vec{X}$ *become independent* from all the others (i.e. their backwards-causal connections are disrupted);
3. as in a Stalnaker-Lewis conditional, this is in some sense *the minimal change that satisfies (1) and (2)*.

Hence, the intervention modalities **CANNOT** be identified with causal conditionals having the **same premise** (of the form $\vec{X} = \vec{a}$)!

The hidden premise: independence

Instead, intervention modalities can be recovered as conditionals **only if we add an independence requirement to the premise**, strengthening it to $\vec{X} = \vec{a} \wedge Indep(\vec{X})$.

The reason is precisely the fact that interventions do *not* just change/reassign values to variables, but also *disrupt causality*, erasing the causal dependence between the (intervened) variables and their parents.

This explains the apparent failure of Weak Centering in Briggs' setting: her formalization simply fails to incorporate the required independence in the premises of her counterfactuals.

Distance between causal models

We'll adopt the first notion of distance: counting the *number of variables* that are being intervened upon.

The **quasi-distance** $d(M, M')$ between two causal models M and M' is *the size of the minimal intervention converting* M *into* M' (i.e. the number of variables whose structural equations are different in M' from M).

This means that

$$d(M, M') \ := \ |\vec{Y}|,$$

if $|\vec{Y}|$ is the length of the minimal tuple of variables $\vec{Y}$ s.t. $M' = M^{\vec{Y} \leftarrow \vec{a}}$ for some tuple of values $\vec{a}$.
If no such tuple exists, then we have

$$d(M, M') \ := \ \infty$$

Note that the quasi-distance is indeed a quasi-metric rather than a metric (since in general symmetry does not hold, and in fact it doesn't make sense: M and M' can be causal submodels of each other only when $M = M'$).

Comparative similarity

But the quasi-distance still gives rise to a comparative similarity relation between "worlds" as $(model, state)$ pairs $w = (M, s)$:

$(M', s') \preceq_{(M,s)} (M'', s'')$ iff $d(M, M') \leq d(M, M'')$ and $s = s' = s''$.

This is really the same natural notion of "distance" proposed by Galles and Pearl, except that we consider it in its appropriate setting: as a measure of similarity between model-state pairs (not between variable assignments).
By unfolding in this setting the Lewis' definition of counterfactuals we obtain the following semantics, for a given model M, state $s = (u, \alpha)$ and formulas φ and ψ:

$$s \models_M \varphi \Box\!\!\rightarrow \psi \quad \text{iff} \quad \text{either } \forall \vec{X}, \vec{a} \left(s \not\models_{M^{\vec{X} \leftarrow \vec{a}}} \varphi \right) \text{ or}$$

$$\exists \vec{X}, \vec{a} \left(s \models_{M^{\vec{X} \leftarrow \vec{a}}} \varphi \text{ and } \forall \vec{Y}, \vec{b} \left(|\vec{Y}| \leq |\vec{X}| \text{ implies } s \models_{M^{\vec{Y} \leftarrow \vec{b}}} \varphi \Rightarrow \psi \right) \right)$$

Full Independence

For every tuple $\vec{X} = (X_{i_1}, \ldots, X_{i_k})$ of endovariables $X_i \in Var - \{X_0\}$, we use the abbreviation

$$Indep(\vec{X}) \ := \bigwedge_{Y \in Var - \{X_0\}} Y \not\rightarrow X_m$$

for the formula saying that every endovariable in $\vec{X}$ is *fully independent* (in the current context) from all other endovariables.

With these notions, we can now spell out the connection between dynamic intervention modality and causal-counterfactual conditionals.

The correct relationship

Proposition Given a causal submodel N of the model M, $N^{\vec{X}\leftarrow\vec{a}}$ is the (unique) closest submodel (most similar) to N satisfying

$$\vec{X} = \vec{a} \wedge Indep(\vec{X}).$$

Corollary Intervention modalities and counterfactual conditionals are mutually definable: every intervention-modality formula of the form $[\vec{X} \leftarrow \vec{a}]\psi$ is logically equivalent to the counterfactual formula

$$\left(\vec{X} = \vec{a} \wedge Indep(\vec{X})\right) \boxdot\!\!\rightarrow \psi.$$

Conversely, every counterfactual formula of the form $\varphi \boxdot\!\!\rightarrow \psi$ is logically equivalent to the intervention-modality formula

$$\left(\bigwedge_{\vec{X},\vec{a}} \neg[\vec{X} \leftarrow \vec{a}]\varphi\right) \vee \bigvee_{\vec{X},\vec{a}} \left([\vec{X} \leftarrow \vec{a}]\varphi \wedge \bigwedge_{\vec{Y},\vec{b}\, s.t.\, |\vec{Y}|\leq|\vec{X}|} [\vec{Y} \leftarrow \vec{b}](\varphi \Rightarrow \psi)\right).$$

3. Causal Intervention Calculus (CIC)

Previous authors (Galles & Pearl 1998), (Halpern 2000, 2019), (Briggs 2012) use languages that **refer directly to values**.

But this presents a number of disadvantages. Defining actual causality and other notions requires **quantifying** over all values (in the range of a causal variable). Since the above formalisms cannot do that, these authors restrict the models to have a **fixed finite range of values** for each variable, so they can simulate quantifiers by just taking conjunctions indexed by values.
This is a severe restriction: it excludes variables such as temperature, position, velocity etc, which take infinitely many (in fact, a continuum) of values.

I chose to explicitly **introduce logical variables** $x_n, y_n, \ldots$, as well FOL quantifiers $\forall x_n$ ranging over all values in $\mathcal{R}_n$.

In addition, I **internalize the structural equations in the syntax**, thus using **complex terms** $F(X_0, \ldots, X_N)$ (in addition to causal and logical variables).

Syntax of CIC: Causal Terms

Terms We simultaneously define, for every natural number $n \geq N$, a set T_n of **terms** t_n **of sort/type** X_n. The definition is by simultaneous recursion, as follows:
for the exogenous type(s) $n = 0$, we simply put

$$t_0 \ ::= \ X_0 \,|\, x_0,$$

where X_0 is our (unique) exogenous variable, and $x_0 \in V_0$ are logical variables of type X_0;
while for $n \geq 1$, we recursively put

$$t_n \ ::= \ X_n \,|\, x_n \,|\, F_n(t_0, \ldots, t_{n-1}, t_{n+1}, \ldots, t_N),$$

where $X_n \in \mathcal{V}$ is the corresponding endogenous variable, $x_n \in V_n$ are logical variables of type X_n, $N := |\mathcal{V}| - 1$, and for each $i \neq n$, t_i is an already formed term of sort X_i.

The **set of all causal terms** is $T := \bigcup_{0 \leq n \leq N} T_n$.

SYNTAX OF CIC: Causal Formulas

The set Φ of *formulas* is given recursively by

$$\varphi \quad ::= \quad t = t' \mid \neg\varphi \mid \varphi \wedge \varphi \mid [X_n \leftarrow t_n]\varphi \mid \forall x_n.\varphi$$

where $t, t' \in T$, $x \in V$, $X_n \in \mathcal{V}$ and $t_n \in T_n$.

Free and bound (occurrences of) variables (in a formula): as usual in (multi-sorted) FOL.

Sentences: closed formulas (all variables have only bounded occurrences).

Constraint: in the construct $[X_n \leftarrow t_n]\varphi$, every variable occurring in t_n must have only free occurrences in φ.

Abbreviations

We have the usual FOL abbreviations: disjunction
$\varphi \vee \psi := \neg(\neg\varphi \wedge \neg\psi)$, $\varphi \Rightarrow \psi := \neg\varphi \vee \psi$,
$\varphi \Leftrightarrow \psi := (\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$, $\exists x_n.\varphi := \neg\forall x_n.\neg\varphi$.

In addition, we can also can define *multi-variable intervention modalities* $[\vec{X} \leftarrow \vec{x}]\varphi$, for any tuple $\vec{X} = (X_{i_1}, \ldots, X_{i_k})$ of endogenous variables and any corresponding tuple $\vec{x} = (x_{i_1}, \ldots, x_{i_k})$ of values of appropriate sorts.

The definition is by recursion on k, using the clause:

$$[\vec{X}, X_{i_{k+1}} \leftarrow \vec{x}, x_{i_{k+1}}]\varphi := [\vec{X} \leftarrow \vec{x}][X_{i_{k+1}} \leftarrow x_{i_{k+1}}]\varphi$$

Semantics: term interpretation

Given a causal model $M = (\mathcal{R}, \mathcal{F})$ for a signature $\mathcal{S} = (\mathcal{V}, V, F)$, we will *evaluate formulas at states* $s = (u, \alpha)$, but note that the assignment α is irrelevant for the evaluation of closed formulas. So *propositions can in fact be evaluated at contexts* $u \in \mathcal{R}_0$.

For this, we first define a **term interpretation map** $s(t)_M$, that extends the assignment α to all terms: given a state $s = (u, \alpha)$ and a term t, we define the *interpretation* $s(t)_M$ *of term* t *in state* s *of model* M. When the model is understood, we skip the subscript M and simply write $s(t)$. The definition is by recursion:

$$s(X_0) = u, \qquad s(x_n) = \alpha(x_n),$$

$$s(X_n) = \mathcal{F}_n(s(X_0), \ldots, s(X_{n-1}), s(X_{n+1}), \ldots s(X_N)),$$

$$s(F_n(t_0, \ldots, t_N)) = \mathcal{F}_n(s(t_0), \ldots, s(X_{n-1}), s(X_{n+1}), \ldots s(t_N)).$$

Semantics: satisfaction (truth)

Second, we define the **satisfaction relation** $s \models_M \varphi$ between states $s = (u, \alpha)$ and formulas φ, by the recursive clauses:

$$s \models_M t = t' \;\text{ iff }\; s(t)_M = s(t')_M$$

$$s \models_M \neg\varphi \;\text{ iff }\; s \not\models_M \varphi$$

$$s \models_M \varphi \wedge \psi \;\text{ iff }\; s \models_M \varphi \text{ and } s \models_M \psi$$

$$s \models_M [X_n \leftarrow t_n]\varphi \;\text{ iff }\; s \models_{M^{X_n \leftarrow s(t_n)_M}} \varphi$$

$$s \models_M \forall x_n.\varphi \;\text{ iff }\; s^{x_n \leftarrow a_n} \models_M \varphi \text{ for all } a_n \in \mathcal{R}_n,$$

where $s^{x_n \leftarrow a_n} := (u, \alpha^{x_n \leftarrow a_n})$ is given by putting $\alpha^{x_n \leftarrow a_n}(x_n) := a_n$ and $\alpha^{x_n \leftarrow a_n}(x_m) := \alpha(x_m)$ for all $m \neq n$.

Once again, when the model is understood, we skip the subscript M and simply write $s \models \varphi$.

Abbreviation: direct dependence

We can express *local* direct dependence in our syntax, as a formula $X_n \rightsquigarrow X_m$, definable as follows:

- for $n, m \neq 0, n \neq m$, we set $X_n \rightsquigarrow X_m$ as an abbreviation for $\exists x_1, \ldots, x_{m-1}, x_{m+1}, \ldots x_N \, \exists x'_n \, F_m(x_0, \ldots, x_n, \ldots, x_{N-1}) \neq F_m(x_0, \ldots, x'_n, \ldots, x_{N-1})$;
- while for $n = 0, m > 0$, we set

$$X_0 \rightsquigarrow X_m \text{ iff } \bigwedge_{0<k<|Var|} \neg(X_k \rightsquigarrow X_m);$$

- and in all other cases, we put

$$X_n \rightsquigarrow X_m := false.$$

We can now easily check that this abbreviation fits the semantics of direct dependence. For all contexts u and assignments α:

$$(u, \alpha) \models X \rightsquigarrow Y \text{ iff } X \rightsquigarrow_u Y.$$

Abbreviation: dynamic/counterfactual value

For any term $t \in T$, causal variable X_n and term t_n of sort X_n, we introduce the notation $[X_n \leftarrow t_n]t$ as an abbreviation for another term, defined by the recursive clauses:

$$[X_n \leftarrow t_n]x := x, \qquad [X_n \leftarrow t_n]X_0 := X_0, \qquad [X_n \leftarrow t_n]X_n := t_n,$$

$$[X_n \leftarrow t_n]X_m := F_m([X_n \leftarrow t_n]X_0, \ldots, [X_n \leftarrow t_n]X_N),$$

$$[X_n \leftarrow t_n]F_n(t_0, \ldots, t_{n-1}, t_{n+1}, \ldots, t_N) := t_n,$$

$$[X_n \leftarrow t_n]F_m(t_0, \ldots, t_N) := F_m([X_n \leftarrow t_n]t_0, \ldots [X_n \leftarrow t_n]t_N),$$

where $m \neq 0, n$ and $N = |\mathcal{V}| - 1$.

Intuitively, $[X_n \leftarrow t_n]t$ denotes the **value that term t would take after the intervention** $[X_n \leftarrow t_n]t$.

The Proof System CIC

- all the axioms and rules of multi-sorted FOL, with equality as its *only* predicate and functional symbols F_n;
- the Structural Equation axiom: for all $n \geq 1$, we have

$$X_n = F_n(X_0, \ldots, X_{n-1}, X_{n+1}, \ldots, X_N).$$

- Necessitation rule for intervention modalities: from φ, infer $[X \leftarrow t]\varphi$;
- The following Recursion/reduction Axioms:

$$[X \leftarrow t](t' = t'') \Leftrightarrow ([X \leftarrow t]t' = [X \leftarrow t]t'')$$

$$[X \leftarrow t]\neg\varphi \Leftrightarrow \neg[X \leftarrow t]\varphi$$

$$[X \leftarrow t](\varphi \wedge \psi) \Leftrightarrow ([X \leftarrow t]\varphi \wedge [X \leftarrow t]\psi)$$

$$[X \leftarrow t]\forall x.\varphi \Leftrightarrow \forall x.[X \leftarrow t]\varphi$$

  provided that in the last axiom x does not occur in t.
- Recursiveness:

$$(X_{i_0} \rightsquigarrow X_{i_1} \rightsquigarrow \ldots \rightsquigarrow X_{i_k}) \Rightarrow \neg(X_{i_k} \rightsquigarrow X_{i_0}).$$

Soundness, Completeness, Expressivity, Complexity

Proposition
The system $\mathbf{CIC}$ is **sound and complete** with respect to the class of recursive models.

In addition, **intervention modalities are eliminable**: *every formula is provable equivalent to one that belongs to the 'static' fragment* (without any intervention modalities).

Finally, **the satisfiability problem for the static fragment is equivalent to the satisfiability problem for multi-sorted FOL** with equality as the only predicate and with functional symbols, and thus it is **undecidable**.

Theorems, relation to previous calculi

All axioms of Halpern's causal calculus can be proven as theorems in the above calculi (in full generality, not just in the case when we are given finite bounds on the size of the ranges $\mathcal{R}_X$, as stated in Halpern's calculus).

In particular, the most interesting and the most complex of Halpern's axioms ("Composition") can be proven using the Normality of Interventions (i.e. Necessitation Rule and the Reduction axiom for negation), together with the following theorem:

$$\text{("Trivial Intervention")} \qquad \vec{X} = \vec{x} \Rightarrow Y = [\vec{X} \leftarrow \vec{x}]Y.$$

In the calculus CIC, this can be proven for all variables Y_n by induction on n (while in $CICR$ the proof is more complex, and proceeds by cases, considering each possible causal graph and doing induction on the structure of the graph).

Causality

Halpern and Pearl proposed over the years *three different definitions of causality* (which Halpern calls "the original HP definition", "the updated HP definition" and "the modified HP definition"), plus *another, simpler notion* ("but-for cause") lying at the intersection of all three.

All four are expressible in our language, as many other such proposals by other authors.

I will *only encode here the "modified HP definition"* (for which Halpern expresses a clear preference in his book, since it seems able to cope better with a wide range of examples), *as well as the special case of "but-for cause"*, which is of independent conceptual interest.

Fix and $Reset$ operators

To easier understand the encodings, it is useful to first define two derived operators, that describe two specific types of interventions.

$Fix_{\vec{X}}$ denotes the type of intervention (on the variables in the string $\vec{X}$) by which the variables $\vec{X}$ are *fixed to their current values* (while being causally disconnected from their parents);

$Reset_{\vec{X}}$ denotes the type of intervention by which the variables $\vec{X}$ are *reassigned some arbitrary values* (while being causally disconnected from their parents).

For both actions, we can consider the associated *existential and universal modalities* (which are de Morgan duals as usual, hence interdefinable), but the existential versions seems to be more useful for applications to causality.

Fix and $Reset$: formal definitions

The Fix operator is given by

$$\langle Fix_{\vec{X}} \rangle \varphi \;:=\; [\vec{X} \leftarrow \vec{X}]\varphi,$$

or equivalently

$$\langle Fix_{\vec{X}} \rangle \varphi \;:=\; \exists \vec{x} \, (\vec{X} = \vec{x} \wedge [\vec{X} \leftarrow \vec{x}]\varphi).$$

The $Reset$ operator is given by

$$\langle Reset_{\vec{X}} \rangle \varphi \;:=\; \exists \vec{x} \, [\vec{X} \leftarrow \vec{x}]\varphi,$$

where $\vec{X} = \vec{x}$ is an abbreviation for $\bigwedge_i X_i = x_i$.

Note that these operations are only defined for strings $\vec{X}$ of *endovariables*.

Encoding causality: "but-for cause"

With these notations, we put

$$c_{\vec{X}}\varphi \;:=\; \varphi \wedge \langle Reset_{\vec{X}} \rangle \neg \varphi.$$

We read $c_{\vec{X}}\varphi$ as "(the current value of) $\vec{X}$ is a *but-for cause* of φ (in the current context)".

Halpern captures this formally in his language, but only in propositional form and only for specific values, by always making the values explicit: i.e., instead of $c_{\vec{X}}\varphi$ he only defines the expression "$\vec{X} = \vec{a}$ is a but-for cause of φ ".

Encoding causality: actual causality

Finally, we can also define **actual causality** $C_{\vec{X}}\varphi$, to be read as "(the current value of) $\vec{X}$ is an *actual cause* of φ (in the current context)":

$$C_{\vec{X}}\varphi \;:=\; \varphi \wedge \bigvee_{\vec{W} \text{ s.t. } \vec{W}\cap\vec{X}=\emptyset} \langle Fix_{\vec{W}}\rangle\langle Reset_{\vec{X}}\rangle\neg\varphi \wedge$$

$$\wedge \bigwedge_{\vec{Y}\subset\vec{X},\vec{W} \text{ s.t. } \vec{W}\cap\vec{Y}=\emptyset} [Fix_{\vec{W}}][Reset_{\vec{Y}}]\varphi$$

This is Halpern and Pearl's "*modified HP definition*" of actual causality. The third conjunct is a minimality condition, which ensures that only those variables in $B\vec{X}$ whose values are essential for φ are considered part of a cause.
Once again, Halpern's formalization captures this formally in his language, but only in propositional form and for specific values: i.e., instead of $C_{\vec{X}}\varphi$ he defines only "$\vec{X} = \vec{x}$ is a cause of φ ", etc.

Causal relations between variables

Although Halpern only applies *causal notions to formulas* (thus implicitly treating them as propositional operators $c_X\varphi$ and $C_X\varphi$), he is in fact also interested in *causal connections between variables*.

This leads us to the following definitions:

$$c_{\vec{X}}Y \;:=\; \exists y\, c_{\vec{X}}(Y = y),$$

saying that "(the current value of) $\vec{X}$ is a *but-for cause* of (the current value of) Y"; and

$$C_{\vec{X}}Y \;:=\; \exists y\, C_{\vec{X}}(Y = y),$$

saying that "(the current value of) $\vec{X}$ is an *actual cause* of (the current value of) Y".

Again, Halpern's language can express these forms of causal dependence between variables *only indirectly and for specific values*: e.g., instead of $C_{\vec{X}}Y$, he writes "$\vec{X} = \vec{x}$ is a cause of $Y = y$".

Comments

It is important to point that the causal relations $c_X Y$ and $C_X Y$ are *not* equivalent to the dependence relation $X \rightsquigarrow Y$ and $X \preceq Y$.

Also, note that Halpern's notion of causality is *reflexive* (i.e. $C_{\vec{X}} X_i$ whenever X_i occurs in the string $\vec{X}$), but *non-transitive*.