Reading Notes of *Actual Causality*

Gao Fangshu

2019.4.27

Chapter 1. Introduction and Overview

There are two notions of caulsality:

- *type causality*: Also called *general causality*. Type causality contains general staments, and allows people to make predictions (*forward-looking*).
- *actual causality*: Also called *token causality* or *specific causality*. Actual causality focus on particular events, and related to words such as "responsebility" and "blame".

Roughly speaking, reasoning about type causality is equivalent to reasoning about *effects of causes* (possible effects of a given event), whereas reasoning about actual causality is equivalent to reasoning about *causes of effects* (possible causes of a particular outcome).

"But-for" definition of causality: A is a cause of B if, but for A, B would not have happened. However, it is not always enough to determine causality, and *Halpern-Pearl definition* solve some problems where but-for test fails.

Chapter 2. The HP Definition of Causality

2.1 Causal Models

Definition of causal model

A *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$:

- $\mathcal{S}$: A *signature*, explicitly lists the endogenous and exogenous variables and characterizes their possible values. A signature $\mathcal{S}$ is a tuple $(\mathcal{U},\mathcal{V},\mathcal{R})$:
  - $\mathcal{U}$: A set of exogenous variables
  - $\mathcal{V}$: A set of endogenous variables
  - $\mathcal{R}$: $\mathcal{R}$ maps variables in $\mathcal{U}$ or $\mathcal{V}$ into possible values for them (i.e., the set of values over which the variable ranges).
- $\mathcal{F}$: A set of *structural equations*. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X maps $\times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Z)$ to $\mathcal{R}(X)$. That means, function F_X captures a relation between all varaibles but for X and X. F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$.

The key role of the structural equations is that they allow us to determine what happens if things had been other than they were, perhaps due to an external intervention. In many examples, there is general agreement as to the appropriate causal model. The structural equations do not express actual causality; rather, they express the effects of interventions or, more generally, of variables taking on values other than their actual values.

There may be uncertainty of about the effects of interventions. All this uncertainty can be described by putting a probability on causal models and on the values of the exogenous variables. We can then talk about the probability that A is a cause of B.

The dependencies between variables in a causal model M can be described using a *causal network* (sometimes called a *causal graph*), consisting of nodes and directed edges. Informally, a model is said to be *recursive* (or *acyclic*) if there are no such dependency cycles (sometimes called *feedback cycles*) among the variables. A *strongly recursive* model is a model that if we know values of the exogenous variables, we know values of all variables else. The values of all variables are determined given a *context*.

Definition of recursive model

First, we define *partial order* $\preceq$: $X \preceq Y$ denotes that X affects Y. The fact that $\preceq$ is a partial order means that $\preceq$ is a *reflexive*, *anti-symmetric*, and *transitive* relation:

- reflexive: $\forall X, X \preceq X$
- anti-symmetric: $X \preceq Y, Y \preceq X \Rightarrow X = Y$
- transitive: $X \preceq Y, Y \preceq Z \Rightarrow X \preceq Z$

Second, we define *independency*: Y is independent of X in $(M, \vec{u})$ if, for all settings $\vec{z}$ of the endogenous variables other than X and Y, and all values x and x' of X, $F_Y(x, \vec{z}, \vec{u}) = F_Y(x', \vec{z}, \vec{u})$. That is, if we control all other variables, changing the value of X cannot affect value of Y.

Then we define *recursive model*: A model M is recursive if, for each context (setting of the exogenous variables) $\vec{u}$, there is a partial order $\preceq_{\vec{u}}$ of the endogenous variables such that unless $X \preceq_{\vec{u}} Y$, Y is independent of X in $(M, \vec{u})$.

If M is a *strongly recursive* model, then we can assume that all the partial orders $\preceq_{\vec{u}}$ are the same; in a recursive model, they may differ.

The choice of model can have a significant impact in determining causality ascriptions. A may be a cause of B relative to $(M, \vec{u})$ and not a cause of B relative to $(M', \vec{u}')$.

2.2 A Formal Definition of Actual Cause

We associate a causal formula φ with a set of context . $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$. A *causal formula* (over $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$) is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where (1) φ is a Boolean combination of primitive events; (2) $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$; (3) $y_i \in \mathcal{R}(Y_i)$. In a causal model, a causal formula φ may be true or false.

For $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, let $\mathcal{L}(\mathcal{S})$ consist of all Boolean combinations of causal formulas, where the variables in the formulas are taken from $\mathcal{V}$ and the sets of possible values of these variables

are determined by $\mathcal{R}$.

Let $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the *causal setting* $(M, \vec{u})$. Given a model M, the model that describes the result of the intervention setting variables in $\vec{Y}$ to $\vec{y}$ is $M_{\vec{Y} \leftarrow \vec{y}}$. If ψ holds after the intervention, we have: $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\psi$ iff $(M_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models \psi$.

The notation $(M, \vec{u}) \models \varphi$ is standard in the logic and philosophy communities. In the example of forest fire, M^d is the disjunctive model. $\big(M^d, (1,1)\big) \models [MD \leftarrow 0](FF = 1)$, or $\big(M^d_{MD \leftarrow 0}, (1,1)\big) \models [L \leftarrow 0; MD \leftarrow 0](FF = 0)$, means that "even if the arsonist is somehow prevented from dropping the match, the forest burns".

Three versions of HP definition of causality

$\vec{X} = \vec{x}$ is an actual cause of φ in the causal setting $(M, \vec{u})$ if the following three conditions hold (three versions of HP definition are only different in AC2):

- AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$
- AC2. See below:
  - AC2(a). There is a partition of $\mathcal{V}$ (the set of endogenous variables) into two disjoint subsets $\vec{Z}$ and $\vec{W}$(so that $\vec{Z} \cap \vec{W} = \emptyset$) with $\vec{X} \subseteq \vec{Z}$ and a setting $\vec{x}'$ and $\vec{w}$ of the variables in $\vec{X}$ and $\vec{W}$, respectively, such that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$
  - AC2(b^o) (original). If $\vec{z}^*$ is such that $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} - \vec{X}$, we have $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\varphi$.
  - AC2(b^u) (updated). If $\vec{z}^*$ is such that $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{W}'$ of $\vec{W}$ and subsets $\vec{Z}'$ of $\vec{Z} - \vec{X}$, we have $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\varphi$
  - AC2(a^m) (modified). There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, then $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*]\neg\varphi$
- AC3. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies condition ACl and AC2, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of φ unless both $\vec{X} = \vec{x}$ and φ actually happen.

AC2(a) is a necessity condition. It says that for $X = x$ to be a cause of φ, there must be a value $x2$ in the range of X such that if X is set to x', φ no longer holds. This is the but-for clause; but for the fact that $X = x$ occurred, φ would not have occurred.

AC2(b^o) says that changing $\vec{Z}$ (may due to setting $\vec{W}$ to $\vec{w}$, or some variables in $\vec{Z}$ are forced to their original values) does not affect φ; φ continues to be true.

The only difference between AC2(b^o) and AC2(b^u) lies in the clause "for all subsets $\vec{W}'$ of $\vec{W}$": AC2(b^u) must hold even if only a subset of $\vec{W}'$ of the variables in $\vec{W}$ are set to their

values in $\vec{w}$. This means that the variables in $\vec{W} - \vec{W}'$ essentially act as they do in the real world; thar is, their values are determined by the structural equations, rather than being set to their values in $\vec{w}$.

AC3 is a minimality condition, which ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential are considered part of a cause; inessential elements are pruned.