Explanation: A(n Abridged!) Survey

Joe Halpern
Cornell University

Includes joint work with Judea Pearl (UCLA)

The Big Picture

Defining explanation is hard!

- People have been trying for millenia
- Lots of examples developed to shoot down the many attempts
  - just as with definitions of causality

The goal of this talk: to present a definition (based on ideas that Judea Pearl and I developed) that involves causality and knowledge, and to discuss *partial* explanations.

- **Basic idea:** an explanation is a fact that, if found to be true, would constitute an actual cause of the *explanandum* (the fact to be explained), regardless of the agent's initial uncertainty.

Explanation: An Abridged History

The classic definitions of explanation (going back to Hempel and Salmon) do not involve causality.

- Very roughly speaking, an explanation consists of some initial conditions from which the explanandum logically follows
- There were later statistical versions
- Van Fraassen and Gärdenfors: the explanation must raise the probability of the explanandum.
  - Problem: these definitions did not take causality into account
  - Example: The barometer falling rapidly is not an explanation of the storm approaching, even though finding it out raises the probability of a storm
    - The barometer falling is not a *cause* of the storm

Why Knowledge Matters

[Van Fraassen:] What counts as an explanation for one person might not count as an explanation for another.

Example: [Gärdenfors:] Suppose that we seek an explanation of why Mr. Johansson has been taken ill with lung cancer. Some possible explanations:

(a) he worked for years in asbestos manufacturing

(b) a causal model describing the connection between asbestos fibres and lung cancer.

If you know (a) and not (b), then (b) is a good explanation; if you know (b) and not (a), then (a) is a good explanation.

Causal models

A *causal model* is a tuple $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$:

- $\mathcal{U}$: set of exogenous variables
- $\mathcal{V}$: set of endogenous variables
- $\mathcal{F}$: set of structural equations (one for each $X \in \mathcal{V}$):
  - E.g., $X = Y \wedge Z$

Causal models

A *causal model* is a tuple $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$:

- $\mathcal{U}$: set of exogenous variables
- $\mathcal{V}$: set of endogenous variables
- $\mathcal{F}$: set of structural equations (one for each $X \in \mathcal{V}$):
  - E.g., $X = Y \wedge Z$

Variable X depends on variable Y if Y can affect the value of X:

- There is a setting of the other variables such that changing the value of Y changes the value of X (according to $\mathcal{F}$)

We focus on *acyclic* models, where the dependency relation is acyclic. Such models can be described using causal networks:

- Like Bayesian networks, except that instead of associating with each node X a conditional probability table, we associate with it the equation that shows how the value of X is determined by the value of its parents

Let $\vec{u}$ be a *context*: a setting of the exogenous variables:

- $(M, \vec{u}) \models Y = y$ if $Y = y$ is unique solution to equations in $\vec{u}$
  - Here we're assjing that the network is acyclic
- $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi$ if $(M_{\vec{X} \leftarrow \vec{x}}, \vec{u}) \models \varphi$.
  - $[\vec{X} \leftarrow \vec{x}]\varphi$ means "after intervening to set $\vec{X}$ to $\vec{x}$, φ holds"
  - $M_{\vec{X} \leftarrow \vec{x}}$ is the causal model after setting $\vec{X}$ to $\vec{x}$:
    - replace the original equations for the variables in $\vec{X}$ by $\vec{X} = \vec{x}$.

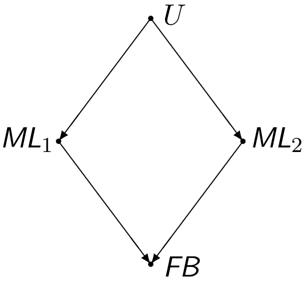
Example 1: Arsonists

Two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios.

1. Disjunctive scenario: either match by itself suffices to burn down the whole forest.
2. Conjunctive scenario: both matches are necessary to burn down the forest

Arsonist scenarios

Same causal network for both scenarios:



- endogenous variables ML_i, $i = 1, 2$:
  - $ML_i = 1$ iff arsonist i drops a match
- exogenous variable $U = u_{j_1 j_2}$
  - $j_i = 1$ iff arsonist i the background conditions are such that arsonist i will drop a match
- endogenous variable FB (forest burns down).
  - For the disjunctive scenario $FB = ML_1 \vee ML_2$
  - For the conjunctive scenario $FB = ML_1 \wedge ML_2$

Sufficient Cause: Definition

Pearl and I defined a notion of *actual causality*, but for explanation, we seem to need a stronger notion: *sufficient causality*

- $\vec{X} = \vec{x}$ is a sufficient cause of φ in $(M, \vec{u})$ if, not only does $\vec{X} = \vec{x}$ bring about φ in context $\vec{u}$, but in all contexts.

Sufficient Cause: Definition

Pearl and I defined a notion of *actual causality*, but for explanation, we seem to need a stronger notion: *sufficient causality*

- $\vec{X} = \vec{x}$ is a sufficient cause of φ in $(M, \vec{u})$ if, not only does $\vec{X} = \vec{x}$ bring about φ in context $\vec{u}$, but in all contexts.

Formally, $\vec{X} = \vec{x}$ is a *sufficient cause of* φ *in in* $(M, \vec{u})$ if

SC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$ (like AC1)

Sufficient Cause: Definition

Pearl and I defined a notion of *actual causality*, but for explanation, we seem to need a stronger notion: *sufficient causality*

- $\vec{X} = \vec{x}$ is a sufficient cause of φ in $(M, \vec{u})$ if, not only does $\vec{X} = \vec{x}$ bring about φ in context $\vec{u}$, but in all contexts.

Formally, $\vec{X} = \vec{x}$ is a *sufficient cause of* φ *in in* $(M, \vec{u})$ if

SC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$ (like AC1)

SC2. (Simplified version:) For some $\vec{x}' \neq \vec{x}$, variables $\vec{Y}$, and setting $\vec{y}$ of these variables, $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{Y} \leftarrow \vec{y}]\neg\varphi$

Sufficient Cause: Definition

Pearl and I defined a notion of *actual causality*, but for explanation, we seem to need a stronger notion: *sufficient causality*

- $\vec{X} = \vec{x}$ is a sufficient cause of φ in $(M, \vec{u})$ if, not only does $\vec{X} = \vec{x}$ bring about φ in context $\vec{u}$, but in all contexts.

Formally, $\vec{X} = \vec{x}$ is a *sufficient cause of* φ *in in* $(M, \vec{u})$ if

SC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$ (like AC1)

SC2. (Simplified version:) For some $\vec{x}' \neq \vec{x}$, variables $\vec{Y}$, and setting $\vec{y}$ of these variables,
$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{Y} \leftarrow \vec{y}]\neg\varphi$

SC3. $(M, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\varphi$ for all contexts $\vec{u}'$.

Sufficient Cause: Definition

Pearl and I defined a notion of *actual causality*, but for explanation, we seem to need a stronger notion: *sufficient causality*

- $\vec{X} = \vec{x}$ is a sufficient cause of φ in $(M, \vec{u})$ if, not only does $\vec{X} = \vec{x}$ bring about φ in context $\vec{u}$, but in all contexts.

Formally, $\vec{X} = \vec{x}$ is a *sufficient cause of* φ *in in* $(M, \vec{u})$ if

SC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$ (like AC1)

SC2. (Simplified version:) For some $\vec{x}' \neq \vec{x}$, variables $\vec{Y}$, and setting $\vec{y}$ of these variables, $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{Y} \leftarrow \vec{y}]\neg\varphi$

SC3. $(M, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\varphi$ for all contexts $\vec{u}'$.

SC4. $\vec{X}$ is minimal; there is no subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}|_{\vec{X}'}$ satisfies conditions SC1, SC2, and SC3

Sufficient Cause: Examples

- In the disjunctive forest fire example, both $ML_1 = 1$ and $ML_2 = 1$ are sufficient causes of the fire
- In the conjunctive forest fire example, $ML_1 = 1 \wedge ML_2 = 1$ is a sufficient cause of the fire

Explanation: The Basic Definition

The definition of explanation is relative to an agent's epistemic state. For now we assume that the causal model M is known.

- An agent's epistemic state is a set $\mathcal{K}$ of contexts with a probability $\Pr$ on them
- $\mathcal{K}$ is the set of contexts that the agent considers possible.

Explanation: The Basic Definition

The definition of explanation is relative to an agent's epistemic state. For now we assume that the causal model M is known.

- An agent's epistemic state is a set $\mathcal{K}$ of contexts with a probability $\Pr$ on them
- $\mathcal{K}$ is the set of contexts that the agent considers possible.

Definition: $\vec{X} = \vec{x}$ *is an explanation of* φ *relative to a set* $\mathcal{K}$ *of contexts* in causal model M if

EX1. $\vec{X} = \vec{x}$ is a sufficient cause of φ in all contexts in $\mathcal{K}$ satisfying $\vec{X} = \vec{x} \wedge \varphi$.

- We "condition" on what we know ($\vec{X} = \vec{x} \wedge \varphi$)
- We consider only contexts in $\mathcal{K}$ in SC3.

EX2. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x} \mid_{\vec{X}'}$ satisfies EX1.

EX3. There exists a context $\vec{u} \in \mathcal{K}$ such that $(M, \vec{u}) \models \vec{X} = \vec{x} \wedge \varphi$.

- The agent consider possible a context where the explanation holds.

Explanation: Examples

- In the disjunctive forest fire example, let u_{ij} be the context where $ML_1 = i$ and $ML_2 = j$.
  - relative to $\mathcal{K} = \{u_{00}, u_{01}, u_{10}, u_{11}\}$, both $ML_1 = 1$ and $ML_2 = 1$ explain the fire
  - relative to $\mathcal{K} = \{u_{00}, u_{10}\}$, $ML_1 = 1$ explains the fire, but $ML_2 = 1$ doesn't
    - EX3 fails: the agent knows that $ML_2 = 1$ doesn't happen
- In the conjunctive forest fire example,
  - $ML_1 = 1 \wedge ML_2 = 1$ is a sufficient cause of the fire relative to $\mathcal{K}$ if $u_{11} \in \mathcal{K}$
  - if $\mathcal{K} = \{u_{01}, u_{11}\}$, $ML_1 = 1$ is an explanation; $ML_1 = 1 \wedge ML_2 = 1$ is not (it violates minimality); $ML_2 = 1$ is not (it violates SC3: it's not a sufficient cause)

Partial Explanations

Not all explanations are equally good.

- There are many different "dimensions" of goodness: simplicity, generality informativeness, ...
  - I focus on three of them here

Partial Explanations

Not all explanations are equally good.

- There are many different "dimensions" of goodness: simplicity, generality informativeness, ...
  - I focus on three of them here

EX1 appeals to sufficient cause, and thus requires SC2 and SC3. for *all* contexts $\vec{u} \in \mathcal{K}$.

- But what if SC2/SC3 hold only for most contexts?

Partial Explanations

Not all explanations are equally good.

- There are many different "dimensions" of goodness: simplicity, generality informativeness, ...
  - I focus on three of them here

EX1 appeals to sufficient cause, and thus requires SC2 and SC3. for *all* contexts $\vec{u} \in \mathcal{K}$.

- But what if SC2/SC3 hold only for most contexts?

Here is where the probability $\Pr$ on $\mathcal{K}$ comes in.

- We can consider the probability that a claimed explanation satisfies SC2/SC3 (i.e., the probability of the set of contexts for which SC2/SC3 hold).

Definition: $\vec{X} = \vec{x}$ is a *partial explanation of* φ *with goodness* (α, β) *relative to* $(\mathcal{K}, \Pr)$ if the set of contexts where SC2 (resp., SC3) holds has probability at least α (resp., β).

Partial Explanations: Examples

Example: Victoria is tanned; I seek an explanation.

- The causal model includes the three variables "Victoria took a vacation in the Canary Islands", "sunny in the Canary Islands", and "went to a tanning salon"
- There are 8 contexts u_{ijk} assigning values (0 or 1) to each of these variables.
- Victoria going to the Canaries is not an explanation of Victoria's tan.
  - It doesn't satisfy SC3 (if it's not sunny, she won't get a tan even if she goes)
  - it may not satisfy SC2 if the reason she got a tan is that she went to the tanning salon

Nevertheless, most people would accept "Victoria took a vacation in the Canary Islands" as a satisfactory explanation of Victoria being tanned.

- It is a partial explanation with high α and β

Likelihood

We often prefer the more likely explanation:

Example: In the disjunctive forest-fire example, if $\mathcal{K} = \{u_{10}, u_{01}\}$ and I give u_{10} higher probability ($ML_1 = 1$ has higher probability than $ML_2 = 1$), then $ML_1 = 1$ is a better explanation of $FB = 1$.

Likelihood

We often prefer the more likely explanation:

Example: In the disjunctive forest-fire example, if $\mathcal{K} = \{u_{10}, u_{01}\}$ and I give u_{10} higher probability (*ML*$_1 = 1$ has higher probability than *ML*$_2 = 1$), then *ML*$_1 = 1$ is a better explanation of *FB* $= 1$.

But that's not the whole story either ...

Example: Suppose there's a fire in a lab; you suspect an arsonist. But one of the variables in the model is O: presence of oxygen.

- if $\mathcal{K}$ consists only of contexts where there is a fire, then $O = 1$ is a sufficient cause for the fire relative to $\mathcal{K}$, so is an explanation of the fire.
- Moreover, the probability that $O = 1$ is high (also conditional on there being a fire).
- But we don't view the presence of oxygen as a very good explanation of the fire.

Explanatory Power

Roughly speaking, we define the *explanatory power* of a partial explanation $\vec{X} = \vec{x}$ for φ relative to $(\mathcal{K}, \Pr)$ as $\Pr(\varphi \mid \vec{X} = \vec{x})$.

- $O = 1$ has low explanatory power for lab fires
  - It's unlikely for a lab fire to occur just because there is oxygen

Explanatory Power

Roughly speaking, we define the *explanatory power* of a partial explanation $\vec{X} = \vec{x}$ for φ relative to $(\mathcal{K}, \Pr)$ as $\Pr(\varphi \mid \vec{X} = \vec{x})$.

- $O = 1$ has low explanatory power for lab fires
  - It's unlikely for a lab fire to occur just because there is oxygen

But this rough definition is not quite right.

- It confounds correlation with causation
  - a falling barometer would have high explanatory power for rain

Explanatory Power

Roughly speaking, we define the *explanatory power* of a partial explanation $\vec{X} = \vec{x}$ for φ relative to $(\mathcal{K}, \Pr)$ as $\Pr(\varphi \mid \vec{X} = \vec{x})$.

- $O = 1$ has low explanatory power for lab fires
  - It's unlikely for a lab fire to occur just because there is oxygen

But this rough definition is not quite right.

- It confounds correlation with causation
  - a falling barometer would have high explanatory power for rain

So we define explanatory power of $\vec{X} = \vec{x}$ for φ relative to $(\mathcal{K}, \Pr)$ as the probability that $\vec{X} = \vec{x}$ is a cause of φ conditional on $\vec{X} = \vec{x}$.

Competing notions of goodness

There is a tension between these notions of goodness:

- goodness of partial explanation
- likelihood of explanation
- explanatory power of explanation

We may not be able to get an explanation that optimizes all three.

There is no obvious way to resolve the tension

- The modeler has to decide what is important.

Conclusion

Explanation is a slippery notion.

- This is not the first or second definition that I tried . . .
  - And it differs from the one in the original Halpern-Pearl paper since it focuses more on sufficient causes
- Since it's not clear how to prove a theorem saying "the definition is right", we must rely on examples to sharpen intuition.

Conclusion

Explanation is a slippery notion.

- This is not the first or second definition that I tried . . .
  - And it differs from the one in the original Halpern-Pearl paper since it focuses more on sufficient causes
- Since it's not clear how to prove a theorem saying "the definition is right", we must rely on examples to sharpen intuition.
- There are many notions of "goodness" for explanations, and a modeler needs to trade them off.

I've only scratched the surface here. For more details, see Chapter 7 in