Actual Causality

Sankaran Vaidyanathan

KDL Lab Meeting, July 27th 2021

What we will cover today

- Motivation and examples for AC
- The Halpern (2015) definition of AC
- Normality and Graded Causation
- Issues with defining AC
- Extra material: responsibility and blame

What is actual causality?

- Assignment of causal responsibility for some event that occurs, based on how events actually play out
  - *type causality*: smoking causes cancer
  - *token causality*: the fact that Willard smoked for 30 years caused him to get cancer

What is actual causality?

- Assignment of causal responsibility for some event that occurs, based on how events actually play out
  - *type causality*: smoking causes cancer
  - *token causality*: the fact that Willard smoked for 30 years caused him to get cancer
  - **It's true that it was pouring rain last night, and I was drunk, but the cause of the accident was the faulty brakes in the car (so I'm suing GM)**

What is actual causality?

- Assignment of causal responsibility for some event that occurs, based on how events actually play out
  - *type causality*: smoking causes cancer
  - *token causality*: the fact that Willard smoked for 30 years caused him to get cancer
  - **It's true that it was pouring rain last night, and I was drunk, but the cause of the accident was the faulty brakes in the car (so I'm suing GM)**
- Why should we care about token/actual causality?

What is actual causality?

- Assignment of causal responsibility for some event that occurs, based on how events actually play out
  - *type causality*: smoking causes cancer
  - *token causality*: the fact that Willard smoked for 30 years caused him to get cancer
  - **It's true that it was pouring rain last night, and I was drunk, but the cause of the accident was the faulty brakes in the car (so I'm suing GM)**
- Why should we care about token/actual causality?
  - Issues of actual causality are omnipresent in the law
  - Arguments about causality are increasingly being called upon within CS in problems of explanation, responsibility and blame attribution/credit assignment

Defining an actual cause

- One idea: use counterfactuals
  - A is a cause of B if it is the case that if A had not happened, B would not have happened
  - A can sometimes be referred to in the literature as a *but-for* cause
  - If the brakes hadn't been faulty, I wouldn't have had the accident

Defining an actual cause

- One idea: use counterfactuals
  - A is a cause of B if it is the case that if A had not happened, B would not have happened
  - A can sometimes be referred to in the literature as a *but-for* cause
  - If the brakes hadn't been faulty, I wouldn't have had the accident
- Pearl and Halpern came up with a definition of actual causality using structural equations to capture counterfactuals
  - Original: Halpern and Pearl (UAI 2001)
  - Halpern and Pearl (BJPS 2005): corrected after a counterexample from Hopkins and Pearl (2003)
  - Halpern (ICJAI 2015): what we will cover today

Are we done?

- The simple counterfactual definition doesn't always work

Are we done?

- The simple counterfactual definition doesn't always work
- An example problem: *preemption*. Example taken from Paul and Hall (2013)
  - Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's throw had not preempted it.

Are we done?

- The simple counterfactual definition doesn't always work
- An example problem: *preemption*. Example taken from Paul and Hall (2013)
  - Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's throw had not preempted it.
- Why is Suzy's throw the cause?
  - If Suzy hadn't thrown under the *contingency* that Billy didn't hit the bottle, then the bottle would have shattered.
  - This contingency where Billy didn't hit the bottle is what actually happened

Are we done?

- The simple counterfactual definition doesn't always work
- An example problem: *preemption*. Example taken from Paul and Hall (2013)
  - Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's throw had not preempted it.
- Why is Suzy's throw the cause?
  - If Suzy hadn't thrown under the *contingency* that Billy didn't hit the bottle, then the bottle would have shattered.
  - This contingency where Billy didn't hit the bottle is what actually happened
- Should Billy's throw also be a cause?
  - Not really, because his rock didn't actually hit the bottle
  - In general, we can use what actually happened to restrict the set of contingencies

An obligatory notation slide

- Basic syntax and semantics
  - Primitive event: setting of a single endogenous variable $X = x$
  - $[\vec{X} \leftarrow \vec{x}]\varphi$ means that after setting $\vec{X}$ to $\vec{x}$, φ holds
  - 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 V in $\mathcal{V}$

An obligatory notation slide

- Basic syntax and semantics
  - Primitive event: setting of a single endogenous variable $X = x$
  - $[\vec{X} \leftarrow \vec{x}]\varphi$ means that after setting $\vec{X}$ to $\vec{x}$, φ holds
  - 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 V in $\mathcal{V}$
- Let $\vec{u}$ be a context: a setting of the exogenous variables:

An obligatory notation slide

- Basic syntax and semantics
  - Primitive event: setting of a single endogenous variable $X = x$
  - $[\vec{X} \leftarrow \vec{x}]\varphi$ means that after setting $\vec{X}$ to $\vec{x}$, φ holds
  - 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 V in $\mathcal{V}$
- Let $\vec{u}$ be a context: a setting of the exogenous variables:
  - $(M, \vec{u}) \models Y = y$ if $Y = y$ is in the unique solution to equations when exogenous variables are set to $\vec{u}$

An obligatory notation slide

- Basic syntax and semantics
  - Primitive event: setting of a single endogenous variable $X = x$
  - $[\vec{X} \leftarrow \vec{x}]\varphi$ means that after setting $\vec{X}$ to $\vec{x}$, φ holds
  - 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 V in $\mathcal{V}$
- Let $\vec{u}$ be a context: a setting of the exogenous variables:
  - $(M, \vec{u}) \models Y = y$ if $Y = y$ is in the unique solution to equations when exogenous variables are set to $\vec{u}$
  - $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi$ if the Boolean predicate φ holds in the setting where the exogenous variables are set to $\vec{u}$ and endogenous variables $\vec{X}$ are set to $\vec{x}$

An obligatory notation slide

- Basic syntax and semantics
  - Primitive event: setting of a single endogenous variable $X = x$
  - $[\vec{X} \leftarrow \vec{x}]\varphi$ means that after setting $\vec{X}$ to $\vec{x}$, φ holds
  - 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 V in $\mathcal{V}$
- Let $\vec{u}$ be a context: a setting of the exogenous variables:
  - $(M, \vec{u}) \models Y = y$ if $Y = y$ is in the unique solution to equations when exogenous variables are set to $\vec{u}$
  - $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi$ if the Boolean predicate φ holds in the setting where the exogenous variables are set to $\vec{u}$ and endogenous variables $\vec{X}$ are set to $\vec{x}$
  - The causal model where variables $\vec{X}$ are set to $\vec{x}$ is denoted as $M_{\vec{X} \leftarrow \vec{x}}$

Defining an actual cause, revisited

- We want to reason about whether A is the cause of B given $(M, \vec{u})$

Defining an actual cause, revisited

- We want to reason about whether A is the cause of B given $(M, \vec{u})$
  - This assumes all relevant facts are given in the structural equation model and context
  - Given these, we have no need for probability
  - Halpern emphasizes that causality is only defined with respect to a model

Defining an actual cause, revisited

- We want to reason about whether A is the cause of B given $(M, \vec{u})$
  - This assumes all relevant facts are given in the structural equation model and context
  - Given these, we have no need for probability
  - Halpern emphasizes that causality is only defined with respect to a model
- The cause (A) is typically restricted to be a conjunction of primitive events, i.e a joint setting of variables $\vec{X} = \vec{x}$

Defining an actual cause, revisited

- We want to reason about whether A is the cause of B given $(M, \vec{u})$
  - This assumes all relevant facts are given in the structural equation model and context
  - Given these, we have no need for probability
  - Halpern emphasizes that causality is only defined with respect to a model
- The cause (A) is typically restricted to be a conjunction of primitive events, i.e a joint setting of variables $\vec{X} = \vec{x}$
- The effect (B) can be any arbitrary Boolean combination φ of primitive events

Formal Definition (Halpern 2015)

$\vec{X} = \vec{x}$ is an actual cause of φ in situation $(M, \vec{u})$ if

- **(AC1)** Both $\vec{X} = \vec{x}$ and φ are true in the real world, i.e $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$

Formal Definition (Halpern 2015)

$\vec{X} = \vec{x}$ is an actual cause of φ in situation $(M, \vec{u})$ if

- **(AC1)** Both $\vec{X} = \vec{x}$ and φ are true in the real world, i.e $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$
- **(AC2)** Coming up on the next slide...

Formal Definition (Halpern 2015)

$\vec{X} = \vec{x}$ is an actual cause of φ in situation $(M, \vec{u})$ if

- **(AC1)** Both $\vec{X} = \vec{x}$ and φ are true in the real world, i.e $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$
- **(AC2)** Coming up on the next slide...
- **(AC3)** $\vec{X}$ is minimal and no irrelevant variables are included in the set. Formally, no subset of $\vec{X}$ will satisfy both AC1 and AC2

AC2

AC2 is meant to capture the counterfactual requirements on the variables involved, and has been revised multiple times in the literature

- There is a set $\overrightarrow{W}$ of variables in $\mathcal{V}$ such that if we keep them fixed at their actual values, then changing $\vec{X}$ can change the outcome φ

AC2

AC2 is meant to capture the counterfactual requirements on the variables involved, and has been revised multiple times in the literature

- There is a set $\overrightarrow{W}$ of variables in $\mathcal{V}$ such that if we keep them fixed at their actual values, then changing $\vec{X}$ can change the outcome φ
- Formally, we have a set $\overrightarrow{W} \subset \mathcal{V}$ and an alternative setting $\vec{x'}$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \overrightarrow{W} = \vec{w}$ then

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}, \overrightarrow{W} \leftarrow \vec{w}]\neg\varphi$$

AC2

AC2 is meant to capture the counterfactual requirements on the variables involved, and has been revised multiple times in the literature

- There is a set $\vec{W}$ of variables in $\mathcal{V}$ such that if we keep them fixed at their actual values, then changing $\vec{X}$ can change the outcome φ
- Formally, we have a set $\vec{W} \subset \mathcal{V}$ and an alternative setting $\vec{x'}$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}$ then

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}, \vec{W} \leftarrow \vec{w}]\neg\varphi$$

- We refer to $\vec{X} = \vec{x'}, \vec{W} = \vec{w}$ as a *witness* for AC2 holding

A note on probabilistic causality

- To talk about the probability that A is a cause of B, Halpern's book (p. 48) suggests we could convert a single causal setting where the equations are probabilistic to a probability over settings, where in each causal setting the equations are deterministic.
- The probabability that A is an actual cause of B is hence the fraction of (deterministic) worlds in which A is an actual cause of B

Throwing rocks, revisited

Here is a simple causal model for the example situation where Suzy and Billy each throw rocks at a bottle

figs/rocks1.png

- We have three binary variables ST (Suzy throws), BT (Billy throws) and BS (bottle shatters). Say that $\mathrm{BS} = \mathrm{ST} \vee \mathrm{BT}$

Throwing rocks, revisited

Here is a simple causal model for the example situation where Suzy and Billy each throw rocks at a bottle

figs/rocks1.png

- We have three binary variables ST (Suzy throws), BT (Billy throws) and BS (bottle shatters). Say that BS $=$ ST $\vee$ BT
- BT and ST play symmetric roles in this model and nothing distinguishes them. We can easily show that both ST$=$1 and BT$=$1 can be causes of BS$=$1

Throwing rocks, revisited

Here is a simple causal model for the example situation where Suzy and Billy each throw rocks at a bottle

`figs/rocks1.png`

- We have three binary variables ST (Suzy throws), BT (Billy throws) and BS (bottle shatters). Say that BS $=$ ST $\vee$ BT
- BT and ST play symmetric roles in this model and nothing distinguishes them. We can easily show that both ST$=$1 and BT$=$1 can be causes of BS$=$1
- To take into account the fact that Suzy's rock hit first, we need to revise the model

Throwing rocks, revisited

Updated model, which is designed to capture the fact that Suzy's throw is the cause by explicitly making Suzy's throw temporally precede Billy's throw

figs/rocks2.png

- SH (Suzy's rock hits the intact bottle)
- BH (Billy's rock hits the intact bottle)
- $BS = BH \vee SH$
- $SH = ST$
- $BH = BT \wedge \neg SH$

Throwing rocks, revisited

figs/rocks2.png

- We have $BS = BH \vee SH$, $SH = ST$ and $BH = BT \wedge \neg SH$

Throwing rocks, revisited

figs/rocks2.png

- We have $BS = BH \vee SH$, $SH = ST$ and $BH = BT \wedge \neg SH$
- We know that what actually happened was ST=1, BT=1, SH=1, BH=0 and BS=1

Throwing rocks, revisited

figs/rocks2.png

- We have $BS = BH \vee SH$, $SH = ST$ and $BH = BT \wedge \neg SH$
- We know that what actually happened was ST=1, BT=1, SH=1, BH=0 and BS=1
- Is ST=1 a cause of BS=1?
  - AC1 is satisfied because we know ST=1 and BS=1 actually happened, and AC3 is trivially satisfied
  - We have that $(M, \vec{u}) \models [ST \leftarrow 0, BH \leftarrow 0](BS = 0)$
  - In other words, we set BH to its actual value and found that changing ST led to a change in BS

Throwing rocks, revisited

figs/rocks2.png

- We have $BS = BH \vee SH$, $SH = ST$ and $BH = BT \wedge \neg SH$
- We know that what actually happened was ST=1, BT=1, SH=1, BH=0 and BS=1
- Is BT=1 a cause of BS=1?
  - No! We can verify that there is no way to set any of the other variables at their actual values and still have both BT=0 and BS=0
  - Note: Halpern and Pearl's example bakes in temporal precedence of Suzy's throw but in general this is not necessary; we could allow who hits first to be determined using the exogenous variables or time-indexed variables

Are we done?

- *The receptionist in the philosophy department keeps their desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. In practice, both assistants and faculty members take the pens. On Monday morning, both an assistant and a professor take pens. Later, the receptionist needs to take an important message, but there are no pens left on their desk. Who is the cause?*

Are we done?

- *The receptionist in the philosophy department keeps their desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. In practice, both assistants and faculty members take the pens. On Monday morning, both an assistant and a professor take pens. Later, the receptionist needs to take an important message, but there are no pens left on their desk. Who is the cause?*
- Think of the naive model where the professor and the assistant play symmetric roles - the graph is the same as the rock-throwing example

Are we done?

- *The receptionist in the philosophy department keeps their desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. In practice, both assistants and faculty members take the pens. On Monday morning, both an assistant and a professor take pens. Later, the receptionist needs to take an important message, but there are no pens left on their desk. Who is the cause?*
- Think of the naive model where the professor and the assistant play symmetric roles - the graph is the same as the rock-throwing example
- We would think it should be the assistant, but we can't lean on temporal precedence to disambiguate

Normality

- To deal with this, Halpern appeals to a theory of *normality* as an ordering on worlds
- (Kahneman and Miller 1986) *"an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it."*

Normality

- To deal with this, Halpern appeals to a theory of *normality* as an ordering on worlds
- (Kahneman and Miller 1986) *"an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it."*
- The world where the professor does not take the pen and the assistant does is more *normal* than the world where the professor takes it and the assistant doesn't.

Normality

- To deal with this, Halpern appeals to a theory of *normality* as an ordering on worlds
- (Kahneman and Miller 1986) *"an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it."*
- The world where the professor does not take the pen and the assistant does is more *normal* than the world where the professor takes it and the assistant doesn't.
- To incorporate this, add a constraint on AC2 that requires the witness world to be at least as *normal* as the actual world

Graded Causation

- Halpern and Hitchcock also suggest using normality conditions to compare and select potential causes of a given outcome based on the normality of their witness worlds

Graded Causation

- Halpern and Hitchcock also suggest using normality conditions to compare and select potential causes of a given outcome based on the normality of their witness worlds
- $\overrightarrow{X_1} = \vec{x_1}$ is a better cause of φ than $\overrightarrow{X_2} = \vec{x_2}$ if the most normal witness for $\overrightarrow{X_1} = \vec{x_1}$ being a cause of φ is *more normal* than $\overrightarrow{X_2} = \vec{x_2}$ being a cause of φ

Throwing rocks, re-revisited

- Suppose we declare the world where Billy throws a rock, Suzy doesn't throw, and Billy does not hit abnormal[1]

[1] J.Y. Halpern (2018), *Actual Causality*, Example 3.2.6

Throwing rocks, re-revisited

- Suppose we declare the world where Billy throws a rock, Suzy doesn't throw, and Billy does not hit abnormal[1]
- This world was needed to show that Suzy throwing is a cause in the previous example

[1] J.Y. Halpern (2018), *Actual Causality*, Example 3.2.6

Throwing rocks, re-revisited

- Suppose we declare the world where Billy throws a rock, Suzy doesn't throw, and Billy does not hit abnormal[1]
- This world was needed to show that Suzy throwing is a cause in the previous example
- Thus, with this normality ordering ST $= 1$ is not a cause, and we can show instead that $ST = 1 \wedge BT = 1$ is the cause

[1] J.Y. Halpern (2018), *Actual Causality*, Example 3.2.6

Throwing rocks, re-revisited

- Halpern justifies this by claiming that worlds are only settings of endogenous variables. In such a case, *"the witness world where $BT = 1$, $BH = 0$, and $ST = 0$ does not seem so abnormal, even if it is abnormal for Billy to throw and miss in a context where he is presumed accurate"*

[2] J.Y. Halpern (2018), *Actual Causality*, Sec. 3.5

Throwing rocks, re-revisited

- Halpern justifies this by claiming that worlds are only settings of endogenous variables. In such a case, *"the witness world where $BT = 1$, $BH = 0$, and $ST = 0$ does not seem so abnormal, even if it is abnormal for Billy to throw and miss in a context where he is presumed accurate"*
- Imposing normality conditions to judge whether a variable is the cause of an outcome can lead to either counterintuitive or empty results

[2] J.Y. Halpern (2018), *Actual Causality*, Sec. 3.5

Throwing rocks, re-revisited

- Halpern justifies this by claiming that worlds are only settings of endogenous variables. In such a case, *"the witness world where $BT = 1$, $BH = 0$, and $ST = 0$ does not seem so abnormal, even if it is abnormal for Billy to throw and miss in a context where he is presumed accurate"*
- Imposing normality conditions to judge whether a variable is the cause of an outcome can lead to either counterintuitive or empty results
- Halpern suggests an alternative by placing normality orderings on events (sets of contexts) instead of worlds[2]

[2] J.Y. Halpern (2018), *Actual Causality*, Sec. 3.5

Are we done?

- Many papers on Actual Causality present their arguments by[3]:
  - coming up with a definition (for example AC2) that captures some variant of the NESS intuition: $X = x$ causes $Y = y$ iff $X = x$ is a Necessary Element of a Sufficient Set (NESS) for $Y = y$
  - demonstrating that this definition offers intuitive verdicts on a number of problematic examples in the literature

[3] S. Beckers (2021), *Causal Sufficiency and Actual Causation*, Journal of Philosophical Logic

Are we done?

- What's the problem?
  - The correctness of a definition is not in itself provable or testable, but we can make progress by testing their consistency and compare them with other definitions
  - There are just too many examples and intuitions: the number of possible statements of actual causality to test blows up exponentially with the number of potential causes
  - Coming up with "natural" restrictions on the set of possible structures will arguably not lead to a tractable number of cases[4]

[4] C. Glymour et. al (2010), *Actual causation: a stone soup essay*, Synthese

How far can we get with examples?

Consider the following SEM:

- The variables are X, Y and Z

How far can we get with examples?

Consider the following SEM:

- The variables are X, Y and Z
- The equations are $X = Y$ and $Z = X \vee Y$

How far can we get with examples?

Consider the following SEM:

- The variables are X, Y and Z
- The equations are $X = Y$ and $Z = X \vee Y$
- The actual case we encountered was X=1, Y=1, Z=1

How far can we get with examples?

Consider the following SEM:

- The variables are X, Y and Z
- The equations are $X = Y$ and $Z = X \vee Y$
- The actual case we encountered was X=1, Y=1, Z=1
- We can verify that X = 1 is a cause of Z = 1, but Y = 1 is not a cause. Y = 1 won't be a partial cause either, because it is not minimal

How far can we get with examples?

Consider the following SEM:

- The variables are X, Y and Z
- The equations are $X = Y$ and $Z = X \vee Y$
- The actual case we encountered was X=1, Y=1, Z=1
- We can verify that X = 1 is a cause of Z = 1, but Y = 1 is not a cause. Y = 1 won't be a partial cause either, because it is not minimal
- This could make intuitive sense in a fault detection context. Think of X as a component whose value is stuck at 1, which leads to Y=1 and Z=1, but the desired behaviour of the system is Z=0. The repair would be to get a new component that sets X=0.

How far can we get with examples?

Consider the following example[5]:

- *An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible?*

[5] I. Rosenberg and C. Glymour (2018), *Review of Joseph Halpern, Actual Causality*

How far can we get with examples?

Consider the following example[5]:

- *An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible?*
- The variables are X (gang members shoot), Y (leader shoots) and Z (victim dies), and the equations are $X = Y$ and $Z = X \vee Y$

[5] I. Rosenberg and C. Glymour (2018), *Review of Joseph Halpern, Actual Causality*

How far can we get with examples?

Consider the following example[5]:

- *An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible?*
- The variables are X (gang members shoot), Y (leader shoots) and Z (victim dies), and the equations are $X = Y$ and $Z = X \vee Y$
- This case is isomorphic to the previous one

[5] I. Rosenberg and C. Glymour (2018), *Review of Joseph Halpern, Actual Causality*

How far can we get with examples?

Consider the following example[5]:

- *An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible?*
- The variables are X (gang members shoot), Y (leader shoots) and Z (victim dies), and the equations are $X = Y$ and $Z = X \vee Y$
- This case is isomorphic to the previous one
- Our answer would be that X=1 is the cause, and Y=1 is neither a cause nor a partial cause

[5] I. Rosenberg and C. Glymour (2018), *Review of Joseph Halpern, Actual Causality*

How far can we get with examples?

Consider the following example[5]:

- *An obedient gang is ordered by its leader to join him in murdering someone, and does so, all of them shooting the victim at the same time. The action of any one of the gang would suffice for the victim's death. If responsibility implies causality, whom among them is responsible?*
- The variables are X (gang members shoot), Y (leader shoots) and Z (victim dies), and the equations are $X = Y$ and $Z = X \vee Y$
- This case is isomorphic to the previous one
- Our answer would be that X=1 is the cause, and Y=1 is neither a cause nor a partial cause
- Can show that we will get the same result even if we have $P(X = 1|Y = 1) = p$ instead of $X = Y$

[5] I. Rosenberg and C. Glymour (2018), *Review of Joseph Halpern, Actual Causality*

Summary

- Questions of actual causality come up in philosophical and legal scenarios, and could be of importance to CS research

Summary

- Questions of actual causality come up in philosophical and legal scenarios, and could be of importance to CS research
- Actual causation is typically described in terms of counterfactuals, which are typically described in terms of SEMs

Summary

- Questions of actual causality come up in philosophical and legal scenarios, and could be of importance to CS research
- Actual causation is typically described in terms of counterfactuals, which are typically described in terms of SEMs
- There have been various revisions to the definition, each of which aim to capture some form of the NESS intuition and cover intuitive examples of interest to the community

Summary

- Questions of actual causality come up in philosophical and legal scenarios, and could be of importance to CS research
- Actual causation is typically described in terms of counterfactuals, which are typically described in terms of SEMs
- There have been various revisions to the definition, each of which aim to capture some form of the NESS intuition and cover intuitive examples of interest to the community
- A sound and complete theory of actual causation has so far proven to be difficult

Summary

- Questions of actual causality come up in philosophical and legal scenarios, and could be of importance to CS research
- Actual causation is typically described in terms of counterfactuals, which are typically described in terms of SEMs
- There have been various revisions to the definition, each of which aim to capture some form of the NESS intuition and cover intuitive examples of interest to the community
- A sound and complete theory of actual causation has so far proven to be difficult
- Extensions to the theory attempt to use considerations about normality of worlds/events to deal with graded causation for a single effect, responsibility, blame and explanation

Responsibility

- Say that A wins an election against B by a vote of 11–0. The outcome variable is $O = 1$ if A wins and zero otherwise.

[6] Gerstenberg et. al (2015), *Responsibility judgments in voting scenarios*, CogSci

Responsibility

- Say that A wins an election against B by a vote of 11–0. The outcome variable is $O = 1$ if A wins and zero otherwise.
- Each voter V_i is a cause of the outcome. But their degree of responsibility is not the same as it would be if the vote were 6-5

[6] Gerstenberg et. al (2015), *Responsibility judgments in voting scenarios*, CogSci

Responsibility

- Say that A wins an election against B by a vote of 11–0. The outcome variable is $O = 1$ if A wins and zero otherwise.
- Each voter V_i is a cause of the outcome. But their degree of responsibility is not the same as it would be if the vote were 6-5
- $V_1 = 1$ is a cause of O=1 in a context where everyone votes for B. We can prove AC2 holds by showing that if $V_1, ..., V_6$ were set to zero then $O = 0$

[6] Gerstenberg et. al (2015), *Responsibility judgments in voting scenarios*, CogSci

Responsibility

- Say that A wins an election against B by a vote of 11–0. The outcome variable is $O = 1$ if A wins and zero otherwise.
- Each voter V_i is a cause of the outcome. But their degree of responsibility is not the same as it would be if the vote were 6-5
- $V_1 = 1$ is a cause of O=1 in a context where everyone votes for B. We can prove AC2 holds by showing that if $V_1, ..., V_6$ were set to zero then $O = 0$
- $V_1 = 1$ is a cause of O=1 in a context where only $V_1, ..., V_6$ vote for B. We can prove AC2 holds by showing that if V_1 were set to zero then $O = 0$

[6] Gerstenberg et. al (2015), *Responsibility judgments in voting scenarios*, CogSci

Responsibility

- Say that A wins an election against B by a vote of 11–0. The outcome variable is $O = 1$ if A wins and zero otherwise.
- Each voter V_i is a cause of the outcome. But their degree of responsibility is not the same as it would be if the vote were 6-5
- $V_1 = 1$ is a cause of O=1 in a context where everyone votes for B. We can prove AC2 holds by showing that if $V_1, ..., V_6$ were set to zero then $O = 0$
- $V_1 = 1$ is a cause of O=1 in a context where only $V_1, ..., V_6$ vote for B. We can prove AC2 holds by showing that if V_1 were set to zero then $O = 0$
- Key idea: use the size of $\overrightarrow{W}$ to determine the degree of responsibility[6]

[6] Gerstenberg et. al (2015), *Responsibility judgments in voting scenarios*, CogSci

Blame

- A doctor's use of a drug to treat a patient may have been the cause of a patient's death. Their degree of responsibility will then be 1, even if they had no idea there would be adverse side effects

Blame

- A doctor's use of a drug to treat a patient may have been the cause of a patient's death. Their degree of responsibility will then be 1, even if they had no idea there would be adverse side effects
- Is the doctor to *blame* for this death?

Blame

- A doctor's use of a drug to treat a patient may have been the cause of a patient's death. Their degree of responsibility will then be 1, even if they had no idea there would be adverse side effects
- Is the doctor to *blame* for this death?
- In legal reasoning, what matters is not only what was known, but what shoulf have been known

Blame

- A doctor's use of a drug to treat a patient may have been the cause of a patient's death. Their degree of responsibility will then be 1, even if they had no idea there would be adverse side effects
- Is the doctor to *blame* for this death?
- In legal reasoning, what matters is not only what was known, but what shoulf have been known
- Halpern suggests defining blame relative to an epistemic state, which is the agent's own distribution of plausibility over scenarios

Blame

- A doctor's use of a drug to treat a patient may have been the cause of a patient's death. Their degree of responsibility will then be 1, even if they had no idea there would be adverse side effects
- Is the doctor to *blame* for this death?
- In legal reasoning, what matters is not only what was known, but what shoulf have been known
- Halpern suggests defining blame relative to an epistemic state, which is the agent's own distribution of plausibility over scenarios
- The degree of blame is the expected degree of responsibility, taken over the situations the agent considers possible

Blame: Example

- Consider a firing squad with 10 excellent marksmen. Only one of them has live bullets in his rifle; the rest have blanks. The marksmen do not know which of them has the bullets.
- All marksmen shoot at once and the prisoner dies
- Only one marksman will be the cause of death and have degree of responsibility 1. All others have zero responsibility
- However, the expected degree of responsibility, or degree of blame, is $1/10$