Subjective Causality in Choice

Andrew Ellis   Heidi Thysen
LSE   NHH

December 31, 2021

Motivation

- Beliefs about which correlations are **causal** affect behavior
  - E.g. why is duration of hospitalization correlated with death?

- Causation **requires** a model, and a **misspecified model** distorts agent's beliefs about her actions

- Can an analyst **reveal** the model that generates the agent's distorted beliefs from her behavior?

Running Example

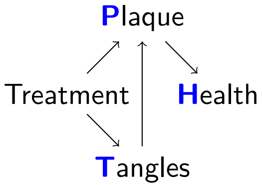
- Doctor (the **DM**) decides between treatments
  - Leads to data on patient's health and several correlates

    Health outcome: Alzheimer's or not
    Correlates: amyloid plaque buildup, neurofibrillary tangles

- **Analyst**, e.g. drug company, sees treatment choices
  - Wants to know why she does not usually choose its treatment
  - Can misspecified causal model explain her decisions?
  - If so, what is doctor's causal model?

[Loosely based on FDA approval of aducanumab]

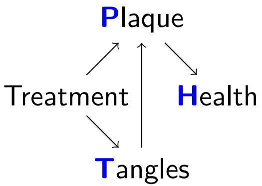
Following Pearl (1995) and Spiegler (2016), we model perception of causality with a **directed acyclic graph (DAG)**



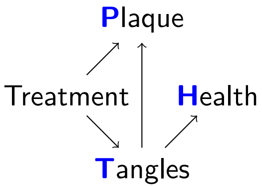
- Arrows indicate **causes**
- **Flexible** & **non-parametric**

- predictions made "as if" each variable independent of its non-descendants given its direct causes

  e.g. Health independent of Treatment & Tangles given Plaque



R_P



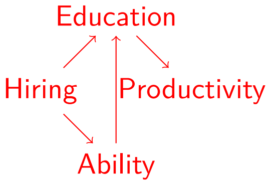
R_T

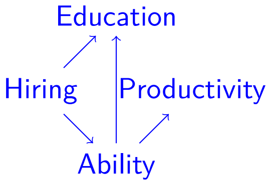
Evidence that a drug decreases Plaque should increase prescriptions with causal model R_P but not R_T

(even if R_P correct!)

▸ Example of an actual DAG used in health research

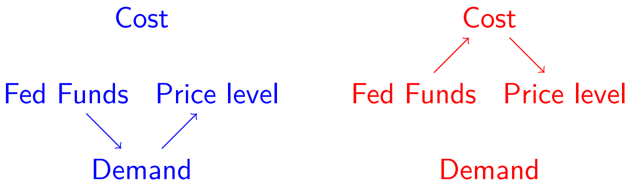
Other Examples





Human capital versus **signaling** models of education

Other Examples



Andre et al (2021) provide survey evidence that experts tend to use a **demand-side model** and laypeople a **supply-side model**

Objectives

A DM has a **subjective causality representation (SCR)** if she

- predicts outcome of her action using an endogenous dataset and a subjective causal model described by a **DAG**
- uses those predictions to maximize expected utility (plus Logit shock)

and her choices form a **personal equilibrium**:

choices affect data affect predictions affect choices affect ...
when model **misspecified**

Objectives

- Analyst observes only the DM's behavior, a **random choice rule** ρ over lotteries, not her model

- **Result 1**: Identifies subjective causal model from ρ

  minimal causal chains revealed by behavior
  Occam's razor – these "simplest" chains pin down model

- **Result 2**: Provides test of model in terms of ρ (axioms)

Implications

- Model accommodates lab evidence of biased inferences
  - selection neglect (Esponda and Vespa, 2018)
  - congruence bias (Wason, 1960)
  - illusion of control (Langer, 1975)
  - patternicity (Shermer, 1998)

- Mechanism: DM's behavior endogenously creates **correlations** that she misinterprets **as causation**

- Leads to two technical challenges in analyzing model:
  - **violation of regularity** (stochastic WARP)
  - **self-confirming behavior** (multiple equilibria)

Big Picture

- Growing literature studying **misspecified** models:
  - **Spiegler (2016)**, Eliaz & Spiegler (2018), Eliaz et al. (2019), Schumacher & Thysen (2020), Spiegler (2020), etc.
  - Esponda & Pouza (2016), He (2018), Bohren and Hauser (2018), Heidhues et al. (2018), Samuelson & Mailath (2019), Frick et al. (2020), Fudenberg et al. (2021), Levy et al. (2021), Montiel Olea et al. (2021), etc.

- Given that the model is **misspecified**, how can we identify what the agent believes it is?
  - Lipman (1999), Ellis & Piccione (2017), Kochov (2018), Ke et al. (2021), Ellis & Masatlioglu (2021), etc.
  - **this paper**

Outline

1. Model
2. Identification
3. Foundations

Setup

- **Single, payoff-relevant** consequence, indexed $n+1$, and n **covariates**, indexed $1, 2, \ldots, n$
  - e.g. two correlates T, P (Tangles, Plaque) and **H**ealth
    DM only cares about health $(H = n+1 = 3)$

- DM's **action** determines their distribution
  action $=$ full support **lottery** over vectors $\{G, B\}^{n+1}$
  - e.g. "given type of patient undergoes treatment"

    lottery captures uncertain effects of treatment

- Analyst observes **random choice rule** indicating how often DM chooses each action a from **every** menu S: $\rho(a, S)$
  - e.g. how often doctor prescribes treatment a when only treatments $S = \{a, b\}$ available

Model

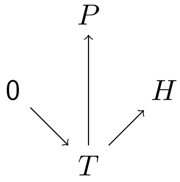
DM calibrates her model using the (endogenous) dataset ρ^S

$$\rho^S(a, x_P, y_T, z_H) = \rho(a, S)a(x_P, y_T, z_H)$$

generated by combining the **lotteries** in a menu S with the **frequency** with which they are chosen

e.g. $\rho^S(a, G, B, G)$ is fraction of patients who got treatment a, had good level of plaque, had bad level of tangles, and good health

Uses ρ^S to estimate the causal effects in the **DAG** R



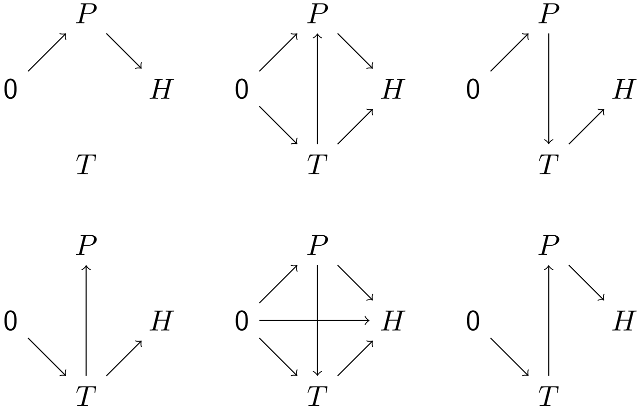
Predicts that chance of good health if she takes action a is

$$\rho^S_R(G_H|a) = \sum_{y=G,B} \rho^S(y_T|a)\left[\sum_{x=G,B} \rho^S(x_P|y_T\cancel{, a})\rho^S(G_H|y_T\cancel{, x_P, a})\right]$$

$$= \sum_{y=G,B} \rho^S(y_T|a)\left[\qquad\qquad \rho^S(G_H|y_T \qquad)\right]$$

General formula

Running Example: DAGs



Definition

Random choice rule ρ has **subjective causality representation (SCR)** if there exists "well-behaved" DAG R and continuous, increasing u so that for every menu S and action $a \in S$:

$$\rho(a, S) = \frac{\exp\left(\sum_{z=G,B} u(z)\rho_R^S(z_H|a)\right)}{\sum_{a' \in S} \exp\left(\sum_{z=G,B} u(z)\rho_R^S(z_H|a')\right)}$$

Equivalent to maximizing

$$\sum_{z=G,B} u(z)\rho_R^S(z_H|a) + \epsilon_a$$

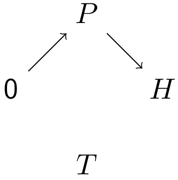
when $\epsilon_a \sim$ iid extreme value

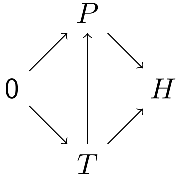
▸ Accommodates biases

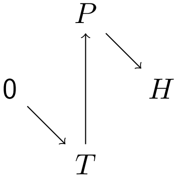
DAG is well-behaved if

1. Nothing causes action (uninformed)
2. Path from action to outcome (nontrivial)
3. Common causes are linked (perfect)
   predicts marginal distribution of each variable correctly

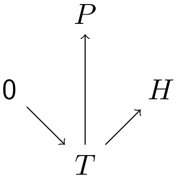
Identify DAG by revealing its $\subseteq$-minimal paths
each path represents a chain of causal relationships

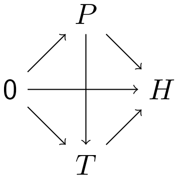


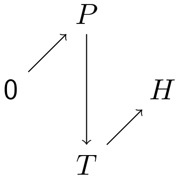




Min Paths: $(0, P, H)$ $(0, P, H), (0, T, H)$ $(0, T, P, H)$







Min Paths: $(0, T, H)$ $(0, H)$ $(0, P, T, H)$

Theorem

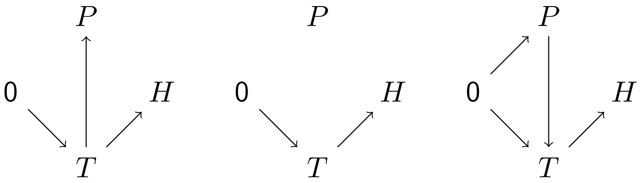
Let ρ have a SCR (R, u) and R' be a well-behaved DAG

Then, ρ also has a SCR (R', u') ***if and only if***
every R' minimal path *is a* ***R minimal path*** *& vice versa*
and there is β so that $u(x) = u'(x) + \beta$ for all $x \in \{G, B\}$

two DMs behave same only if agree on causal chains

1. Any DAGs that represent ρ agree on minimal paths
   - Occam's razor
2. Only the variables in some minimal path matter
   - The rest can be dropped without loss
3. Links in minimal paths capture most causal relationships
   - Only exceptions are those between common parents

Intuition: Occam's razor



R $\qquad$ $\hat{R}^\rho$ $\qquad$ R'

ρ represented by $R \iff$ rep'ed by $\hat{R}^\rho \iff$ rep'ed by R':

- for R, saw $\rho^S_R(G_H|a) = \rho^S_{\hat{R}^\rho}(G_H|a)$
- for R', y_P integrates out so $\rho^S_R(x_T|a) = \rho^S_{R'}(x_T|a)$ for any x

Proof

Two steps to reveal the minimal paths:

1. Reveal the sets of variables that intersect every path to consequence from action

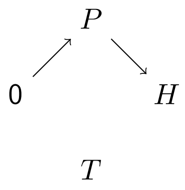
   They "**separate**" or "block all paths to" outcome from action

2. **Order** minimal separators to recover which causes which

Key observation: If two variables are ρ^S-**independent** of each other, then DM estimates the causal effect to be **zero**

Illustration: Separates

- Suppose that plaque is **independent** of all other variables regardless of how DM chooses from $\{a, b\}$
- If DM **doesn't believe** either action or tangles causes Health, then **infers** that all treatments lead to same expected health
- Hence indifferent: $\rho(a, \{a, b\}) = \rho(b, \{a, b\})$
- $\{P\}$ **separates** if indifferent for all such $\{a, b\}$; **reveals** that **all** paths from 0 to H go through P



Proof: Separates

Definition

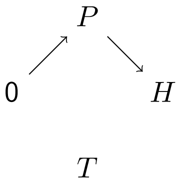
A subset of covariates I **separates** if $\rho(a,\{a,b\}) = \frac{1}{2}$, i.e. DM is indifferent, whenever the variables indexed by I are **independent** of all others, regardless of how DM chooses from $\{a,b\}$

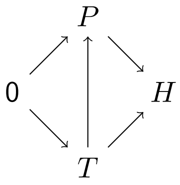
Lemma

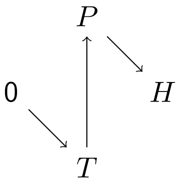
*A subset of covariates I **separates** **if and only if** **every path** from 0 to $n+1$ in the DM's DAG **intersects** I*

Notation: $\mathcal{A} = \{I \subset N : I \text{ is a minimal set that separates}\}$

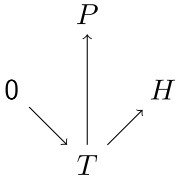
Illustration: Separators

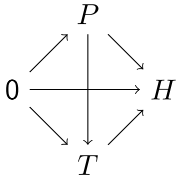


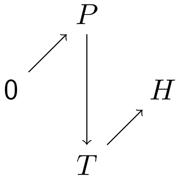




$\mathcal{A}$: $\{\{P\},\{H\}\}$ $\{\{T,P\},\{H\}\}$ $\{\{T\},\{P\},\{H\}\}$







$\mathcal{A}$: $\{\{T\},\{H\}\}$ $\{\{H\}\}$ $\{\{T\},\{P\},\{H\}\}$

Illustration: Ordering

- in illustration, if $\mathcal{A} \neq \{\{P\},\{T\},\{H\}\}$, done; otherwise, need to know whether DM believes plaque or tangles **directly** cause disease
- If we know direction of causality, done; otherwise, we need to reveal it

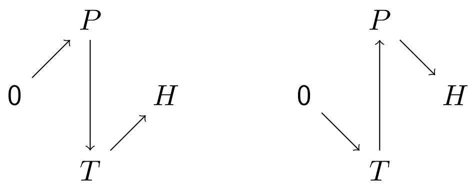


Illustration: Ordering

Intuition: if a variable is **independent** of the next variable in the chain, then the causal chain is **"broken"** so DM is **indifferent**

- H causes nothing, so H comes last
- if **equally likely** to choose either whenever Plaque buildup **independent of** Alzheimer's, then P is third
- if **equally likely** to choose either whenever presence of Tangles **independent of** Alzheimer's, then T is third
- remaining correlate is second
- action is first

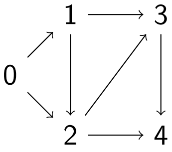
Proof: Ordering

Lemma

$\mathcal{A}$ can be (uniquely) ordered $\mathcal{A} = \{A_1^*, \ldots, A_{|\mathcal{A}|}^*\}$ so that $(i_0, \ldots, i_m)$ ***is a minimal path*** *from 0 to $n+1$* ***if and only if*** *for every j there exists k so that*

$$i_j \in A_k^* \setminus A_{k+1}^* \;\&\; i_{j+1} \in A_{k+1}^* \setminus A_k^*$$

where $A_0^* = \{0\}$.



$$\Longrightarrow \; A_1^* = \{1,2\},\; A_2^* = \{2,3\},\; A_3^* = \{4\}$$

Proof: Ordering

$\{n+1\}$ always in $\mathcal{A}$ and causes nothing, so $A^*_{|\mathcal{A}|} = \{n+1\}$

Definition

The set $A^*_i \in \mathcal{A} \setminus \{A^*_{i+1}, \ldots, A^*_{|\mathcal{A}|}\}$ causes A^*_{i+1} if

- $\rho(a, \{a, b\}) = \frac{1}{2}$ whenever the variables indexed by A^*_{i+1} are **independent** of those indexed by A^*_i regardless of how DM chooses from $\{a, b\}$, and
- A^*_i has $\subseteq$-largest intersection with A^*_{i+1}

- first condition generalizes illustration
  - requires variables in $A^*_i \cap A^*_{i+1}$ are independent of others
- if second condition failed, then we would have a cycle

Implications

Definition

For $j, k \in \{0, 1, \ldots, n+1\}$, j **is revealed to cause** k **for** ρ, written $j\hat{R}^\rho k$, if jRk **whenever** ρ has SCR (R, u)

- If more than one DAG represents ρ, $\hat{R}^\rho$ common to all
- Similar to Masatlioglu et al (2012)

Proposition

If ρ has a SCR, then $j\hat{R}^\rho k$ if and only if there exists i so that $j \in A_i^*$ and $k \in A_{i+1}^* \setminus A_i^*$ when $A_0^* = \{0\}$.

Implications

Proposition

If ρ has a SCR (R, u), then

$$\rho_R^S\left(x_{\cup_{i=1}^{|\mathcal{A}|}A_i^*}|a\right) = a\left(x_{A_1^*}\right)\prod_{i=1}^{|\mathcal{A}|-1}\rho^S\left(x_{A_{i+1}^*\setminus A_i^*}|x_{A_i^*}\right)$$

for ever $x \in \{G, B\}^{n+1}$

(x_E for $E \subset \{1, ..., n+1\}$ is projection of x onto E)

Lessons so far

When ρ has a SCR (R, u), then the DM's choices reveal the **minimal causal chains** in her model

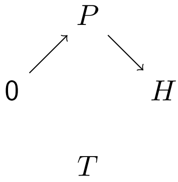
- **Identify** and then **order** the set of minimal separators

Minimal causal chains suffice to recover **every causal** relationship that **must** be in model and **pin down** the DAGs that represent ρ

How can we **refute** the model?

Behavioral Characterization

- Provide axioms equivalent to ρ having SCR
- Statements given within context of running example with $\mathcal{A} = \{\{P\}, \{H\}\}$



Axiom (Full-Support)

For any menu S and action $a \in S$, $\rho(a, S) > 0$.

- every treatment prescribed with positive probability

Axiom (Bounded Misperception)

The Luce ratio is bounded:

$$\sup_{S;a,b\in S} \frac{\rho(a,S)}{\rho(b,S)} < \infty$$

- relative frequency of prescribing a pair of treatments

  is **bounded above** by that of

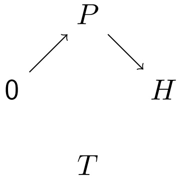
  treatment **known** to lead to good health and one **known** to lead to bad health

- limits how much DM misperceives distribution

  similar to Weak Monotonicity in Ellis and Piccione (2017)

Axiom (Consistent Revealed Causes)

The sets $\{H\}$ and $\{P\}$ are the only sets that separate



- **General:** possible to order $\mathcal{A}$ as in Lemma

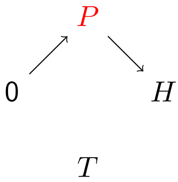
▸ Statement

Axiom (I5)

(Indifferent If Identical Immediate Implications)

For any $a, b \in S$ and any S,
if $\text{marg}_P\, a = \text{marg}_P\, b$, then $\rho(a, S) = \rho(b, S)$.

- equally likely to perform any pair of treatments with same chance of plaque buildup



- **General:** identical distribution over variables immediately caused by action, i.e. in A_1^*, implies indifference ▸ Statement

Axiom (LCI)

(Luce's Choice Axiom Given Inferences)
For any menus $S, S_1, S_2, \ldots$ with $a, b \in S_m \cap S$ for each m, if

$$\rho^{S_m}(x_H|z_P) = \rho^{S}(x_H|z_P) \text{ for all } x, z \in \{G, B\},$$

then

$$\frac{\rho(a, S_m)}{\rho(b, S_m)} = \frac{\rho(a, S)}{\rho(b, S)}$$

- if relationship between plaque and Health is (almost) the same, then (almost) satisfies Luce's choice axiom
- **General:** almost same inferences implies LCA close to holding
  ▸ Statement

Axiom (LCI)

(Luce's Choice Axiom Given Inferences)
For any menus $S, S_1, S_2, \ldots$ with $a, b \in S_m \cap S$ for each m, if

$$\rho^{S_m}(x_H|z_P) \to \rho^S(x_H|z_P) \text{ for all } x, z \in \{G, B\},$$

then

$$\frac{\rho(a, S_m)}{\rho(b, S_m)} \to \frac{\rho(a, S)}{\rho(b, S)}$$

- if relationship between plaque and Health is (**almost**) the same, then (**almost**) satisfies Luce's choice axiom
- **General:** almost same inferences implies LCA close to holding
  ▸ Statement

Definition

A menu S is **(revealed to be) correctly perceived** if

$$a(x_H|z_P) = b(x_H|z_P)$$

for all $x, z \in \{G, B\}$ and $a, b \in S$

General: menu correctly perceived whenever revealed causes match actual conditional independence structure of menu

$X_{A^*_{i+1} \setminus A^*_i} \perp X_{A^*_j \setminus A^*_i}$ given $X_{A^*_i}$ for all $j < i$ regardless of how DM chooses ▸ Statement

Axiom (Correctly Perceived Logit)

There exists a increasing $u : \{G, B\} \to \mathbb{R}$ so that

$$\rho(a, S) = \frac{\exp\left(\sum_{x=G,B} u(x)a(x_H)\right)}{\sum_{b \in S} \exp\left(\sum_{x=G,B} u(x)b(x_H)\right)}$$

for every correctly perceived menu S

- When DM perceives menus correctly, acts "rationally"
- "independence," "monotonicity," and "continuity"

THEOREM

A random choice rule ρ has a SCR ***if and only if***
ρ satisfies Full-Support, Bounded Misperception, Consistent Revealed Causes, I5, LCI, and Correctly Perceived Logit

Axioms make clear when & why ρ differs from Logit-EU

- BM – maximum deviation from Luce rule
- LCI – only violates Luce rule when inferences change
- I5 – same (incorrect) perception implies same frequency
- Correctly Perceived Logit

Corollary

A random choice rule ρ is Logit-EU ***if and only if***
ρ has a SCR and $\mathcal{A} = \{\{n+1\}\}$ ***if and only if***
$\mathcal{A} = \{\{n+1\}\}$ and ρ satisfies Full-support, I5, Bounded Misperception, LCI, and Correctly Perceived Logit

When $0 \; R \; n+1$:

- $A_1^* = \{n+1\}$ so coincidence of a and b's consequence distribution implies equal probability of choosing either
- hypothesis of LCI holds vacuously, so choice axiom holds

Proof Outline

1. For any S, let S'_1 be a correctly perceived copy of S: for each $a \in S$ there is $a' \in S'_1$ so that $a'(x) = a(x_{A_1^*}) \prod_i \rho^S(x_{A_{i+1}^* \setminus A_i^*} | x_{A_i^*})$
2. Form a nested sequence of menus $(S'_m)_m$ so that each S'_m is correctly perceived and $|S'_m| \to \infty$; $\rho^{S'_m}(y_{A_{i+1}^* \setminus A_i^*} | y_{A_i^*}) = \rho^S(y_{A_{i+1}^* \setminus A_i^*} | y_{A_i^*})$ for all i, y

   Pick any $a, b \in S$

3. BM implies $\rho(S, S \cup S'_m) \to 0$ and so $\rho^{S'_m \cup S}(y_{A_{i+1}^* \setminus A_i^*} | y_{A_i^*}) \to \rho^S(y_{A_{i+1}^* \setminus A_i^*} | y_{A_i^*})$ for all i, y
4. I5 implies $\frac{\rho(a', S'_m \cup S)}{\rho(b', S'_m \cup S)} = \frac{\rho(a, S'_m \cup S)}{\rho(b, S'_m \cup S)}$ for each m
5. LCI implies that $\frac{\rho(a', S'_m \cup S)}{\rho(b', S'_m \cup S)} = \frac{\rho(a, S'_m \cup S)}{\rho(b, S'_m \cup S)} \to \frac{\rho(a', S'_1)}{\rho(b', S'_1)}$
6. LCI also implies $\frac{\rho(a, S'_m \cup S)}{\rho(b, S'_m \cup S)} = \frac{\rho(a', S'_m \cup S)}{\rho(b', S'_m \cup S)} \to \frac{\rho(a, S)}{\rho(b, S)}$

   Applying Correctly Perceived Logit completes proof

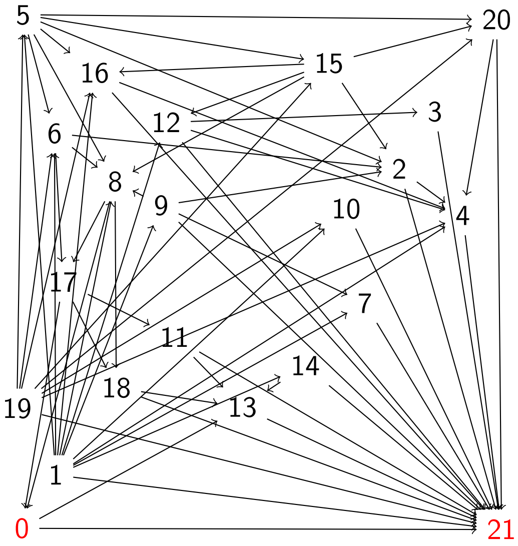
Other Literature

- **DAG as causal model:** Pearl (2009)
- **Axioms for normative model of causality:** Schenone (2020)
- **Boundedly Rational Stochastic Choice:** Manzini & Mariotti (2014), Brady & Rehbeck (2016), Cattaneo et al. (2020), etc.
- **Stochastic Choice with rich alternatives:** Gul & Pesendorfer (2006), Lu (2016), Apesteguia & Ballester (2018), etc.

Wrap up

- Provided identification and axiomatization result for model of subjective causality
- Extensions:
  - Relationships to cognitive biases Details
  - Comparative statics on dropping nodes/links in DAG
  - Deterministic vs. random choice
  - Exogenous datasets
- Directions for future research:
  - Information
  - Imperfect DAGs
  - Empirical/experimental implementation

Thank you



Actual DAG used by Evandt et al. (2017, *Environmental Health*)

General formula

Let $R(i) = \{j : jRi\}$, called parents of i

Given $p \in \Delta\mathcal{X}$, $p_R \in \Delta\mathcal{X}$ is such that

$$p_R(x) = p(x_0) \prod_{j=1}^{n+1} p(x_j | x_{R(j)})$$

when nothing causes 0

For comparison,

$$p(x) = p(x_0) \prod_{j=1}^{n+1} p(x_j | x_{\{0,\ldots,j-1\}})$$

Back

Biases

Consider a doctor who thinks plaque alone causes Alzheimer's, **BUT** both actually caused by tangles

- treatment-π has good P iff good T
- treatment-ι has P independent of T
- Then, **predicted** effect of lowering Plaque increases with fraction treated using π

- May incentivize more prescription of π, reinforcing the effect
- Can cause a **regularity violation**:

$$\rho(\iota, \{\iota, \pi\}) < \rho(\iota, \{\iota, \pi, \nu\})$$

Explains illusion of control and patternicity

Biases

Consider a doctor who thinks plaque alone causes Alzheimer's, **BUT** both actually caused by tangles

- Status quo: no effect on disease and plaque buildup only if diseased
- New drug: prevents the disease and plaque buildup only if healthy

- If usually takes status quo treatment, then predicts new drug increases chance of plaque buildup and so also disease
- If usually prescribes new drug, then predicts status quo treatment harms patients

Explains congruence bias and status quo bias Back

Axiom (Consistent Revealed Causes)

$\{n+1\} \in \mathcal{A}$, and for $i = 1, \ldots, |\mathcal{A}|$, A_i^* exists.

That is, exactly one of the following holds:

(1) $\{H\}$ and $\{P\}$ are the only sets that separate, so $A_1^* = \{H\}$ and $A_2^* = \{P\}$

(2) $\{T\}$ and $\{P\}$ are the only sets that separate, so $A_1^* = \{T\}$ and $A_2^* = \{P\}$

...

(6) $\{H\}$, $\{T\}$, and $\{P\}$ are the only sets that separate and $X_T \perp_{\{a,b\}} X_H$ implies $\rho(a, \{a, b\}) = \frac{1}{2}$, so $A_1^* = \{H\}$, $A_2^* = \{T\}$, and $A_2^* = \{P\}$

◂ Back

Axiom (I5)

For any $a, b \in S$ and any S,
if $\operatorname{marg}_{A_1^*} a = \operatorname{marg}_{A_1^*} b$, then $\rho(a, S) = \rho(b, S)$.

Back

LCI

- Recall: variables in A_i^* revealed to cause those in $A_{i+1}^* \setminus A_i^*$
- Reveals menus for which DM makes same or similar inferences
- if for all i and x

$$\rho^{S'}(x_{A_{i+1}^* \setminus A_i^*} | x_{A_i^*}) = \rho^{S}(x_{A_{i+1}^* \setminus A_i^*} | x_{A_i^*}),$$

  then DM makes the **same inferences** about health outcome of each treatment when facing S as when facing S'

- and if for all i and x

$$\rho^{S'}(x_{A_{i+1}^* \setminus A_i^*} | x_{A_i^*}) \approx \rho^{S}(x_{A_{i+1}^* \setminus A_i^*} | x_{A_i^*}),$$

  then DM makes **almost the same inferences** about health outcome of each treatment when facing S as when facing S'

Axiom (LCI)

For any $S, S_1, S_2, \dots \in \mathcal{S}$ with $a, b \in S_m \cap S$ for each m: if

$$\rho^{S_m}(x_{A^*_{i+1} \setminus A^*_i} | x_{A^*_i}) \to \rho^{S}(x_{A^*_{i+1} \setminus A^*_i} | x_{A^*_i})$$

for every $i = 1, \dots, |\mathcal{A}| - 1$ and $x \in \{G, B\}^{n+1}$, then $\frac{\rho(a,S_m)}{\rho(b,S_m)} \to \frac{\rho(a,S)}{\rho(b,S)}$

Back

Definition (General)

A menu S is **(revealed to be) correctly perceived** if

$$b(x_{\cup_{i=1}^{|\mathcal{A}|} A_i^*}) = b(x_{A_1^*}) \prod_{i=1}^{|\mathcal{A}|-1} a(x_{A_{i+1}^* \setminus A_i^*} | x_{A_i^*})$$

for any $a, b \in S$ and $x \in \{G, B\}^{n+1}$

Back