Informative Counterfactuals

Adam Bjorndahl (CMU, Philosophy) & Todd Snider (Cornell, Linguistics)

SALT 25, Stanford

Carnegie Mellon University

Cognitive Science @ Cornell

Overview

- We use structural equation models (SEMs) to interpret counterfactuals
- SEMs model dependencies between events
- Formally, counterfactuals denote sets of such dependencies
- Intuitively, an SEM can be thought of as a possible *explanation* for a given counterfactual
- We classify such explanations into four categories

Counterfactuals

- We use counterfactuals to talk about things we know to be false

  (1) If the movie hadn't been so boring, I wouldn't have fallen asleep.

- And to talk about things we're uncertain about

  (2) If Sam were angry, Pat would have been angry, too. (But I don't know if she was.)

- Counterfactuals describe some relationship between the events
- There are many ways for two events to be related

  (3) If Alice had gone to the party, Bob would have stayed home.

- Does Bob try to avoid Alice?
  - Maybe he's shy
  - Maybe he doesn't like her
- Do other circumstances prevent them from attending parties together?
  - Maybe they're a couple on a tight budget
  - Maybe Bob is actually Alice in disguise
- Does Alice try to avoid Bob?
  - Unlike the other scenarios, this one doesn't seem to jive with (3)
- To understand a counterfactual, we have to capture this range of relationships

Modeling Relationships

- To capture relationships between events, we use *structured possible worlds* (Starr 2014)
- Worlds are event variables, their values, and **dependencies** between them
  - Just like truth values, we can use the (non)existence of dependencies to discriminate among worlds
- We model these dependencies using Structural Equation Models as formalized in Pearl 2000
- Nodes as events, arrows as dependencies

Rejecting Explanations

- There are many reasons to reject an explanation (including the simple direct dependency)
- It might contradict prior knowledge (encoded in the common ground)
- It might violate a law of good explanations
  - e.g., by positing an effect temporally prior to its cause
- Moreover, only some explanations are compatible with a given counterfactual
  - e.g., Alice avoiding Bob doesn't work, as we saw
  - But can be allowed with, e.g., the double aux construction

Key Contrast

We take the SEM not as a *given* but as a *goal*: rather than evaluating truth in a fully specified model, we think of counterfactual statements as *restricting the space of possible models*.

Graph as given

Start with a fully specified model

Intervene on the model

Evaluate truth in updated model

Hear B > H

Graph as goal

Start with a space of models

Consider possible explanations

Restrict possible models

Hear B > H

A Typology of Explanatory Strategies

Direct Cause

$B = \neg A$

- A simple direct dependency
- The 'default' assumption
- A is B's parent; B's value is set by A

Additional Cause

$B = \neg(A \wedge C)$

- Positing an additional causal factor
- A & B covary in the right C conditions
- Ex: Bob dislikes Alice

Common Cause

$A = C \qquad B = \neg C$

- Positing a shared cause
- No direct relation between A and B
- Ex: Coin flip to determine who attends

Intermediate Cause

$C = A \qquad B = \neg C$

- Positing a mediating factor
- A & B are related, but not directly
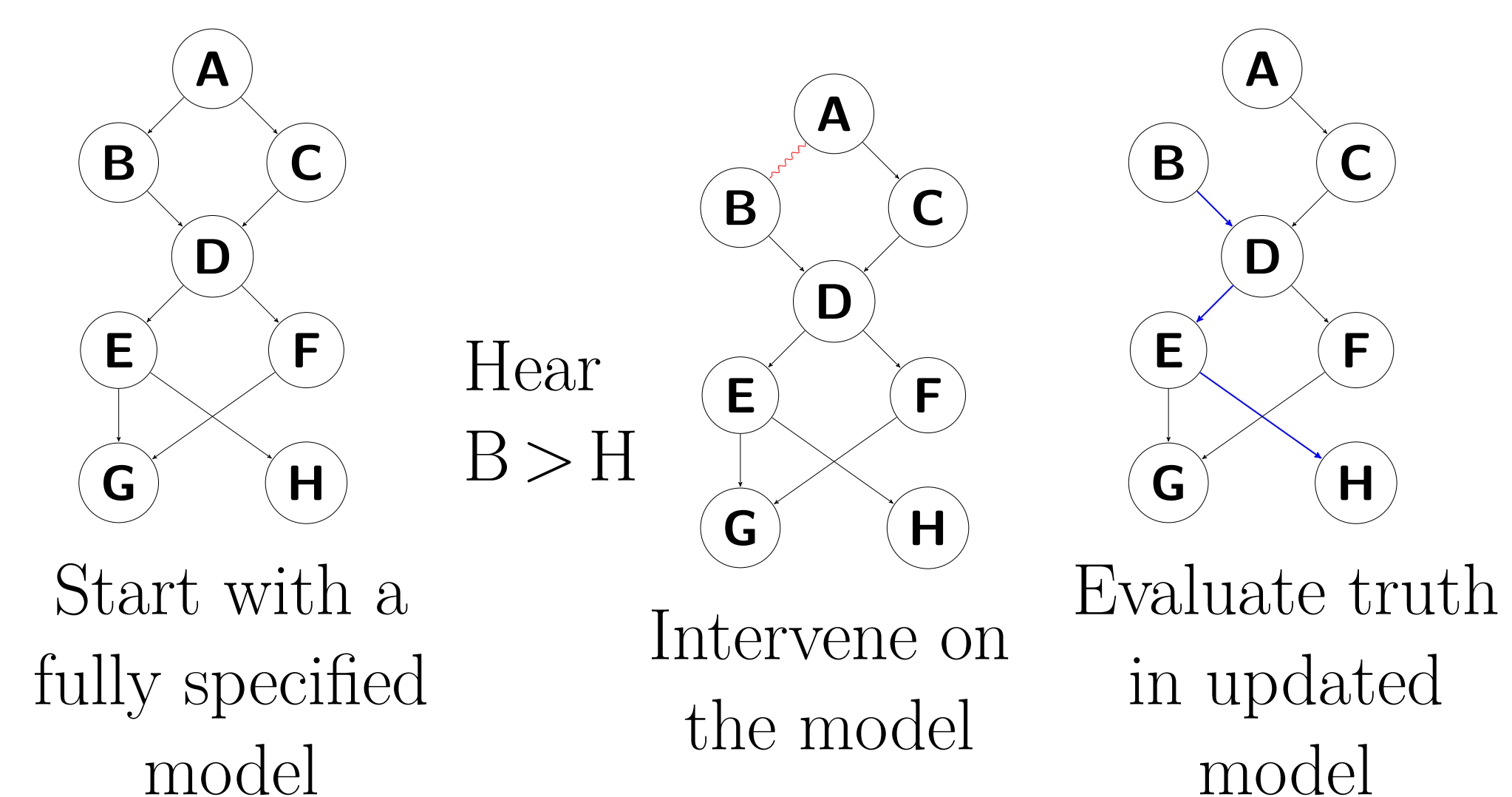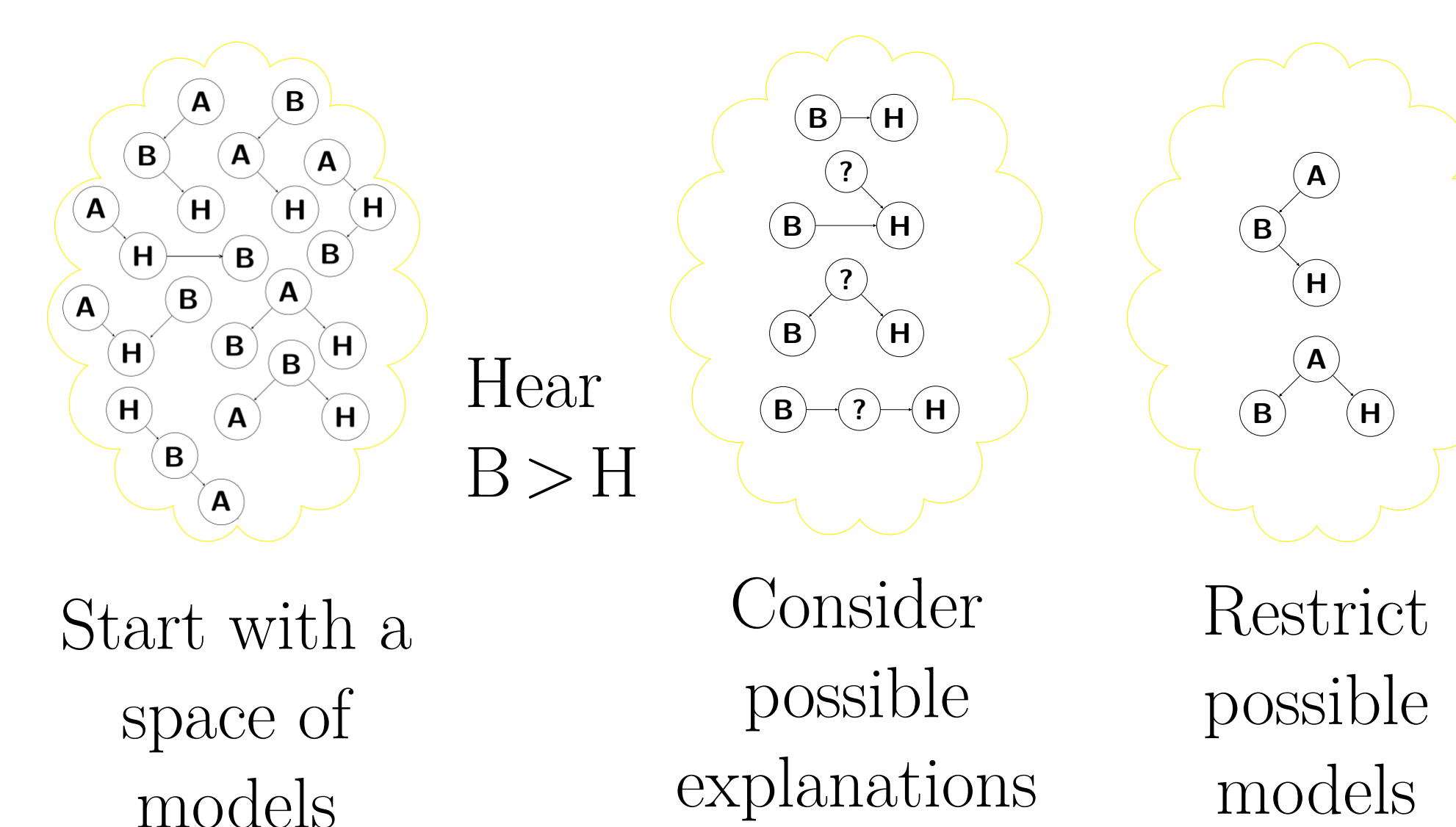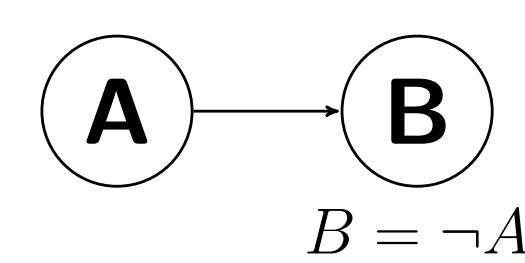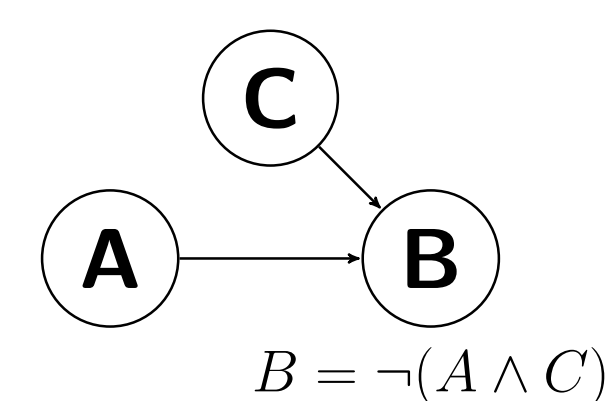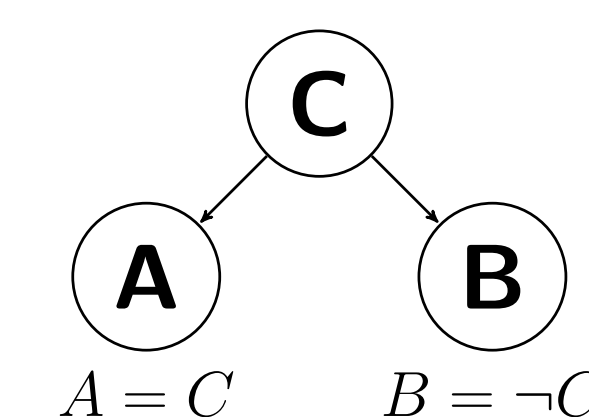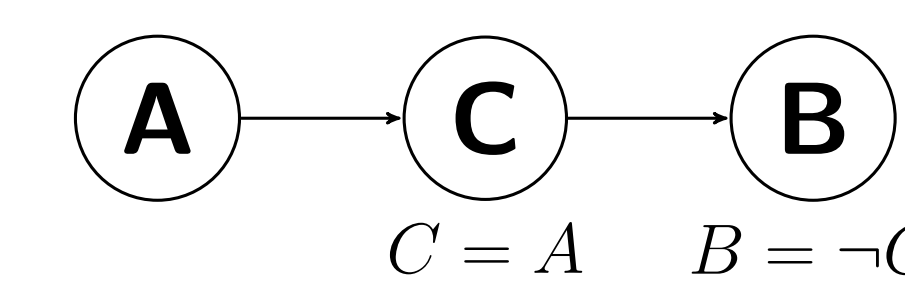- Ex: Bob is allergic to Alice's cat

Mutual Incompatibility

- Some counterfactuals which are individually felicitous are jointly infelicitous
- Consider a world where Alice and Bob are married, and live with their young son Doug

  (3) If Alice had gone to the party, Bob would have stayed home.

  (4) If Alice had gone to the party, Doug would have been home alone.

- Updating with (3) adds a covariance between A and $\neg B$ to the common ground
- Updating with (4) requires that A and B have the same value
- The models compatible with some explanation of (3) are not compatible with any explanation of (4)

Conclusion

- We can use structured possible worlds to model dependencies, and thus counterfactuals
- Doing so provides a natural way to typologize explanatory strategies
- Also yields insight into the mechanism that explains mutually infelicitous counterfactuals

Selected References

Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge Univ Press, 2000.

William B. Starr. Structured possible worlds. Ms. Cornell University, 2014.

Acknowledgements

Contact Information

- Web: http://conf.ling.cornell.edu/tsnider
- Email: tns35@cornell.edu