Minimal Intervention and Counterfactual Cognition

William B. Starr :: Assistant Professor, Philosophy :: Cornell University

will.starr@cornell.edu :: http://williamstarr.net

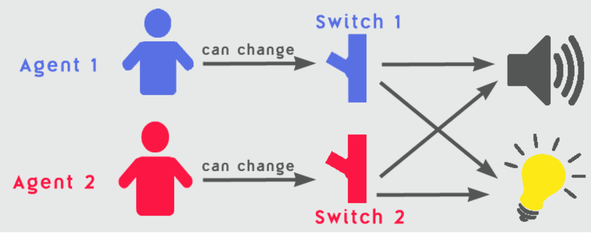
Society for Philosophy and Psychology :: June 2-4, 2016 :: UT Austin

OVERVIEW

- Analyze counterfactuals like 'if the switch had been up, the light would be on'
- Drawing on recent experimental literature and work on Bayesian Networks
- Highlight new examples and limitations of existing approaches

DATA

Context



Judgements

(1) *Switch 1 up, 2 down, light/music off*
If Switch 2 were up, the light would be on
- True (Sloman & Laganado 2005)

(2) *Switch 1 up, 2 up, light/music on*
If the light weren't on, the music wouldn't be on
- True (Hiddleston 2005, Rips 2010)

(3) *Switch 1 up, 2 up, light/music on*
If the light had failed, the music wouldn't be on
- False (Sloman & Laganado 2005)

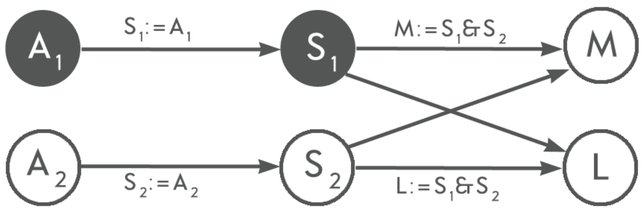
(4) If the light were off, then it would come on if both switches were flipped up
- True (Intuition to be tested)

BAYESIAN NETWORKS

Motivation: combinatorial explosion in probabilistic model
- *N* variables requires 2^N probability values

Solution: store only relations of conditional dependence
- If $P(A) \neq P(A|B) \neq P(A|{\sim}B)$ store $P(A|B)$ instead of each Boolean combination of *A* and *B* (Pearl 1993)
- Dependencies as graphs, edges nature of dependence

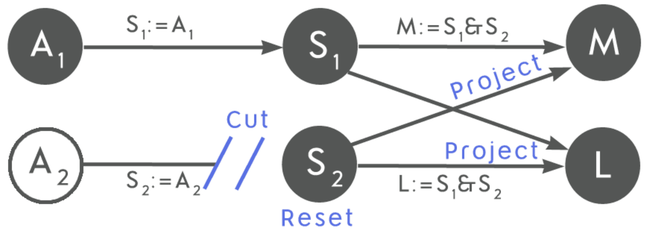
● True ○ False



INTERVENTIONIST THEORY

Basic View (Pearl 2000)
Evaluate 'if X had been true, Y would have been true' by cutting arrows into X, resetting it to True, and projecting consequences

If S_2 had been true:



Predictions

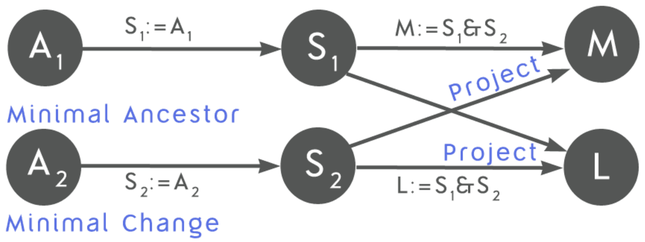
(1) True ✓ (2) False ✗ (3) False ✓ (4) False ✗

- Kemp & West (2015) modified interventionist account corrects (2) using 'node stability' but not (4)
- Other embeddings problematic (Briggs 2012)

MINIMAL NETWORK THEORY

Basic View (Hiddleston 2005)
To evaluate 'if X had been, Y would have been' find minimal ancestors of X, project each minimal change to them making X True

If S_2 had been true:



Predictions

(1) True ✓ (2) True ✓ (3) True ✗ (4) True ✓

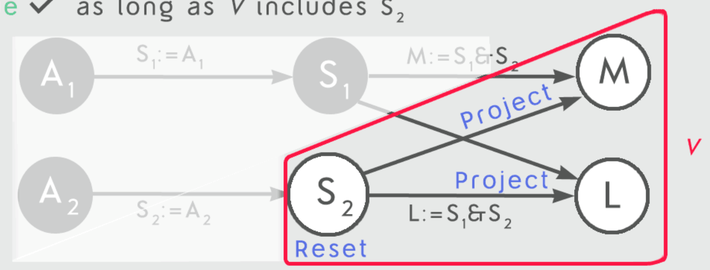
- Rips & Edwards (2013) modified MN theory corrects (3) using 'hypothesized cause node' for light's failure
  - Faces overgeneration issues
- **Bad Prediction**: both switches up, `If the light were off, either S_1 or S_2 would be up' (Kemp & West 2015)
- **Bad Prediction**: requires deepest `backtracking'

MINIMAL INTERVENTION THEORY

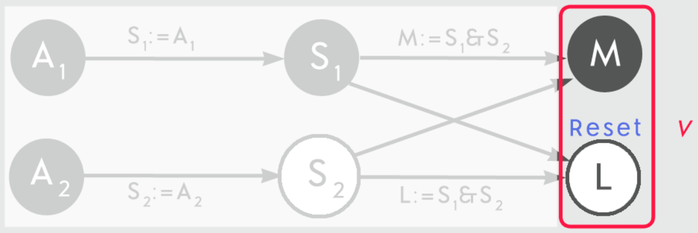
Proposed View
Evaluate 'if X had been, Y would have been' given salient variables *V* by finding minimal graph(s) covering *V*, minimal ancestors of X and projecting each setting of them making X True
- *V* is set of perceptually and linguisticially salient variables, and includes any variables in X and Y.

Predictions

(1) True ✓ regardless of *V*, since it must include S_2 and L
(2) True ✓ as long as *V* includes S_2



(3) False ✓ when *V* excludes S_1 and S_2
- Lexical semantics of `failed' excludes input to L



(4) True ✓ when *V* includes S_1 and S_2

CONCLUSION

- Proposed theory predicts (1)-(4) better than competitors
- (4) is important new example type, needs experimental confirmation with non-expert population
- Salience of variables does a lot of work, needs to be better empirically defined

Download: http://williamstarr.net/spp16.pdf

REFERENCES • Briggs, R (2012). 'Interventionist counterfactuals.' *Philosophical Studies*, 160(1) • Hiddleston (2005). 'A Causal Theory of Counterfactuals.' *Noûs*, 39(4). • Lucas & Kemp (2015). 'An Improved Probabilistic Account of Counterfactual Reasoning.' *Psychological Review*, 122(4) • Rips (2010). 'Two Causal Theories of Counterfactual Conditionals.' *Cognitive Science*, 34(2): 175–221. • Rips & Edwards (2013). 'Inference and Explanation in Counterfactual Reasoning.' *Cognitive Science*. • Sloman & Lagnado (2005). 'Do We "do"?' Cognitive Science, 29(1): 5–39. • Pearl (2000). *Causality*. CUP.