Structured Possible Worlds

Will Starr – Cornell University
will.starr@cornell.edu :: http://williamstarr.net

May 15, 2014

1 Plot Overview

- The Players:
  - Lewis-Stalnaker analysis of counterfactuals in terms of similarity
  - Analyses of causal counterfactuals in terms of structural equations/causal networks:
    - Intervention analysis: Pearl, Spirtes *et. al.*, Schulz, Briggs
    - Minimal network analysis: Hiddleston
- Critical points:
  - For similarity analysis:
    1. Problem of laws vs. matters of fact
    2. Communication/metasemantic issues
    3. Theoretical appropriateness concerns
  - For structural equations analyses:
    1. Counterexamples to interventionism; MP problem
    2. Counterexamples to minimal network analysis
- Positive points:
  1. Structural equations **can** provide an all-purpose semantics of counterfactual modality
     - Not just causal counterfactuals
     - Equations govern relations of dependence between facts
     - Compatible w/metaphysically light (Humean) and hefty spin on 'dependencies'
  2. Propose a structural equation semantics that integrates ideas of interventionist/minimal network theories
     - **But** addresses counterexamples to those analyses
  3. *Some* insights about counterlegals
  4. Gesture at the technical implementation of the proposal

2 Against Similarity

Lewis-Stalnaker Semantics
(Stalnaker 1968; Stalnaker & Thomason 1970; Lewis 1973)

- $\phi > \psi$ is true at w just in case all of the ϕ-worlds most **similar** to w are ψ-worlds
  - Most similar according to the selection function f
  - f takes a proposition p and a world w and returns the p-worlds most similar to w
- $[\![\phi > \psi]\!]_f = \{w \mid f(w, [\![\phi]\!]_f) \subseteq [\![\psi]\!]_f\}$

(Making the 'Limit Assumption': there are most similar worlds)

2.1 What is Similarity?

- Lewis (1973: §4.2) is quite clear: truth-conditions of counterfactuals are determined by *comparative overall similarity of possible worlds*

> Somehow, we *do* have a familiar notion of comparative similarity, even of comparative similarity of big, complicated, variegated things like whole people, whole cities, or even – I think – whole possible worlds. However mysterious that notion may be, if we can analyze counterfactuals in terms of it we will be left with one mystery in place of two.
> (Lewis 1973: 92)

> Even if we take the selection function as the basic primitive semantic determinant in the analysis of conditionals, we still must rely on some more or less independently understood notion of similarity of closeness of worlds to describe the intuitive basis on which the selection is made. The intuitive idea is something like this: the function selects a possible world in which the antecedent is true but which otherwise is as much like the actual world, in *relevant respects*, as possible. (Stalnaker 1984: 141)

- This is important to our understanding of the semantics
  - What makes counterfactuals true?
    - Comparative overall similarity of worlds
  - What are we estimating when we are judging counterfactuals?
    - Intuitive overall similarity of worlds

- Lewis is also clear that we should resist trying to define precise measures of similarity over structured worlds:

> It is tempting to try to define some exact measure of the similarity 'distance' among worlds, using the mathematical ersatz worlds introduced in Section 4.1... We must resist temptation. The exact measure thus defined cannot be expected to correspond well to our own opinions about comparative similarity. Some of the similarities and differences most important to us involved idiosyncratic, subtle, Gestalt properties. It is impossible in practice, and perhaps in principle, to express these respects of comparison in terms of the distribution of matter over space-time (or the like), even if the distribution of matter suffices to determine them. Consider a similar proposal to measure the visual similarity of faces...
> (Lewis 1973: 94-5)

- How well did this position stand the test of time? Not well:

> Sometimes a pair of counterfactuals of the following form seem true: *If A, the world would be very different; but if A and B, the world would not be very different.* Only if the similarity relation governing counterfactuals disagrees with that governing explicit judgments of what is 'very different' can such a pair be true... (I owe this argument to Pavel Tichy and, in a slightly different form, to Richard J. Hall.) It seems to me no surprise, given the instability even of explicit judgments of similarity, that two different comparative similarity relations should enter into the interpretation of a single sentence. (Lewis 1979: 466)

- Lewis' response:

> ...[W]e must use what we know about the truth and falsity of counterfactuals to see if we can find some sort of similarity relation – not necessarily the first one that springs to mind – that combines with [the similarity analysis] to yield the proper truth conditions. It is this combination that can be tested against our knowledge of counterfactuals, not [the similarity analysis] by itself. In looking for a combination that will stand up to the test, we must use what we know about counterfactuals to find out about the appropriate similarity relation – not the other way around.
> (Lewis 1979: 466-7)

- Lewis' final position:
  - *Not* our intuitive conception of overall similarity at work
  - Whatever concept of similarity is at work needs to be cooked up to fit our counterfactuals
  - Further, these similarity relations we cook up aren't going to be fully general.
    - ► There's going to be a heterogenous class of them
- Why did Lewis' position change?
  - Counterexamples from Fine, Tichý and others.
  - These show that intuitive overall comparative similarity of worlds is not what is at play in the evaluation of counterfactuals
  - Intuitively, a world where Nixon pushed the button and some mechanical failure prevented a nuclear holocaust is more similar to our own than one where a nuclear holocaust happens. Yet, it seems true that if Nixon had pushed the button, there would have been nuclear holocaust. (Fine 1975)
  - Tichý (1976: 271):

(1) a. Invariably, if it is raining, Jones wears his hat
b. If it is not raining, Jones wears his hat at random
c. Today, it is raining and so Jones is wearing his hat
d. But, even if it had not been raining, Jones would have been wearing his hat

  - Given (1a-c), (1d) is judged to be incoherent/false/bad
  - These can be handled by Lewis' **Standard Resolution** of similarity

- **Lewis' Standard Resolution**:

S1. First importance: avoid big, widespread, diverse violations of law. ('big miracles')

S2. Second importance: maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
  - Maximize exact match in matters of particular fact *before* antecedent turns out true

S3. Third importance: avoid even small, localized, simple violations of law. ('little miracles')

S4. Little or no importance: secure approximate similarity of particular fact, even in matters that concern us greatly.

- Essential for solving Fine/Tichý cases that S4 be ‘no importance’
    - After little miracle needed to make Nixon push button, don’t care about matching particular facts, e.g. lack of nuclear holocaust.
    - After little miracle needed to change the rain, don’t care about matching particular facts, e.g. Jones wearing his hat
- More theses, more problems:
    1. Particular matters of fact do, sometimes, matter
    2. Bleached of all intuitive content, one must wonder how speakers manage to resolve the similarity parameter in context
    3. Lewis’ new methodology suggests curve-fitting similarity to make particular counterfactuals true; two worries:
        - This makes the theory logically/semantically unexplanatory
        - Even if logically/semantically explanatory, it does not unify counterfactual reasoning with rational agency, imagination, causation, etc.

2.2 When Matters of Particular Fact Matter

- But as even Lewis notes, some cases go the other way:

(2) [You’re invited to bet heads on a coin-toss. You decline. The coin comes up heads.] See, if you had bet heads you would have won! (Slote 1978: 27 fn33; reporting Morgenbesser)

- Particular fact held fixed: toss outcome

(3) If we had bought one more artichoke this morning, we would have had one for everyone at dinner tonight (Sanford 1989: 173)

- Particular fact held fixed: number of dinner guests

(4) [t_1: switch 1 is up, switch 2 is down and the light is off. t_2: I flip switch 1 down (light remains off). t_3: I flip switch 2 up (light remains off).] If I hadn’t flipped switch 1 down, the light would be on. (My variant on Lifschitz’s example in Schulz 2007: 101)

- Particular fact held fixed: switch two is up

- These are a problem for S4, since they seem to make particular matters of fact after the antecedent event highly important
- Lewis (1979: 472): “I would like to know why”.

- Veltman (2005: 164) riffs on Tichý:

(5) a. Jones always flips a coin before he opens the curtain to see what the weather is like
b. If it’s not raining and the coin comes up heads, he wears his hat
c. If it’s not raining and the coin comes up tails, he doesn’t wear his hat
d. Invariably, if it is raining he wears his hat
e. Today, the coin came up heads and it is raining, so Jones is wearing his hat
f. But, even if it hadn’t been raining, Jones would have been wearing his hat

- Given (5a-e), (5f) seems right
- So why do we give up the fact that Jones is wearing his hat in (1) but not in (5)?
- Diagnosis by Veltman (2005: 164):

> Similarity of particular fact is important, but only for facts that do not **depend** on other facts. Facts stand and fall together. In making a counterfactual assumption, we are prepared to give up everything that depends on something that we must give up to maintain consistency. But, we want to keep in as many independent facts as we can.

- In (1), Jones wearing his hat depended on it raining, but the counterfactual made us give that up.
    - ► So we also give up the fact that he is wearing his hat
- In (5), Jones wearing his hat does not depend *just* on it raining
    - ► It depends also on the fact that the coin came up heads
    - ► Se we are not forced to give up the fact that he is wearing his hat when we give up the fact that it’s raining

- It is hard to see how to supplement the similarity analysis without appealing to dependence of some sort
- As I will discuss in §3, once one has **dependence** there’s no work left for similarity to do in an analysis of counterfactuals
- More motivation to talk in terms of dependence comes from counterexamples that violate S2

(6) [Lighting this firecracker leads to a simultaneous flash and bang. I didn't light it, so there was neither a flash nor bang.] If there had been a flash (just now), there would have been a bang (just now). (Hiddleston 2005: 644)

  - We don't hold fixed all facts leading up to the flash
  - We imagine the flash being brought about in the way made salient in the context

- Explanation in terms of dependence:
  - When A's dependence on B is salient, we imagine A being brought about by B
  - But if other facts also depend on B, then they will be brought about as well

2.3 Communication Problems

- **Problem 1** (Problem of Access):
  - It's unclear how agents like us could mutually fixate on particular values of f, since by Lewis' own lights f does not track any known intuitive concept of similarity
    - ► There are a LOT of different values of f! Assuming there are only 4 possible worlds, there'd still be 100s
  - It's unclear how agents like us express propositions, or even a coherent range of them, on a semantics which treats f as a contextual parameter
  - It's unclear how agents like us communicate and comprehend counterfactuals
- **Problem 2** (Problem of Informativity):
  - Strictly speaking, f (together with world of evaluation) contains *more* information than the truth of one counterfactual
  - Given a value of f and a world of evaluation, one knows the truth-conditions of *every* counterfactual
  - So once it is common ground which f is being used, it should be common ground which counterfactuals are true
  - But then uttering any counterfactual, on the similarity view, would seem redundant
  - Contrast with indexical I:
    - ► Need to know who speaker is to know what's communicated by *I'm tired*
    - ► But that knowing who speaker is isn't sufficient for deducing truth or falsity of sentence

2.4 Theoretical Appropriateness

- Counterfactuals are a essential tool for thinking about rational agency and scientific explanation
- It is not clear why similarity, even of the intuitive variety, would be at the heart of either phenomenon
- It is just not clear what contemporary picture of rational agency or scientific explanation one could hold that would exalt the extremely particularized, case-by-case, notion of similarity Lewis ends up with
- Worry: this is asking too much of a semantics!
  - Perhaps it is asking more than we should demand
  - But if another semantics offers it, it seems that would be some reason to prefer it
- What's our world like?
  - Does similarity matter?
  - Any important 'similarities' seem reducible to more basic notions, e.g. laws, dependence, etc.
  - Consider Loewer's (2007) proposal: counterfactuals ideally track probability that B occurs given A, the fundamental dynamical laws, and an initial boundary condition of the universe.
- What are we like?
  - Why would similarity matter to agents with limited resources, trying to bring about conditions favorable to them in an uncertain world?

3 Structural Equations

- How to proceed?
  - Conditional probability (Edgington 2004)
  - Dependence in premise semantics (Kratzer 1989; Veltman 2005)
  - Use explanation to constrain similarity (Kment 2006)
  - Introduce talk of 'states' and changes btwn them (Fine 2012)
- Each of these diverse strategies has limitations and all bear a striking resemblance to work on counterfactuals in causal models and structural equations (Spirtes *et al.* 1993, 2000; Pearl 2000, 2009)

> [...]the closest-world semantics still leaves two questions unanswered. (1) What choice of distance measure would make counterfactual reasoning compatible with ordinary conceptions of cause and effect? (2) What mental representation of inter world distances would render the computation of counterfactuals manageable and practical (for both humans and machines)? These two questions are answered by the structural model approach expanded in Chapter 7.
> (Pearl 2009: 35)

- Why structural equations are promising:
  1. Connects plausibly to accounts of explanation and causation (Woodward 2003; Hitchcock & Woodward 2003; Woodward & Hitchcock 2003; Halpern & Pearl 2005a,b)
     - For dissent: Hall (2007); Cartwright (2007)
  2. Emerged quite naturally from a large literature on rational agency and representation
     - E.g. Pearl (2000, 2009); Sloman (2005)
- I'll be focussing on the latter motivation

3.1 Rational Agency (Pearl 2000, 2009; Sloman 2005)

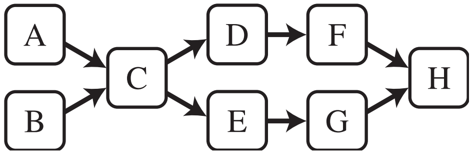
- Background picture: we are agents who can influence the world
  - Some states of affairs are better/worse for us than others
  - We have imperfect information about the world
  - And limited resources (time, energy, strength) to settle on and execute actions
- Uncertainty: Bayesian methods
  - That is, credences defined over a range of events (binary facts) satisfying Kolmogorov axioms; updated by conditionalization.
  - Suppose you have a system with 8 events in play: A, B, C, D, E, F, G, H.
  - A complete probabilistic description of this system is the joint distribution: $P(A, B, C, D, E, F, G, H)$
    - ▶ This requires estimating probability of all possible combinations of event outcomes:
      $P(A = 1, B = 1, C = 1, D = 1, E = 1, G = 1, H = 1) = n_0$,
      $P(A = 1, B = 1, C = 1, D = 1, E = 1, G = 1, H = 0) = n_1, \ldots$
    - ▶ This requires storing $2^8 - 1 (= 255)$ values, for just 8 events!

- Resource sensitivity: conditional (in)dependence
  - Agents need to compress joint probability distribution
  - Fortunately, probability of some variables depends on others
    - ▶ E.g. $P(B \mid A) \neq P(B)$, $P(D \mid C, B) \neq P(D)$

  Probabilistic Independence
  $P(A)$ is independent of $P(B)$ iff $P(A) = P(A \mid B) = P(A \mid \neg B)$

  Probabilistic Dependence
  $P(A)$ is dependent on $P(B)$ iff $P(A) \neq P(A \mid B) \neq P(A \mid \neg B)$

  - Pearl's discovery was that this affords a very compact representation of joint probabilities
  - $P(A, B, C, D, E, F, G, H) =$
    $P(A) \cdot P(B) \cdot P(C \mid A, B) \cdot P(D \mid C) \cdot P(E \mid C) \cdot P(F \mid D) \cdot P(G \mid E) \cdot P(H \mid F, G)$
    - ▶ Assuming A and B are independent of all other variables, and all other conditional probabilities register relations of conditional dependence
  - This reduces 255 to 18!
  - Pearl notes that chains of probabilistic dependence form a directed acyclic graph:



    - ▶ Arrows signify (sometimes joint) dependence
- Agency:
  - Conditional dependencies are not just *any* way of compactly representing joint probabilities
  - Leverage for an agent looking to influence the world
  - Knowing what A depends on is rather useful information for bringing it about when A is not under our direct influence
  - This invites the idea that agents reason by considering certain hypothetical changes to the values of nodes
    - ▶ And letting their effects flow through the graph
- This has inspired two analyses of counterfactuals in this framework
- **Interventionist Analysis** (Pearl, Schulz, Briggs)
  - To evaluate A > B, sever all links coming in to A, change it's value to 1 and let that change percolate through the graph

- **Minimal Network Analysis** (Hiddleston)
  - To evaluate A > B, find the smallest set of independent ancestors of A and check that each way of changing them which makes A true, also makes B true.
- To see how these theories work, and how it solves Lewis problem of particular facts, consider the switches and light case.

3.2 Intervention, Minimal Networks, Minimal Inverventions

3.2.1 Problems for Interventionism

- Consider *Flash* > *Bang*, given:

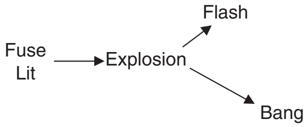


Figure 4. A cannon.

  - Surely: if there had been a flash, there would have been a bang
  - Hiddleston: this a problem for Lewis and interventionism
- Similarly, consider an example where TV 1 and TV 2 display pictures exactly when a broadcast tower is sending a signal. Right now, the tower isn't sending a signal, so neither TV is displaying a picture. But, it seems true that if TV 1 were displaying a picture right now, TV 2 would be displaying a picture.
- Here's an example without common causes:

(7) If the light were off and the switch were flipped, the light would turn on

  - An interventionist gives the wrong account here
  - They cut the influence of the switch on the light to make first antecedent true, so when the second antecedent changes the switch, it no longer influences the light!
- Briggs considers Pearl's executioner scenario (Fig.4) and the following:

(8) If A had fired, then (even) if the captain had not signaled, the prisoner would have died.

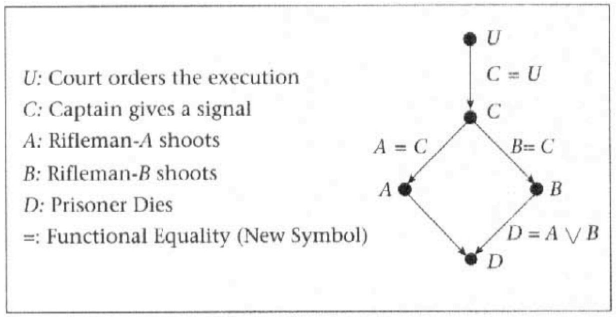

U: Court orders the execution
C: Captain gives a signal
A: Rifleman-A shoots
B: Rifleman-B shoots
D: Prisoner Dies
=: Functional Equality (New Symbol)
U
C = U
C
A = C
B= C
A
B
D = A ∨ B
D

Figure 4. Causal Models at Work (the Impatient Firing Squad).

  - Now suppose the court did give the order, the captain gave the signal, both *A* and *B* fired and the prisoner died.
  - The whole counterfactual (8) will be true, but the embedded one false, even though antecedent is true.
  - Evaluating (8) will first remove the arrow from C to A, setting A to 1, and then remove the link from U to C, setting C to 0.
  - Because A was set to 1 by intervention, D will still be 1.

3.2.2 Problems for Minimal Network

- Minimal network theory seems to give the right predictions
- But minimal network theory has it's own issues:
  - What about counterfactuals describing worlds extending indefinitely into the past and future
  - There is no minimal network that could make this true: *if the universe extends infinitely into the past, it is even older than I thought.*
- In general, it allows infinite backtracking:

(9) If I had been born in Scotland, the initial conditions of the universe would (have to) have been different

  - Perhaps (9) is true, but it is a manifestly odd thing to say without a lot of setup
  - Setup which makes salient the connection between those initial conditions and my birth
- Also, examples that force an intervention-style reading:

(10) (The light is off and the switch down.) If the light were miraculously on, the switch would be up.

3.2.3 Minimal Intervention (Starr)

- Here's a hypothesis that blends the approaches:
  - In every context, there is a salient set of atomic variables A
  - In evaluating a counterfactual, we assume any variables it contains are added to A
  - Further, we find the smallest subgraph of w that connects a maximal number of the variables in A
  - We then consider each minimal intervention to the top nodes that makes the antecedent true, and check that the consequent comes out true
- This will perform exactly like interventionism when $A = \{X, Y\}$ and you are intervening to evaluate $X > Y$ in a graph where $Y := X$
- It will perform exactly like minimal network theory when A includes *all* of the variables in the network
- Think of it as another step in making the structural equations analysis sensitive to our limited cognitive resources (in this case attention)

3.3 Structured Possible Worlds

- In standard propositional logics, atomic sentences are assigned independent truth-values
  - A valuation v, is a simple function from atomics to truth-values
    - ▶ E.g. $v(A, w) = 1, v(B, w) = 0, \ldots$
- Causal models give up this assumption
  - The truth-value of an atomic D can **depend** on the truth-value others A and B
- These dependencies are *functional*
  - If D depends only on A and B, then Ds' truth-value is uniquely determined by D and A
- D's truth depends on both A and C being true, $\mathsf{D} := \mathsf{A} \wedge \mathsf{C}$
  - Or D's truth depends on one of them being true, $\mathsf{D} := \mathsf{A} \vee \mathsf{C}$
- You can picture the models underlying these equations as directed graphs
- Starting point: classical possible worlds are valuations (situations are partial valuations)

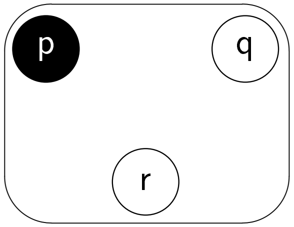


Figure 1: Classical possible world w

$$w(\mathsf{p}) = 0$$
$$w(\mathsf{q}) = 1$$
$$w(\mathsf{r}) = 1$$

Figure 2: System of equations for w

  - Worlds fix the truth values of each atomic sentence
  - Picture each atomic as a dot, which is black if false, white if true.
- Now depart from the classical picture:
  - The dependencies between facts endow worlds with a structure

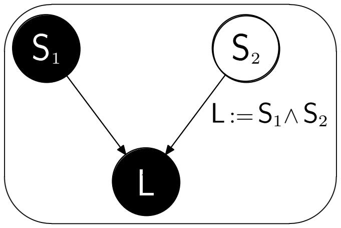


Figure 3: A structured possible world w

$$w(\mathsf{S}_1) = 0 \qquad (11)$$
$$w(\mathsf{S}_2) = 1 \qquad (12)$$
$$w(\mathsf{L}) = w(\mathsf{S}_1) \cdot w(\mathsf{S}_2) \qquad (13)$$
$$= 0$$

Figure 4: Equations for w

- We write our equations keeping in mind that ¬, ∧ and ∨ all have arithmetic counterparts operating on 1 and 0

$\neg$	$\wedge$	$\vee$
$1 - x$	$x \cdot y$	$(x + y) - (x \cdot y)$

- To evaluate the counterfactual $\mathsf{S}_1 > \mathsf{L}$, create world w_{S_1}
  - Step 1: intervention
    - ▶ Eliminate old assignment for S_1, line (14)
    - ▶ Make S_1 1, line (15)
  - Step 2: projection

- ▶ Apply equation (17) to solve for L
- ▶ New result: $w_{\mathsf{S}_1}(\mathsf{L}) = 1$!

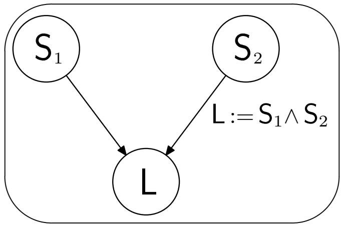


Figure 5: The New World w_{S_1}

$$w(\mathsf{S}_1)\cancel{=0} \tag{14}$$

$$w_{\mathsf{S}_1}(\mathsf{S}_1) = 1 \tag{15}$$

$$w_{\mathsf{S}_1}(\mathsf{S}_2) = w(\mathsf{S}_2) = 1 \tag{16}$$

$$w_{\mathsf{S}_1}(\mathsf{L}) = w(\mathsf{L}) = w_{\mathsf{S}_1}(\mathsf{S}_1) \cdot w_{\mathsf{S}_1}(\mathsf{S}_2) \tag{17}$$

$$= \cancel{0}$$

$$= 1 \tag{18}$$

Figure 6: Equations for w_{S_1}

- To see how this works better, consider a slightly modified scenario:
  - Switch 1 turns on a servo that controls switch 2: $\mathsf{S}_2 := \mathsf{S}_1$
  - Switch 2 turns on the light: $\mathsf{L} := \mathsf{S}_2$
  - Currently, switch 1 is up, so 2 is up and the light is on

(19) If switch 2 were up, the light would be off

- This comes out true, because after setting S_2 to 1, the equation connecting it will make L come out as 1 too
  - Lesson: you only keep fixed facts which are not determined by facts you are counterfactually giving up

3.4 What are Worlds?

- The idea spelled out for w:
  - Each independent atomic is mapped to a truth-value: $\{\langle \mathsf{S}_2, 1\rangle, \langle \mathsf{S}_1, 0\rangle\}$
  - Pair L with a dependency function d that determines the truth of L from a pairing of $\mathsf{S}_1, \mathsf{S}_2$ with truth-values:

S1	S2	L
1	1	1
1	0	0
1	0	0
1	1	0

  - d maps L and $\{\langle \mathsf{S}_1, 1\rangle, \langle \mathsf{S}_2, 1\rangle\}$ to 1
  - d maps L and $\{\langle \mathsf{S}_1, 1\rangle, \langle \mathsf{S}_2, 0\rangle\}$ to 0
  - d maps L and $\{\langle \mathsf{S}_1, 0\rangle, \langle \mathsf{S}_2, 1\rangle\}$ to 0
  - d maps L and $\{\langle \mathsf{S}_1, 0\rangle, \langle \mathsf{S}_2, 0\rangle\}$ to 0

- So, a world is a function from some independent atomics to truth-values (independent facts), together with a dependence function applying to each dependent atomic

- $w = \{\langle \mathsf{S}_2, 1\rangle, \langle \mathsf{S}_1, 0\rangle, d\}$
  - $d = \{ \quad \langle\, \langle \mathsf{L}, \{\langle \mathsf{S}_1, 1\rangle, \langle \mathsf{S}_2, 1\rangle\}\rangle, 1\rangle,$
    $\langle\, \langle \mathsf{L}, \{\langle \mathsf{S}_1, 1\rangle, \langle \mathsf{S}_2, 0\rangle\}\rangle, 0\rangle,$
    $\langle\, \langle \mathsf{L}, \{\langle \mathsf{S}_1, 0\rangle, \langle \mathsf{S}_2, 1\rangle\}\rangle, 0\rangle,$
    $\langle\, \langle \mathsf{L}, \{\langle \mathsf{S}_1, 0\rangle, \langle \mathsf{S}_2, 0\rangle\}\rangle, 0\rangle \quad \}$
- Dependence functions map pairs of atomics and situations (p, s) to truth-values
  - Particular dependencies can be identified with the sub-function that maps a single atomic and some situations to truth-values, e.g. there's only one dependency in d
- In three layer networks, dependence functions will have situations in their domain that are not part of the independent facts of the world
  - What's important is that those situations are determined by some other part of the dependence function and the independent facts

3.5 Complex Antecedents

- We'd like to define the notation w_ϕ for any non-counterfactual ϕ
- Let's first try just with compounds of atomics:
  - w_A is the world exactly like w except that it assigns A to 1
  - $w_{\neg \mathsf{A}}$ is the world exactly like w except that it assigns A to 0
  - $w_{\mathsf{A} \wedge \mathsf{B}}$ is the world exactly like w except that it assigns A, B to 1
  - $w_{\mathsf{A} \vee \mathsf{B}}$ is the world exactly like w except that it assigns ?????????
  - $w_{\neg\phi}$ is the world exactly like w except it assigns ??????
- An idea inspired by Dynamic Logic Harel *et al.* (2000):
  - Distinguish between ways of a formula being true, and ways of making it true
    - ▶ On analogy with: ways a machine can be when a program has run, versus ways a machine can run a program
  - Ways: transitions between worlds $\langle w, w'\rangle$
    - ▶ Transitions: thought of as minimal network intervention
  - A way of A being true: $\langle w, w\rangle$ where $w(A) = 1$
  - A way of making A true $\langle w, w'\rangle$ where $w(A) = 0$, $w'(A) = 1$ and w differs from w' by an minimal network intervention to make A true.

- Thus: $[\![p]\!]_A$ will be a set of $\langle w, w' \rangle$ where either $w = w'$ and $w(p) = 1$ or $w' \in w_p^A$ where w_p^A is the set of minimal network interventions on the subgraph connecting A to make p true.

- This generalizes to connectives:

Definition 1 (Semantics)

$$\begin{array}{lrcl}
(1) & [\![p]\!]_A & = & \{\langle w, w' \rangle \mid w = w' \text{ if } w(p) = 1 \;\&\; w' \in w_p^A \text{ if } w(p) = 0\} \\
(2) & [\![\neg\phi]\!]_A & = & \{\langle w', w \rangle \mid w = w' \;\&\; \langle w, w' \rangle \notin [\![\phi]\!]_A \text{ or } \langle w, w' \rangle \in [\![\phi]\!]_A\} \\
(3) & [\![\phi \wedge \psi]\!]_A & = & \{\langle w, w'' \rangle \mid \exists w' \colon \langle w, w' \rangle \in [\![\phi]\!]_A \;\&\; \langle w', w'' \rangle \in [\![\psi]\!]_A\} \\
(4) & [\![\phi \vee \psi]\!]_A & = & [\![\phi]\!]_A \cup [\![\psi]\!]_A \\
(5) & [\![\phi > \psi]\!]_A & = & \{\langle w, w \rangle \mid \langle w', w' \rangle \in [\![\psi]\!]_A \text{ if } \langle w, w' \rangle \in [\![\phi]\!]_A\} \\
(6) & [\![p := \phi]\!]_A & = & \ldots
\end{array}$$

Dependency Semantics for Counterfactuals

- $[\![\phi > \psi]\!] = \{w \mid w_\phi \subseteq [\![\psi]\!]\}$
- $\phi > \psi$ is true iff either ψ is independent of ϕ and true, or else ϕ is sufficient for bringing about ψ when holding fixed all those facts that do not depend upon ϕ.[1]
    - Entailment, truth, etc. defined classically?

3.6 Remaining Issues

- The problem of informativity is solved by putting structural equations 'in the world'
    - You are effectively ruling out similarity measures by ruling out worlds with certain structures
    - Only form of context sensitivity is the 'variables in play'
- Experimental work evaluating structural equation theories **???**
- What about counter-legals?
    - Consider: $(\mathsf{L} := \mathsf{S}_1 \vee \mathsf{S}_2) > \mathsf{L}$
        - ▶ $(\mathsf{L} := \mathsf{S}_1 \vee \mathsf{S}_2)$ denotes a dependency d
            - ▷ $d = \{ \; \langle \langle \mathsf{L}, \{\langle \mathsf{S}_1, 1\rangle, \langle \mathsf{S}_2, 1\rangle\}\rangle, 1\rangle,$
              $\langle \langle \mathsf{L}, \{\langle \mathsf{S}_1, 1\rangle, \langle \mathsf{S}_2, 0\rangle\}\rangle, 1\rangle,$
              $\langle \langle \mathsf{L}, \{\langle \mathsf{S}_1, 0\rangle, \langle \mathsf{S}_2, 1\rangle\}\rangle, 1\rangle,$
              $\langle \langle \mathsf{L}, \{\langle \mathsf{S}_1, 0\rangle, \langle \mathsf{S}_2, 0\rangle\}\rangle, 0\rangle \; \}$
    - We can then define $w_{A:=\phi}$ as: w with $\langle A, x \rangle$ removed and the dependency denoted by $A := \phi$ added in its place

[1] This intuitive paraphrase is from Cumming (2009: 1).

- Consider a case where one switch controls a light; the switch is up and the light on:
    - But if the light had been off, then if you had flipped the switch up, the light would have come on

A Logic of Structural Counterfactuals

Definition 2 (Situations) $S : A \mapsto \{1, 0\}$ where $A \subseteq \mathcal{A}t$

Definition 3 (Dependencies) $D : (A \times S) \mapsto \{1, 0\}$ where $A \subset \mathcal{A}t$

Definition 4 (Worlds)
$W = \{s \cup d \mid s \in S \;\&\; d \in D \;\&\; \operatorname{dom} s \cap \operatorname{dom} d = \emptyset \;\&\; \operatorname{dom} s \cap \operatorname{dom} d = \mathcal{A}t\}$

- Needed constraint: d is recursive

Definition 5 (Atomic Truth in a World)
$w(p) = s(p)$ if $s \in w$ and $p \in \operatorname{dom} s$. Otherwise $w(p) = d(p, s')$, where $d \in w$ and either $s = s'$ or s' is determined by d and s'

- s' is determined by d and s...

Definition 6 (Minimal Changes)
w_p^A is the set of worlds w' s.t. $w(p) = 1$ and $w = w'$ or g is the largest subgraph of w connecting every sentence in A, w' results from a minimal changes to the top nodes of g and $w'(p) = 1$

Definition 7 (Semantics)

$$\begin{array}{lrcl}
(1) & [\![p]\!]_A & = & \{\langle w, w' \rangle \mid w = w' \text{ if } w(p) = 1 \;\&\; w' \in w_p^A \text{ if } w(p) = 0\} \\
(2) & [\![\neg\phi]\!]_A & = & \{\langle w', w \rangle \mid w = w' \;\&\; \langle w, w' \rangle \notin [\![\phi]\!]_A \text{ or } \langle w, w' \rangle \in [\![\phi]\!]_A\} \\
(3) & [\![\phi \wedge \psi]\!]_A & = & \{\langle w, w'' \rangle \mid \exists w' \colon \langle w, w' \rangle \in [\![\phi]\!]_A \;\&\; \langle w', w'' \rangle \in [\![\psi]\!]_A\} \\
(4) & [\![\phi \vee \psi]\!]_A & = & [\![\phi]\!]_A \cup [\![\psi]\!]_A \\
(5) & [\![\phi > \psi]\!]_A & = & \{\langle w, w \rangle \mid \langle w', w' \rangle \in [\![\psi]\!]_A \text{ if } \langle w, w' \rangle \in [\![\phi]\!]_A\} \\
(6) & [\![p := \phi]\!]_A & = & \ldots
\end{array}$$

References

Briggs, R (2012). 'Interventionist counterfactuals.' *Philosophical Studies*, **160(1):** 139–166. URL http://dx.doi.org/10.1007/s11098-012-9908-5.

Cartwright, N (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics.* New York: Cambridge University Press.

Cumming, S (2009). 'On What Counterfactuals Depend.' Ms. UCLA.

Edgington, D (2004). 'Counterfactuals and the benefit of hindsight.' In PDP Noordhof (ed.), *Cause and Chance: Causation in an Indeterministic World*, 12–27. New York: Routledge.

Fine, K (1975). 'Review of Lewis' *Counterfactuals.*' *Mind*, **84**: 451–8.

Fine, K (2012). 'Counterfactuals Without Possible Worlds.' *Journal of Philosophy*, **109(3)**: 221-246.

Hall, N (2007). 'Structural equations and causation.' *Philosophical Studies*, **132(1)**: 109–136. URL http://dx.doi.org/10.1007/s11098-006-9057-9.

Halpern, J & Pearl, J (2005a). 'Causes and Explanations: A Structural-Model Approach. Part I: Causes.' *British Journal for Philosophy of Science*, **56**.

Halpern, J & Pearl, J (2005b). 'Causes and Explanations: A Structural-Model Approach. Part II: Explanations.' *British Journal for Philosophy of Science*, **56**: 889–911.

Harel, D, Kozen, D & Tiuryn, J (2000). *Dynamic Logic.* Cambridge, MA: MIT Press.

Hiddleston, E (2005). 'A Causal Theory of Counterfactuals.' *Noûs*, **39(4)**: 632–657. URL http://dx.doi.org/10.1111/j.0029-4624.2005.00542.x.

Hitchcock, C & Woodward, J (2003). 'Explanatory Generalizations, Part II: Plumbing Explanatory Depth.' *Noûs*, **37(2)**: 181–199. URL http://dx.doi.org/10.1111/1468-0068.00435.

Kment, B (2006). 'Counterfactuals and Explanation.' *Mind*, **115(458)**: 261–310. URL http://mind.oxfordjournals.org/content/115/458/261.

Kratzer, A (1989). 'An Investigation of the Lumps of Thought.' *Linguistics and Philosophy*, **12(5)**: 607–653.

Lewis, DK (1973). *Counterfactuals.* Cambridge, Massachusetts: Harvard University Press.

Lewis, DK (1979). 'Counterfactual Dependence and Time's Arrow.' *Noûs*, **13**: 455–476.

Loewer, B (2007). 'Counterfactuals and the Second Law.' In H Price & R Corry (eds.), *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*, 293–326. New York: Oxford University Press.

Pearl, J (2000). *Causality: Models, Reasoning, and Inference.* Cambridge, England: Cambridge University Press.

Pearl, J (2009). *Causality: Models, Reasoning, and Inference.* 2nd edn. Cambridge, England: Cambridge University Press.

Pollock, JL (1981). 'A Refined Theory of Conditionals.' *Journal of Philosophical Logic*, **10(2)**: 239–266.

Sanford, DH (1989). *If P then Q: Conditionals and the Foundations of Reasoning.* London: Routledge.

Schulz, K (2007). *Minimal Models in Semantics and Pragmatics: Free choice, Exhaustivity, and Conditionals.* Ph.D. thesis, University of Amsterdam: Institute for Logic, Language and Computation, Amsterdam. URL http://www.illc.uva.nl/Publications/Dissertations/DS-2007-04.text.pdf.

Sloman, S (2005). *Causal Models: How People Think About the World and Its Alternatives.* New York: OUP.

Slote, M (1978). 'Time in Counterfactuals.' *Philosophical Review*, **7(1)**: 3–27.

Spirtes, P, Glymour, C & Scheines, R (1993). *Causation, Prediction, and Search.* Berlin: Springer-Verlag.

Spirtes, P, Glymour, C & Scheines, R (2000). *Causation, Prediction, and Search.* 2 edn. Cambridge, Massachusetts: The MIT Press.

Stalnaker, R (1968). 'A Theory of Conditionals.' In N Rescher (ed.), *Studies in Logical Theory*, 98–112. Oxford: Basil Blackwell.

Stalnaker, RC (1984). *Inquiry.* Cambridge, MA: MIT Press.

Stalnaker, RC & Thomason, RH (1970). 'A Semantic Analysis of Conditional Logic.' *Theoria*, **36**: 23–42.

Tichý, P (1976). 'A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals.' *Philosophical Studies*, **29**: 271–273.

Veltman, F (2005). 'Making Counterfactual Assumptions.' *Journal of Semantics*, **22**: 159–180. URL http://staff.science.uva.nl/~veltman/papers/FVeltman-mca.pdf.

Woodward, J (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

Woodward, J & Hitchcock, C (2003). 'Explanatory Generalizations, Part I: A Counterfactual Account.' *Noûs*, **37(1)**: 1–24. URL http://dx.doi.org/10.1111/1468-0068.00426.