Counterfactuals

Marco Degano

Philosophical Logic 2025
24 & 26 November 2025

Readings

Suggested:

- Frank Veltman lecture notes on counterfactuals. `https://staff.fnwi.uva.nl/f.j.m.m.veltman/papers/Notes_Counterfactuals.pdf`
- SEP Entry 'Counterfactuals'
- Lecture notes: chapter 6

Plan

1. Data
2. Similarity Analyses
3. Correspondence Theory
4. Challenges
5. Premise Semantics
6. Truthmakers

Outline

Conditionals

Material conditional: $p \supset q$
true iff $\neg p \vee q$**.**

Indicative conditional: $p \rightarrow q$
If it rains, the sky is gray.

Counterfactual/subjunctive: $p \rightsquigarrow q$
If it had rained, the sky would have been gray.

Template for a counterfactual:

> *If it had been the case that* φ*, it would have been the case that* ψ*.*

What are counterfactuals?

We can take counterfactuals as conditionals about what *would* have been the case if the world were different (*often* contrary to fact).

Note however that we can felicitously use counterfactuals or subjunctive mood even when the antecedent is (or turns out) true, or is epistemically open. Can you think of some examples?

(1) If the glass had been dropped from two metres, it would have shattered.

(2) If this metal were heated to 1000°C, it would melt.

(3) If the maid had been upstairs at midnight, she would have heard the noise.

Counterfactuals vs. material implication

Counterfactuals $p \rightsquigarrow q$ *differ from* the material conditional $p \supset q$ in several ways:

- no trivial truth on false antecedents
- non-truth-functionality
- failure of monotonicity
- failure of contraposition and (sometimes) transitivity
- context/dependence on a similarity ordering

No triviality

For $p \supset q$, a false p makes the conditional true. If counterfactuals worked like that, most counterfactuals would be vacuously true.

(4) If the moon had been red, I would not exist.

No truth-functionality

Two conditionals can share the same (false) antecedent yet differ in truth value.

The antecedent of (5) and (6) is false, but the conditionals differ in plausibility:

(5) If I had put the heating on, the room would have been warm.

(6) If I had put the heating on, the room would have exploded.

No monotonicity / strengthening the antecedent

From $\varphi \rightsquigarrow \psi$ it need not follow that $(\varphi \wedge \chi) \rightsquigarrow \psi$.

(7) a. If I had put sugar in my coffee, it would have tasted better. $\varphi \rightsquigarrow \psi$

b. If I had put sugar *and diesel oil* in my coffee, it would have tasted better. $(\varphi \wedge \chi) \rightsquigarrow \psi$

No contraposition

Generally, $\varphi \rightsquigarrow \psi$ does *not* entail $\neg\psi \rightsquigarrow \neg\varphi$.

Suppose two wolves (A, B) attacked a sheep yesterday.

(8) a. If wolf A had not been around, the sheep would have (still) been killed. $\varphi \rightsquigarrow \psi$

b. If the sheep had not been killed, then wolf A would have been around. $\neg\psi \rightsquigarrow \neg\varphi$

No transitivity

Transitivity can fail (from Sider 2016).

(9) a. If I hadn't been born, Mike would have been my parents' oldest child. $\varphi \rightsquigarrow \psi$

b. If my parents had never met, I wouldn't have been born. $\chi \rightsquigarrow \varphi$

c. If my parents had never met, Mike would have been my parents' oldest child. $\chi \rightsquigarrow \psi$

Context-dependence / ambiguity

(10) If Amsterdam had been Rome, the weather would have been better.

Possible readings:

- Amsterdam located where Rome is,
- Amsterdam (in the Netherlands) simply named “Rome,”
- Amsterdam as capital of the Roman Empire,
- …

Different similarity assumptions yield different truth conditions.

The First Lewis (strict implication)

C. I. Lewis (1883–1964)

C. I. Lewis (1912) proposed analyzing counterfactuals as *strict* conditionals.

(11) a. If I had put the heating on, the room would be warm.
b. Analyze as $\Box(\varphi \supset \psi)$
c. $M, w \models \Box(\varphi \supset \psi)$ iff $\forall w' \, wRw' \Rightarrow M, w' \models \varphi \supset \psi)$

Strict conditionals: what they (wrongly) predict

With K and necessitation, many of the previous facts are not met.

- **Monotonicity** (strengthening the antecedent)

$$\Box(\varphi \supset \psi) \models \Box((\varphi \wedge \chi) \supset \psi)$$

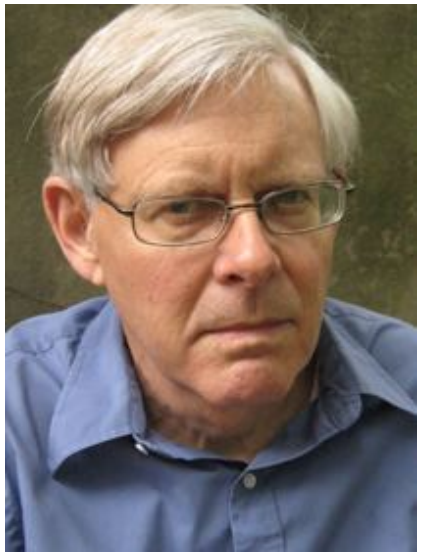
- **Contraposition**

$$\Box(\varphi \supset \psi) \models \Box(\neg\psi \supset \neg\varphi)$$

- **Transitivity**
  from $\Box(\varphi \supset \psi)$ and $\Box(\psi \supset \chi)$ derive $\Box(\varphi \supset \chi)$.
- **Vacuity** (relative to accessibility)
  If no accessible w' with wRw' satisfies φ, then every such w' satisfies $(\varphi \supset \psi)$. Hence $\Box(\varphi \supset \psi)$.

Outline

The Second Lewis and Stalnaker



Robert Stalnaker

David Lewis (1941–2001)

Stalnaker (1968) and D. Lewis (1973) analyze counterfactuals via *similarity* to the actual world.

If it had been the case that φ, it would have been the case that ψ

We evaluate ψ at the φ-world(s) **most similar** to the actual world w.

Language

We work with a propositional language $\mathcal{L}(\rightsquigarrow)$ with a binary connective $\rightsquigarrow$ for counterfactuals..

$$\varphi \ ::= \ p \ \mid \ \neg\varphi \ \mid \ (\varphi \wedge \varphi) \ \mid \ (\varphi \vee \varphi) \ \mid \ (\varphi \supset \varphi) \ \mid \ (\varphi \rightsquigarrow \varphi)$$

Frames

We interpret formulas on *similarity frames* $F = \langle W, \prec \rangle$, where:

- $W \neq \emptyset$ is a set of possible worlds.
- $\prec$ assigns to each $w \in W$ a strict partial order $\prec_w$ on some subset $W_w \subseteq W$.
- Read $u \prec_w v$ as: u *is more similar to* w *than* v *is*.
- We write $x \preceq_w y$ for the *reflexive closure* of $\prec_w$: $x \preceq_w y$ iff $x \prec_w y$ or $x = y$.

Similarity relation

- For each $w \in W$, $\prec_w$ is a *strict partial order* (transitive, irreflexive, hence asymmetric).
- The *field* W_w ("accessible" worlds for w) is the domain on which $\prec_w$ is defined.

$W_w = \{x \in W : \exists y \in W\,(x \prec_w y)\} \cup \{y \in W : \exists x \in W\,(x \prec_w y)\} =$
$\{u \in W : \exists v \in W\,(\,u \prec_w v \text{ or } v \prec_w u\,)\}$

(We will later relate structural constraints on $\prec_w$ to axioms for $\rightsquigarrow$.)

Models

A model $M = \langle W, \prec, V \rangle$, where $\langle W, \prec \rangle$ is a frame and V assigns truth values to atomic sentences at each world.

$M, w \models \varphi$ means: φ is true at world w (in M).

$[\![\varphi]\!]_M := \{w \in W \mid M, w \models \varphi\}$ is the proposition expressed by φ in M.

Semantic Clauses (standard)

$$
\begin{array}{lll}
M, w \models p & \text{iff} & V(p, w) = 1 \\
M, w \models \neg\varphi & \text{iff} & M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \text{iff} & M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models \varphi \vee \psi & \text{iff} & M, w \models \varphi \text{ or } M, w \models \psi \\
M, w \models \varphi \supset \psi & \text{iff} & M, w \not\models \varphi \text{ or } M, w \models \psi
\end{array}
$$

We use $\subset\supset$ for the material biconditional.

Definition (Logical consequence)

$\Gamma \models \varphi$ iff for every model M and every world w of M, if $M, w \models \gamma$ for each $\gamma \in \Gamma$, then $M, w \models \varphi$.

Semantic Clauses ($\rightsquigarrow$)

Definition (Counterfactual: general clause)

$$M, w \models \varphi \rightsquigarrow \psi \iff \forall u \in W_w \cap [\![\varphi]\!] \ \exists u' \in [\![\varphi]\!] \text{ such that}$$
$$\text{(i)} \quad u' \preceq_w u$$
$$\text{(ii)} \quad \forall u'' \in [\![\varphi]\!] \ (u'' \preceq_w u' \Rightarrow M, u'' \models \psi)$$

For every accessible φ-world u (relative to w), there is a φ-world u' at least as similar to w as u such that *every* φ-world u'' that is at least as similar to w as u' makes ψ true.

No matter how close to w you go among the φ-worlds, you never encounter a φ-world where ψ is false.

Simple case for $p \rightsquigarrow q$ at w

$$M, w \models \varphi \rightsquigarrow \psi \text{ iff } \forall u \in W_w \cap [\![\varphi]\!] \ \exists u' \in [\![\varphi]\!] \text{ such that } u' \preceq_w u \text{ and } \forall u'' \in [\![\varphi]\!] \ (u'' \preceq_w u' \Rightarrow M, u'' \models \psi).$$

- $W = \{w, a, b, c, d, e, f, g\}$
- $a \prec_w c$, $b \prec_w c$, $c \prec_w d$, $e \prec_w a$, with transitive closure
- p true only at a, b, c, d; q true only at a, b, c

We show $M, w \models p \rightsquigarrow q$ by checking each $u \in W_w \cap [\![p]\!] = \{a, b, c, d\}$. **(blackboard)**

a and b are the p-closest worlds. When checking for c and d, we picked $u' = a$ (or b) and checked that the consequent holds in that world.

So why not considering only the p-closest worlds?

Why not considering only the closest worlds

$$M, w \models \varphi \rightsquigarrow \psi \text{ iff } \forall u \in W_w \cap [\![\varphi]\!] \ : \ \exists u' \in [\![\varphi]\!] \text{ such that } u' \preceq_w u \text{ and } \forall u'' \in [\![\varphi]\!] \ (u'' \preceq_w u' \Rightarrow M, u'' \models \psi).$$

$W = \{w\} \cup \{u_n : n \in \mathbb{N}_{\geq 1}\}, \quad W_w = \{u_n : n \geq 1\}, \quad u_{n+1} \prec_w u_n$ for all n.

For all $n \geq 1$: $M, u_n \models p$ and $M, u_n \models q$ (and $M, w \models \neg p, \neg q$).

$M, w \models p \rightsquigarrow q$ *even though there is no closest* p*-world*.

Fix any $u = u_k \in W_w \cap [\![p]\!]$. Let $u' = u_k$. Then every $u'' \in [\![p]\!]$ with $u'' \preceq_w u'$ is some u_m with $m \geq k$, and $M, u_m \models q$. Thus the clause holds for each u_k, so $M, w \models p \rightsquigarrow q$.

A “closest-worlds” clause would be undefined here: the full clause still yields the right verdict.

The Limit Assumption

Definition (Limit Assumption)

For every $w \in W$, the relation $\prec_w$ on W_w is *well-founded*.

Equivalent characterizations:

- *No infinite descent:* there is no sequence $u_1, u_2, \cdots \in W_w$ with $u_{n+1} \prec_w u_n$ for all n.
- *Minimal elements:* every nonempty $X \subseteq W_w$ has a $\prec_w$-minimal element.

For $X \subseteq W_w$, let $\mathrm{Min}_w(X) := \{u \in X : \neg\exists v \in X\ (v \prec_w u)\}$

We write $\mathrm{Min}_w(\varphi)$ for $\mathrm{Min}_w(W_w \cap [\![\varphi]\!])$.

Semantic clause with the Limit Assumption

Under the limit assumption, the counterfactual reduces to a "closest-worlds" clause:

Definition (Counterfactual: closest-worlds / limit clause)

$$M, w \models \varphi \rightsquigarrow \psi \quad \Longleftrightarrow \quad \forall u \in Min_w(\varphi)\ \ M, u \models \psi,$$

where $Min_w(\varphi) := \{u \in W_w \cap [\![\varphi]\!] : \neg\exists v \in W_w \cap [\![\varphi]\!]\ (v \prec_w u)\}$.

- If $\mathrm{Min}_w(\varphi)$ has multiple members, *all* must satisfy ψ.
- If $W_w \cap [\![\varphi]\!] = \varnothing$, then $\mathrm{Min}_w(\varphi) = \varnothing$ and the universal condition is vacuously true.

Lewis rejects assuming well-foundedness in general. Why?

Lewis's "longer than one inch" example

"If the line had been longer than one inch, it would have been one hundred miles long."

Consider $\varphi =$ "the line is > 1 inch." For any $\epsilon > 0$, there is a φ-world where the line is $1 + \epsilon$, and one even *closer* with length $1 + \epsilon/2$.

$$\cdots \prec_w u_{1/4} \prec_w u_{1/2} \prec_w u_1, \quad \text{where } u_\epsilon \models (\text{length } = 1 + \epsilon).$$

There is *no* $\prec_w$-minimal φ-world: an infinite descending chain

An example

$$(p \rightsquigarrow q) \wedge (p \rightsquigarrow r) \models (p \wedge q) \rightsquigarrow r$$

Fix w and assume $M, w \models p \rightsquigarrow q$ and $M, w \models p \rightsquigarrow r$.

- If $W_w \cap [\![p]\!] = \varnothing$, then also $W_w \cap [\![p \wedge q]\!] = \varnothing$, so $(p \wedge q) \rightsquigarrow r$ holds vacuously.
- Otherwise, let $u \in Min_w(p \wedge q)$ be arbitrary. We show $u \in Min_w(p)$.
  - Suppose not. By well-foundedness there is $v \in Min_w(p)$ with $v \prec_w u$.
  - Since $p \rightsquigarrow q$, every $x \in Min_w(p)$ satisfies q. In particular $v \models q$.
  - Hence $v \in W_w \cap [\![p \wedge q]\!]$ and $v \prec_w u$, contradicting $u \in Min_w(p \wedge q)$.
- Thus $u \in Min_w(p)$.
- From $p \rightsquigarrow r$ we have: every $x \in Min_w(p)$ satisfies r. In particular $u \models r$.

Failure of monotonicity: a countermodel

$\not\models \ (p \rightsquigarrow r) \ \supset \ ((p \wedge q) \rightsquigarrow r)$

$W = \{w, a, b\}$, $W_w = \{a, b\}$ with $a \prec_w b$.

$$M, a \models p \wedge \neg q \wedge r \qquad M, b \models p \wedge q \wedge \neg r \qquad M, w \models \neg p$$

Then

$$\mathrm{Min}_w(p) = \{a\} \quad \Rightarrow \quad M, w \models p \rightsquigarrow r$$

but

$$\mathrm{Min}_w(p \wedge q) = \{b\} \quad \Rightarrow \quad M, w \not\models (p \wedge q) \rightsquigarrow r$$

Hence $M, w \models p \rightsquigarrow r$ but $M, w \not\models (p \wedge q) \rightsquigarrow r$.

Axiomatization: system $\mathbf{P}$

Axioms For all formulas φ, ψ, χ:

TAUT	Every propositional tautology.
CI	$\varphi \rightsquigarrow \varphi$
CC	$(\varphi \rightsquigarrow \psi) \wedge (\varphi \rightsquigarrow \chi) \supset (\varphi \rightsquigarrow (\psi \wedge \chi))$
CW	$(\varphi \rightsquigarrow \psi) \supset (\varphi \rightsquigarrow (\psi \vee \chi))$
ASC	$(\varphi \rightsquigarrow \psi) \wedge (\varphi \rightsquigarrow \chi) \supset ((\varphi \wedge \psi) \rightsquigarrow \chi)$
AD	$(\varphi \rightsquigarrow \chi) \wedge (\psi \rightsquigarrow \chi) \supset ((\varphi \vee \psi) \rightsquigarrow \chi)$

Rules

MP	If $\vdash \varphi$ and $\vdash \varphi \supset \psi$, then $\vdash \psi$.
REA	If $\vdash \varphi \subset\supset \psi$, then $\vdash (\varphi \rightsquigarrow \chi) \subset\supset (\psi \rightsquigarrow \chi)$
REC	If $\vdash \varphi \subset\supset \psi$, then $\vdash (\chi \rightsquigarrow \varphi) \subset\supset (\chi \rightsquigarrow \psi)$

$\mathbf{P}$ is *sound and complete* for the similarity semantics of $\rightsquigarrow$

Exercise

A. Show the axioms and rules of $\mathbf{P}$ are valid on all similarity frames.

B. Show these schemas are <u>not</u> valid (build countermodels):

Monotonicity:	$(\varphi \rightsquigarrow \psi) \supset ((\varphi \wedge \chi) \rightsquigarrow \psi)$
Contraposition:	$(\varphi \rightsquigarrow \psi) \supset (\neg\psi \rightsquigarrow \neg\varphi)$
Transitivity:	$[(\varphi \rightsquigarrow \psi) \wedge (\psi \rightsquigarrow \chi)] \supset (\varphi \rightsquigarrow \chi)$
Vacuity (triviality):	$\neg\varphi \supset (\varphi \rightsquigarrow \psi)$

The relevance of the Limit Assumption

Does adding the *Limit Assumption* (well-founded $\prec_w$) change the logic?

Two classes of frames:

$\mathcal{K}_{\text{all}} = \{\text{all similarity frames}\}$
$\mathcal{K}_{\text{lim}} = \{\text{frames with the Limit Assumption}\}$

- **Finite consequence:** *no difference*[1]
  If Γ is finite, $\quad \Gamma \models_{\mathcal{K}_{\text{all}}} \varphi \iff \Gamma \models_{\mathcal{K}_{\text{lim}}} \varphi$
- But it does matter with **infinitely many premises**.
  Compactness fails over $\mathcal{K}_{\text{lim}}$: there exists infinite Δ such that every finite $\Gamma \subseteq \Delta$ is satisfiable on a limit frame, but Δ itself is not.

[1] This invariance holds for the plain class of similarity frames. However, adding *additional* structural constraints can change the finitary logic when combined with the Limit Assumption.

Compactness and limit assumption

Let $p_1, p_2, \ldots$ be pairwise distinct atoms. We write

$$\alpha_k \;:=\; \bigvee_{i=1}^{k} p_i \qquad (k \geq 1)$$

We define, for each $k \geq 1$,

$$\varphi_k \;:=\; (\alpha_{k+1} \rightsquigarrow \neg\alpha_k) \qquad \text{and} \qquad \psi_k \;:=\; \neg(\alpha_{k+1} \rightsquigarrow \alpha_k).$$

Let $\Delta \;:=\; \{\varphi_k,\ \psi_k : k \geq 1\}$.

Δ is satisfiable *iff* the Limit Assumption fails.

Main facts to prove

For a fixed world w. If the limit assumption holds at w (i.e. $\prec_w$ is well-founded on W_w):

1. Δ is **not** satisfiable at w.[2]
2. Every finite $\Sigma \subset \Delta$ is satisfiable at w.

(On blackboard)

Corollary (Compactness failure under the Limit Assumption)

Every finite $\Sigma \subset \Delta$ is satisfiable (even with well-founded $\prec_w$), but Δ is not. Hence compactness fails.

[2] A formula is satisfiable if there is a model $M = \langle W, \prec, V \rangle$ and world $w \in W$ s.t. $M, w \models \varphi$. Likewise a set Δ is satisfiable if there is a model $M = \langle W, \prec, V \rangle$ and world $w \in W$ s.t. for all $\delta \in \Delta$, $M, w \models \delta$. In the slides, we say satisfiable at w to make the link with well-foundness of $\prec_w$.

Exercise

Show that Δ is satisfiable in the class of all similarity frames. For a fixed world w. If the limit assumption fails at w (i.e. there is an infinite descending $\prec_w$-chain in W_w), then there is a model M with $M, w \models \delta$ for each $\delta \in \Delta$.

Exercise

If you have taken modal logic, show the following:

Let $\mathcal{K}_{all}$ be all similarity frames and $\mathcal{K}_{lim}$ those satisfying the Limit Assumption. Show that for any *finite* Γ and any φ,

$$\Gamma \models_{\mathcal{K}_{all}} \varphi \quad \Longleftrightarrow \quad \Gamma \models_{\mathcal{K}_{lim}} \varphi$$

(Define the the *counterfactual depth* of a formula φ as the maximal nesting of $\rightsquigarrow$ in φ, and apply 'depth-bounded' filtration to obtain a well-founded companion model.)

Outline

Correspondence: frame conditions vs. axioms

Relate *structural* properties of the similarity ordering $\prec$ to *valid* counterfactual principles.

Constraints on $\prec$ $\iff$ Valid principles for $\rightsquigarrow$

Weak Centering

Weak Centering

$$\forall w \in W: \quad w \in W_w \text{ and } \neg\exists v \in W_w\ (v \prec_w w).$$

Nothing is strictly closer to w than w.

$\mathrm{MP}^{\rightsquigarrow}$

$$(\varphi \rightsquigarrow \psi) \ \wedge \ \varphi \ \models \ \psi$$

WC $\Rightarrow \mathrm{MP}^{\rightsquigarrow}$

$$M, w \models \varphi \rightsquigarrow \psi \iff \forall u \in W_w \cap [\![\varphi]\!]\, \exists u' \in [\![\varphi]\!]\, (u' \preceq_w u \wedge \forall u'' \in [\![\varphi]\!]\, (u'' \preceq_w u' \Rightarrow M, u'' \models \psi)),$$

where $x \preceq_w y$ abbreviates $x \prec_w y$ or $x = y$.

- Since WC gives $w \in W_w$, and $M, w \models \varphi$, we have $w \in W_w \cap [\![\varphi]\!]$.
- Take $u := w$ to get some $u' \in [\![\varphi]\!]$ such that $u' \preceq_w w$ and $\forall u'' \in [\![\varphi]\!]\, (u'' \preceq_w u' \Rightarrow M, u'' \models \psi)$.
- By WC there is no $v \prec_w w$, hence $u' = w$. Taking $u'' := w$ ($w \preceq_w w$) yields $M, w \models \psi$.

$\mathrm{MP}^{\rightsquigarrow} \Rightarrow$ WC

We prove the contrapositive.

Case 1: $w \notin W_w$.

- Pick p, q with $M, w \models p \wedge \neg q$ and $M, u \models \neg p$ for all $u \in W_w$.
- Then $W_w \cap [\![p]\!] = \varnothing$, so $M, w \models p \rightsquigarrow q$ vacuously.
- Hence $M, w \models (p \rightsquigarrow q) \wedge p$ but $M, w \not\models q$.

Case 2: $\exists v \in W_w$ with $v \prec_w w$.

- Pick p, q and set $M, w \models p \wedge \neg q$, $M, v \models p \wedge q$, and $M, u \models \neg p$ for all $u \notin \{w, v\}$.
- Then the only accessible p-world(s) are v (and possibly w). For $u = v$ choose $u' = v$. For $u = w$ (if $w \in W_w$) choose $u' = v$ since $v \preceq_w w$.
- In both subcases the clause for $p \rightsquigarrow q$ is satisfied at w, so $M, w \models p \rightsquigarrow q$, yet $M, w \models p \wedge \neg q$.

Strong Centering

Strong Centering

$$\forall w \in W: \quad w \in W_w \text{ and } \forall v \in W_w \setminus \{w\}: \; w \prec_w v$$

(In words: w is the *unique* closest world to itself; no ties at w.)

Conjunctive Sufficiency

$$(\varphi \wedge \psi) \rightarrow (\varphi \rightsquigarrow \psi)$$

If φ (and ψ) already hold at w, then the closest φ-world(s) are just w itself. Hence the counterfactual $\varphi \rightsquigarrow \psi$ is forced.

Uniqueness (Connectedness + Limit Assumption)

Connectedness (totality) at w

$$\forall u, v \in W_w \; (u \neq v \;\Rightarrow\; u \prec_w v \;\text{ or }\; v \prec_w u)$$

(No two distinct worlds are incomparable: $\prec_w$ is a strict *total* order on W_w.)

CEM

$$(\varphi \rightsquigarrow \psi) \quad \vee \quad (\varphi \rightsquigarrow \neg\psi)$$

Given an antecedent φ, there is a unique closest φ-world, so ψ is either true there or its negation is.

Disjunction without commitment (Bizet and Verdi)

$C :=$ Compatriots(Bizet, Verdi) $\quad I :=$ Italian(Bizet)
$F :=$ French(Verdi)

(12) If Bizet and Verdi had been compatriots, Bizet would have been Italian. $C \rightsquigarrow I$

(13) If Bizet and Verdi had been compatriots, Verdi would have been French. $C \rightsquigarrow F$

(14) If Bizet and Verdi had been compatriots, *either* Verdi would have been French *or* Bizet would have been Italian. $C \rightsquigarrow (F \vee I)$

If we accept only the weaker conditional $C \rightsquigarrow (F \vee I)$ but neither $C \rightsquigarrow F$ nor $C \rightsquigarrow I$, we must *reject Uniqueness*

Almost-Connectedness (and ASP)

Almost-Connectedness

$$\forall z \in W\ \forall u, v, w \in W_z : \quad (u \prec_z w) \ \Rightarrow\ (u \prec_z v \ \ \vee \ \ v \prec_z w)$$

This is weaker than Connectedness: ties are allowed; it just forces every v to be comparable with the "interval" determined by $u \prec_z w$.

ASP: Strengthening with a Possibility

$$(\neg(\varphi \rightsquigarrow \neg\psi)\ \wedge\ (\varphi \rightsquigarrow \chi)) \ \ \rightarrow \ \ ((\varphi \wedge \psi)\ \rightsquigarrow\ \chi)$$

$\neg(\varphi \rightsquigarrow \neg\psi)$ can be read as *If it had been the case that φ, it might have been the case that ψ.* Then any $\varphi \rightsquigarrow \chi$ remains true when we strengthen φ by such a *compatible* ψ.

Intuitively, AC orders *ties* into blocks: if we set $u \simeq_z v$ iff neither $u \prec_z v$ nor $v \prec_z u$, then the $\simeq_z$-classes are linearly ordered by $\prec_z$.

Counterexample to ASP (Bizet, Verdi and Satie)

(15) a. It's not the case that if Verdi and Satie had been compatriots, Satie and Bizet would not have been compatriots. $\neg(\varphi \rightsquigarrow \neg\psi)$

b. If Verdi and Satie had been compatriots, Bizet would have been French. $\varphi \rightsquigarrow \chi$

Many accept (a) and (b) but hesitate about the conclusion:

(16) If both Verdi and Satie, and Satie and Bizet had been compatriots, Bizet would have been French. $(\varphi \wedge \psi) \rightsquigarrow \chi$

Outline

Tichý's hat example (Tichý 1976)

If the weather is bad, Jones always wears a hat. If the weather is fine, he may or may not wear a hat at random. Actually the weather is bad and he is wearing a hat.

B = bad weather, F = fine weather, H = Jones wearing a hat.
(Actual world: $B \wedge H$)

We hesitate to assert:

$$\text{(T)} \qquad F \rightsquigarrow H$$

Lewis's similarity ordering tends to keep H fixed when switching $B \to F$, predicting $F \rightsquigarrow H$ (closest F-world preserves H). That clashes with the intended "random under F" reading.

Treating H as a "fact" to be preserved makes (T) *true* by similarity, but intuitively we want only the *might* claim $\neg(F \rightsquigarrow \neg H)$.

Fine's Nixon button (Fine 1975)

Nixon does not press the button for launching the nuclear bomb. The launch mechanism is functioning. There is no holocaust.

P = "Nixon presses the button", W = "launch mechanism works normally", H = "nuclear holocaust".

We found plausible to assert:

$$\text{(F)} \qquad P \rightsquigarrow H$$

To make P true, compare:

1. $P \wedge W \Rightarrow H$ (massive divergence: widespread change).
2. $P \wedge \neg W \Rightarrow \neg H$ (local failure prevents divergence).

Clearly, (1) is more similar to the actual world than (2). Hence (F) should come out false.

The Oswald/Kennedy case

If Oswald had not shot Kennedy, someone else would have.

We judge this counterfactual false. But On Lewis's similarity ordering, this counterfactual is plausibly *true*.

- Actual world: Oswald shoots Kennedy. Kennedy dies and history unfolds roughly as it actually did.
- To make the antecedent true, we consider worlds where Oswald does *not* shoot Kennedy.
- Among such worlds, the most similar can be determined by:
  - some *other* small change occurs (another assassin shoots Kennedy),
  - so that Kennedy is still killed and large-scale history remains similar,
  - rather than worlds where Kennedy survives and history diverges massively.

Lewis's system of weights (Lewis 1979)

Lewis Heuristics for $\prec_w$:

1. Avoid big, widespread, diverse law-violations ("big miracles").
2. Maximize exact spatio-temporal match of particular facts with w.
3. Avoid even small, local law-violations ("little miracles").
4. Give *low* weight to *approximate* similarity of particular facts.

- Actual world (Tichý case): $B \wedge H$ (bad weather, Jones wears a hat). To make F true, we need a small "miracle" changing the weather from bad to fine.
- Among F-worlds, compare $F \wedge H$ and $F \wedge \neg H$. After we changed the weather, there is no strict law linking fine weather to (not) wearing a hat, so both are equally law-respecting.
- Highly contingent particular facts (such as whether Jones happens to wear his hat) should matter little for similarity.

Are we treating the relevance of particular facts in an ad hoc way?

Problem for Lewis's system of weights: particular facts vs. causal influence (Morgenbesser conditionals)

I decline to bet on a coin toss. It is tossed anyway. It lands heads.

B = "I bet on heads"
H = "coin lands heads"
W = "I win" (take $W := B \wedge H$)

(Actual world: $\neg B \wedge H$)

Often accepted: $B \rightsquigarrow W$

Problem for Lewis's system of weights: particular facts vs. causal influence (Morgenbesser conditionals)

Sometimes this **is reasonable (independence):**

- *No causal path* $B \rightarrow H$ (my betting doesn't affect the toss).
- We hold H fixed.

Sometimes this **is not (possible influence):**

- B might influence the toss (timing, set-up, psychology),
- We *cannot* hold H fixed when moving to B-worlds.
- At best one may get a "might" claim.

There is a tension between Lewis's "preserve particular facts" heuristic and our causality-sensitive intuitions about counterfactuals.

Disjunctive antecedents

$$(\varphi \rightsquigarrow \chi) \ \wedge \ (\psi \rightsquigarrow \chi) \ \Rightarrow \ (\varphi \vee \psi) \rightsquigarrow \chi \quad \text{(AD, valid)}$$

$$(\varphi \vee \psi) \rightsquigarrow \chi \ \Rightarrow \ (\varphi \rightsquigarrow \chi) \ \wedge \ (\psi \rightsquigarrow \chi) \quad \text{(SDA, not valid)}$$

SDA seems acceptable in most cases:

(a) If there had been a fire or an earthquake in this building, the alarm would have gone off.

(b) If there had been a fire in this building, the alarm would have gone off, and if there had been an earthquake in this building, the alarm would have gone off.

Disjunctive antecedents

But in other cases, we do not accept it (McKay and van Inwagen 1977):

(a) If Spain had entered the Second World War on one side or the other, it would have fought on the side of the Axis powers.

(b) If Spain had entered the Second World War on the side of the Allied powers, it would have fought on the side of the Axis powers.

The closest $(\text{Axis} \vee \text{Allies})$-world may be an Axis world. Yet the closest Allies-world can be different (and farther), where “Axis” is false. Thus

$$(\varphi \vee \psi) \rightsquigarrow \chi \;\not\Rightarrow\; \psi \rightsquigarrow \chi$$

Think: in which cases do we accept SDA, and in which cases do we reject SDA?

Outline

Premise Semantics (Kratzer & Veltman)

Frank Veltman

Angelika Kratzer

Counterfactuals are evaluated relative to a body of *premises*: background facts, laws, norms, etc.

Ramsey Test (Stalnaker 1968): To evaluate *If* α*, then* β:

1. Add α hypothetically to your beliefs.
2. Revise your beliefs *minimally* to restore consistency, keeping α.
3. Check whether β holds in the revised belief state.

Premise semantics:

- taking $\mathcal{H}_w$ as your basic premises at w,
- looking at all *maximal* subsets of $\mathcal{H}_w$ that are consistent with α,
- $\alpha \rightsquigarrow \beta$ *is true at* w *iff every such "maximal* α*-revision" still forces* β*.*

A toy formal example

Let $W = \{w_1, w_2, w_3\}$ and fix $w := w_1$. Suppose

$$[\![\varphi]\!] = \{w_1, w_2\} \qquad [\![\psi]\!] = \{w_1, w_3\} \qquad \mathcal{H}_w = \{[\![\varphi]\!], [\![\psi]\!]\}$$

Let the antecedent be α with

$$[\![\alpha]\!] = \{w_2, w_3\}$$

For $\Delta \subseteq \mathcal{H}_w$, call Δ *α-consistent* iff

$$\bigcap(\Delta \cup \{[\![\alpha]\!]\}) \neq \emptyset$$

Call Δ *maximally α-compatible* iff Δ is α-consistent and there is no Δ' with $\Delta \subsetneq \Delta' \subseteq \mathcal{H}_w$ that is still α-consistent. Let $\mathrm{Max}_w(\alpha)$ be the set of all maximally α-compatible Δ.

A toy formal example

Δ	$\bigcap(\Delta \cup \{[\![\alpha]\!]\})$	α-consistent?	maximal?
$\emptyset$	$\{w_2, w_3\}$	yes	no
$\{[\![\varphi]\!]\}$	$\{w_2\}$	yes	✓
$\{[\![\psi]\!]\}$	$\{w_3\}$	yes	✓
$\{[\![\varphi]\!], [\![\psi]\!]\}$	$\emptyset$	no	–

Thus we have three α-consistent subsets ($\emptyset$, $\{[\![\varphi]\!]\}$, $\{[\![\psi]\!]\}$), and exactly two of them are maximally α-compatible:

$$\mathrm{Max}_w(\alpha) = \{\{[\![\varphi]\!]\},\ \{[\![\psi]\!]\}\}$$

Premise semantics for $\rightsquigarrow$ (relative to $\mathcal{H}_w$)

$M, w \models \alpha \rightsquigarrow \beta$ iff for every $\Delta \in \mathrm{Max}_w(\alpha)$,

$$\bigcap(\Delta \cup \{[\![\alpha]\!]\}) \subseteq [\![\beta]\!]$$

A toy formal example

Now pick two consequents with

$$[\![\beta]\!] = \{w_1, w_2, w_3\} \quad \text{and} \quad [\![\gamma]\!] = \{w_1, w_2\}$$

For each $\Delta \in \mathrm{Max}_w(\alpha)$:

$$\Delta = \{[\![\varphi]\!]\} : \quad \bigcap(\Delta \cup \{[\![\alpha]\!]\}) = \{w_2\}$$
$$\Delta = \{[\![\psi]\!]\} : \quad \bigcap(\Delta \cup \{[\![\alpha]\!]\}) = \{w_3\}$$

- $\alpha \rightsquigarrow \beta$ is true at w since $\{w_2\} \subseteq [\![\beta]\!]$ and $\{w_3\} \subseteq [\![\beta]\!]$.
- $\alpha \rightsquigarrow \gamma$ is false at w since $\{w_3\} \not\subseteq [\![\gamma]\!]$.

A natural language example (Kratzer's bridge)[3]

Angelika and Regina must cross a bridge one after the other.

- β: *Angelika takes one minute to cross.*
- γ: *Regina waits one minute before crossing.*
- α: *Angelika takes 40 seconds.* (incompatible with β)

Actual situation: β and γ.

If Angelika had taken 40 seconds, would Regina have waited one minute?

Formally: do we have $\alpha \rightsquigarrow \gamma$ at the actual world w?

[3]The reprinted version of Kratzer (1981) in Kratzer (2012) is more pleasant to read.

The idea of lumping

1. Separate premises: $\mathcal{H}_w = \{[\![\beta]\!], [\![\gamma]\!]\}$.
   - To make α true we must give up β, but can keep γ.
   - Maximal α-compatible subsets all still contain $[\![\gamma]\!]$, so in every α-revision, γ holds.
   - Thus $\alpha \rightsquigarrow \gamma$ *true*.

2. Lumped premise: $\mathcal{H}_w = \{[\![\beta \wedge \gamma]\!]\}$.
   - α is incompatible with $\beta \wedge \gamma$, so we must drop that whole premise.
   - The maximal α-compatible set contains no information about γ.
   - Thus $\alpha \rightsquigarrow \gamma$ *false*.

Equivalence with ordering semantics

Lewis (1981) proved that premise semantics and similarity/ordering semantics are **truth-conditionally equivalent**.

In the case of **premise semantics** (Veltman 1976, Kratzer 1981), we have for each w, a set $\mathcal{H}_w$ of propositions (all true at w).

In the case of **similiarity/ordering semantics** (Stalnaker 1968, Lewis 1973), we have for each w, a similarity relation $\prec_w$ on W_w.

From premises to orderings

- For a fixed w, compare two worlds u and v in W_w by looking at which members of $\mathcal{H}_w$ they satisfy.
- Say that u is more similar to w than v just in case
  - every proposition in $\mathcal{H}_w$ that is true at v is also true at u,
  - and at least one proposition in $\mathcal{H}_w$ is true at u but not at v.
- Thus the ordering ranks worlds by how well they preserve the background premises at w.

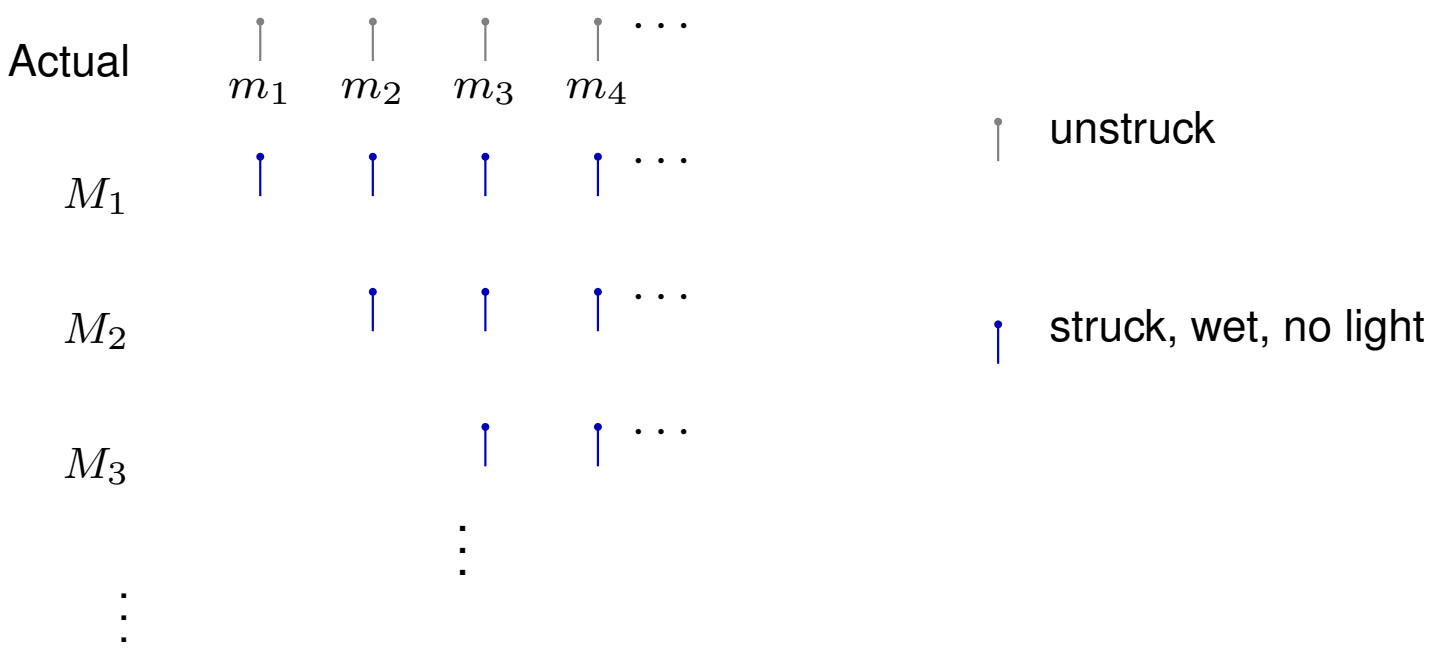
From ordering to premises

- For a fixed w, and for each world v in W_w, consider the set of worlds that are at least as similar to w as v is according to $\prec_w$.
- Treat each such set as a single proposition saying that "the actual world is no less similar to w than v is".
- Let $\mathcal{H}_w$ be the collection of all these propositions.
- With this choice of $\mathcal{H}_w$, premise semantics makes exactly the same selections of "closest" antecedent-worlds, and hence delivers the same truth conditions for counterfactuals as the original similarity ordering semantics.

Outline

Matches (Fine 2011, 2012)[4]

- We have infinitely many matches $m_1, m_2, m_3, \ldots$ in different regions.
- Actual world: No match is struck.
- M_1: All matches are struck; from m_1 onward they are wet and do *not* light.
- M_2: All matches are struck; from m_2 onward they are wet and do not light.
- M_3: m_1, m_2 light; from m_3 onward they are wet and do not light.
- M_4: ...



[4] You will not be tested on the contents of this section.

The argument

Let M_n be the sentence "scenario M_n holds", and set $S := \bigvee_{n\geq 1} M_n$.

- **Positive Effect:** for all n $M_n \rightsquigarrow M_{n+1}$.
- **Negative Effect:** for all n $M_{n+1} \rightsquigarrow \neg M_n$.
- S is a coherent counterfactual supposition (i.e., $\neg(S \rightsquigarrow \neg S)$).

We assume:

- Substitution of equivalents in antecedent position.
- Transitivity, Disjunction, and (infinitary) Conjunction for $\rightsquigarrow$.

Then we can derive, for every k,

$$S \rightsquigarrow (M_{k+1} \vee M_{k+2} \vee \dots)$$

Hence

$$S \rightsquigarrow \bigwedge_{n\geq 1} \neg M_n, \quad \text{i.e.} \quad S \rightsquigarrow \neg S$$

contradicting the assumption that S is a coherent supposition.

Logical principles used

Substitution in antecedent

If $\varphi \subset\!\supset \varphi'$ (classically equivalent), then:

$$(\varphi \rightsquigarrow \psi) \subset\!\supset (\varphi' \rightsquigarrow \psi)$$

Transitivity

From $(\varphi \rightsquigarrow \psi)$ and $(\varphi \wedge \psi) \rightsquigarrow \chi$ infer:

$$\varphi \rightsquigarrow \chi$$

Disjunction

If φ and ψ are mutually inconsistent and

$$\varphi \rightsquigarrow \chi \quad \text{and} \quad \psi \rightsquigarrow \chi, \text{ then } (\varphi \vee \psi) \rightsquigarrow \chi$$

(Infinitary) Conjunction

From $\varphi \rightsquigarrow \psi_n$ for all $n \geq 1$, infer

$$\varphi \rightsquigarrow \bigwedge_{n \geq 1} \psi_n$$

Two derived rules

Weakening (of the consequent)
If $\varphi \rightsquigarrow \psi$ and $\psi \models \chi$, then

$$\varphi \rightsquigarrow \chi$$

Transitivity′
If

$$\varphi \rightsquigarrow \psi, \qquad \psi \rightsquigarrow \chi, \qquad \text{and } \psi \models \varphi$$

then

$$\varphi \rightsquigarrow \chi$$

A local lemma for M_1

1. From *Negative Effect* (for $n = 1$): $M_2 \rightsquigarrow \neg M_1$
2. Note: M_2 is equivalent to $(M_1 \wedge M_2) \vee (\neg M_1 \wedge M_2)$.
3. By *Substitution*:
$$((M_1 \wedge M_2) \vee (\neg M_1 \wedge M_2)) \rightsquigarrow \neg M_1$$
4. From *Positive Effect* (for $n = 1$): $M_1 \rightsquigarrow M_2$
5. Trivially also $M_1 \rightsquigarrow M_1$. By *Conjunction*:
$$M_1 \rightsquigarrow (M_1 \wedge M_2)$$
6. Since $(M_1 \wedge M_2) \models (M_1 \wedge M_2) \vee (\neg M_1 \wedge M_2)$, by Weakening:
$$M_1 \rightsquigarrow ((M_1 \wedge M_2) \vee (\neg M_1 \wedge M_2))$$
7. By *Entailment*, also
$$(\neg M_1 \wedge M_2) \rightsquigarrow ((M_1 \wedge M_2) \vee (\neg M_1 \wedge M_2))$$
8. Apply *Disjunction* to the last two lines (antecedents M_1 and $\neg M_1 \wedge M_2$ are incompatible):
$$(M_1 \vee (\neg M_1 \wedge M_2)) \rightsquigarrow ((M_1 \wedge M_2) \vee (\neg M_1 \wedge M_2))$$
9. Combine this with step (3) via *Transitivity'* to obtain:
$$(M_1 \vee (\neg M_1 \wedge M_2)) \rightsquigarrow \neg M_1$$

From the lemma to $S \rightsquigarrow \neg M_1$

We already have

$$(M_1 \vee (\neg M_1 \wedge M_2)) \rightsquigarrow \neg M_1$$

- By *Entailment*,

$$(\neg M_1 \wedge \neg M_2 \wedge (M_3 \vee M_4 \vee \dots)) \rightsquigarrow \neg M_1$$

- By *Disjunction* (antecedents incompatible):

$$(M_1 \ \vee \ (\neg M_1 \wedge M_2) \ \vee \ (\neg M_1 \wedge \neg M_2 \wedge (M_3 \vee M_4 \vee \dots))) \rightsquigarrow \neg M_1$$

- But the big disjunction in the antecedent is just a partition of

$$S \ = \ M_1 \vee M_2 \vee M_3 \vee \dots$$

(either M_1, or not-M_1 but M_2, or neither M_1 nor M_2 but some later M_n).

- So by *Substitution* in the antecedent:

$$S \rightsquigarrow \neg M_1$$

Pushing the argument along the sequence

From $S \rightsquigarrow \neg M_1$ we can strengthen the consequent and then iterate.

(i) Exclude M_1 and move to later M_n:

- From $S \rightsquigarrow \neg M_1$ and trivial $S \rightsquigarrow S$, by *Conjunction*:
$$S \rightsquigarrow (\neg M_1 \wedge S)$$
- But $(\neg M_1 \wedge S) \models (M_2 \vee M_3 \vee \dots)$, so by *Weakening*:
$$S \rightsquigarrow (M_2 \vee M_3 \vee \dots)$$

(ii) Repeat the previous pattern starting at M_2:

- Using *Positive* and *Negative Effect* for M_2, M_3 and the same derivation as before, we obtain:
$$(M_2 \vee M_3 \vee \dots) \rightsquigarrow \neg M_2.$$
- Now apply *Transitivity'* to get:
$$S \rightsquigarrow \neg M_2$$

By repeating the argument starting at each M_n, we obtain, for every $n \geq 1$,

$$S \rightsquigarrow \neg M_n$$

The contradiction

We have derived, for each $n \geq 1$,

$$S \rightsquigarrow \neg M_n$$

Using the *(infinitary) Conjunction* rule, we get:

$$S \rightsquigarrow \bigwedge_{n \geq 1} \neg M_n$$

But the conjunction $\bigwedge_{n \geq 1} \neg M_n$ just says that *none* of the M_n scenarios holds, i.e.

$$\bigwedge_{n \geq 1} \neg M_n \;\; \subset\!\supset \;\; \neg S$$

By *Weakening*, we conclude:

$$S \rightsquigarrow \neg S$$

This contradicts our assumption that S is a coherent counterfactual supposition $\neg(S \rightsquigarrow \neg S)$

Fine's critique of similarity semantics

- Possible-worlds semantics validates: if $\varphi \subset\!\supset \varphi'$ then $(\varphi \rightsquigarrow \psi) \subset\!\supset (\varphi' \rightsquigarrow \psi)$.
- But Fine's *matches* cases suggest that:
  - we want to keep principles like Conjunction and Disjunction (and a strong *infinitary Conjunction rule*),
  - we want to keep intuitively compelling counterfactual judgments about the cases,
  - yet we *must* reject full Substitution to avoid contradiction.
- Lewis's reaction: the Limit Assumption (no closest worlds) to block the *infinitary Conjunction rule*
- Fine's diagnosis: the real culprit is the idea that only the *set of worlds* where φ is true matters. We also need to care about *how* φ is made true: truthmakers!

Fine's truth-conditions for counterfactuals

We add a transition relation $t \rightarrow_w u$:

- t is a state we *impose* on world w (a way of making the antecedent true),
- u is a *possible outcome* of imposing t at w.

We want:

- **Universal realizability of the antecedent:**
  Check *every* state t that exactly verifies φ (every admissible way φ might have held).
- **Universal verifiability of the consequent:**
  For each such t, look at *every* possible outcome u of imposing t at w and require that u inexactly verify ($\models_i$) ψ.

(Recall that $\mathcal{M}, s \models_i \phi$ iff for some $s' \leq s$, $\mathcal{M}, s' \models_e \phi$.)

Example

Let $T(\varphi)$ be the set of states that *exactly* verify φ.

- Suppose $T(\varphi) = \{t_1, t_2\}$ and for each $i = 1, 2$ there are (possibly many) outcomes $u_{i1}, u_{i2}, \ldots$ such that

$$t_i \to_w u_{ij} \quad \text{for all } j, \qquad \text{and} \qquad M, u_{ij} \models_i \psi \text{ for all } j.$$

  Then every admissible way of making φ true *and* every possible outcome of that way satisfies ψ, so $M, w \models \varphi \rightsquigarrow \psi$.

- If instead there is some $t^* \in T(\varphi)$ and u^* with

$$t^* \to_w u^* \quad \text{and} \quad M, u^* \not\models_i \psi,$$

  then *that* way of realizing φ yields a bad outcome, and $M, w \not\models \varphi \rightsquigarrow \psi$.

Evaluating a counterfactual

Definition (Counterfactual: truthmaker semantics)

$M, w \models \varphi \rightsquigarrow \psi$

iff for every state t with t exactly verifying φ ($t \models_e \varphi$), and every outcome u with $t \rightarrow_w u$, we have u inexactly verifying ψ ($u \models_i \psi$).

Whichever way φ might have come about, *whatever* would then have happened, ψ would (in that outcome) be true.

Logic on Fine's semantics

- **Conjunction** (including infinitary forms):
  If $\varphi \rightsquigarrow \psi_i$ for all i, then $\varphi \rightsquigarrow \bigwedge_i \psi_i$.
- **Inclusive Disjunction** (with an extra premise):
  From $\varphi \rightsquigarrow \chi$, $\psi \rightsquigarrow \chi$, and $(\varphi \wedge \psi) \rightsquigarrow \chi$, infer $(\varphi \vee \psi) \rightsquigarrow \chi$.
- **Exact strengthening of the antecedent**:
  If φ' is an *exact logical consequence* of φ and $\varphi \rightsquigarrow \psi$, then $\varphi' \rightsquigarrow \psi$.
- **Inexact weakening of the consequent**:
  If ψ *inexactly entails* χ and $\varphi \rightsquigarrow \psi$, then $\varphi \rightsquigarrow \chi$.

- But **Substitution of equivalents** *fails*: the truth of $\varphi \rightsquigarrow \psi$ can depend on *how* φ is made true, not just on the set $[\![\varphi]\!]$.

Solving the match puzzle

For the match scenarios $M_1, M_2, \ldots$ we assumed:

- Positive Effect: $M_n \rightsquigarrow M_{n+1}$
- Negative Effect: $M_{n+1} \rightsquigarrow \neg M_n$
- Counterfactual Possibility: some M_n is a coherent counterfactual supposition.
- Logical rules: Entailment, Transitivity, infinitary Conjunction, ...
- Substitution of equivalent antecedents.

- **Reject full Substitution for antecedents**[5]

[5] A similar concern is stated by Ciardelli, Zhang and Champollion (2018), who showed experimentally that counterfactuals with premises $\neg\varphi \vee \neg\psi$ and $\neg(\varphi \wedge \psi)$ receive different evaluations. Truthmaker semantics does not resolve this issue, since De Morgan's laws hold.

Bonus: belief revision[6]

$\varphi \rightsquigarrow \psi$:= *'After revising by φ, it is believed that ψ.'*

- How an agent should change their beliefs in the face of new (possibly conflicting) information.
- φ is treated as incoming information: if $\varphi \rightsquigarrow \psi$ is true, then ψ is accepted after the revision by φ.
- Many axioms for $\rightsquigarrow$ now read very naturally:
  - **CI** ($\varphi \rightsquigarrow \varphi$): after revising by φ, you believe φ.
  - **AD**: $(\varphi \rightsquigarrow \chi) \wedge (\psi \rightsquigarrow \chi) \supset ((\varphi \vee \psi) \rightsquigarrow \chi)$
    "If both a revision by φ and a revision by ψ lead you to accept χ, then so does a revision by $\varphi \vee \psi$."

[6] This section of the slides is optional. It also follows quite closely Frank Veltman's notes listed at the beginning of the slides.

AGM belief revision in a nutshell

- **Expansion** $K + \varphi$: just add φ and close under consequence.
- **Contraction** $K - \varphi$: give up φ while changing K as little as possible.
- **Revision** $K * \varphi$: incorporate φ (possibly contradicting K) by minimally changing K so that φ is believed.

AGM postulates for revision $K * \varphi$

For all belief sets K and sentences φ, ψ:

(K*1) Closure: $K * \varphi$ is a belief set.

(K*2) Success: $\varphi \in K * \varphi$.

(K*3) Inclusion: $K * \varphi \subseteq K + \varphi$.

(K*4) Vacuity: If $\neg\varphi \notin K$, then $K + \varphi \subseteq K * \varphi$.

(K*5) Consistency: $K * \varphi$ is inconsistent only if $\vdash \neg\varphi$.

(K*6) Extensionality: If $\vdash \varphi \equiv \psi$, then $K * \varphi = K * \psi$.

(K*7) Superexpansion: $K * (\varphi \wedge \psi) \subseteq (K * \varphi) + \psi$.

(K*8) Subexpansion / Rational Monotonicity: If $\neg\psi \notin K * \varphi$, then $(K * \varphi) + \psi \subseteq K * (\varphi \wedge \psi)$.

Similarity models as belief revision models

The *same* ordering semantics used for counterfactuals can represent AGM-style belief revision.

Fix a belief set K and build a model $M_K = \langle W, \prec, V \rangle$:

- W = set of all maximal consistent theories in the underlying language.
- $\prec$ a *well-founded*, *almost-connected* strict partial order on W ("more plausible / closer to K").
- The $\prec$-minimal worlds in W are exactly the maximal consistent extensions of K.
- $V(p, w) = 1$ iff $p \in w$.

Similarity models as belief revision models

Now define the revision of K by φ via *closest φ-worlds*:

$$\psi \in K * \varphi \quad \text{iff} \quad \psi \in w \text{ for every } w \in \mathrm{Min}_{\prec}([\![\varphi]\!]).$$

- With this definition, $K*$ satisfies all AGM postulates (K*1–K*8).
- Moreover, whenever $\psi \in K * \varphi$, the corresponding model M_K makes the counterfactual $\varphi \rightsquigarrow \psi$ true (at all "K-worlds").

The similarity ordering $\prec$ that we used for counterfactuals can also be read as encoding a policy for belief revision.

From belief revision back to similarity orderings

Conversely, suppose we start from a belief set K and a revision function $K*$ satisfying AGM (K*1-K*8).

We can reconstruct a *similarity model* $M_K = \langle W, \prec, V \rangle$ such that:

$$\psi \in K * \varphi \quad \Longleftrightarrow \quad \psi \text{ holds at all } \prec\text{-minimal } \varphi\text{-worlds.}$$

Sketch of the construction:

- W = set of all maximal consistent theories.
- $V(p, w) = 1$ iff $p \in w$.
- For each world w define a certain set $\tau(w) \subseteq W$ determined by how w relates to the various revised belief sets $K * \varphi$.
- Set
$$u \prec w \quad \text{iff} \quad \tau(w) \subseteq \tau(u) \text{ and } u \notin \tau(w)$$
- One can show that this $\prec$ is well-founded and almost connected, and recovers $K*$ via minimal φ-worlds.

Revision vs. counterfactual supposition

Even though the formal connections are tight, *revising* your beliefs by φ is not the same cognitive act as *supposing* that φ (had) been the case.

Counterfactual supposition:
"If Oswald had not killed Kennedy, Kennedy might still be alive." You imagine a scenario in which $\neg\varphi$ holds and in which it is open that Kennedy survives.

Actual belief revision: If you were to *discover* that φ is false and revise your beliefs accordingly (in light of all your other evidence about 1963), you would almost certainly still believe that Kennedy is dead.

- Counterfactual reasoning: exploring hypothetical alternatives, often holding fixed much of what we actually believe.
- AGM revision: correcting your belief set in light of new, trustworthy information.