Counterfactuals Modulo Temporal Logics

Bernd Finkbeiner and *Julian Siber*

CISPA Helmholtz Center for Information Security

LPAR-24 | 9 June 2023

Counterfactuals

"If φ had been true, then ψ would have been true, too."

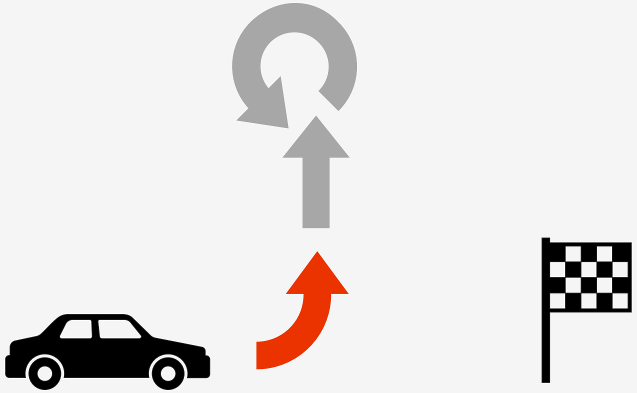
$$\varphi \mathrel{\Box\!\!\rightarrow} \psi$$

Counterfactuals Modulo Temporal Logics

*"If the car had **moved straight** at the first time point, then it would have **reached its goal eventually**."*

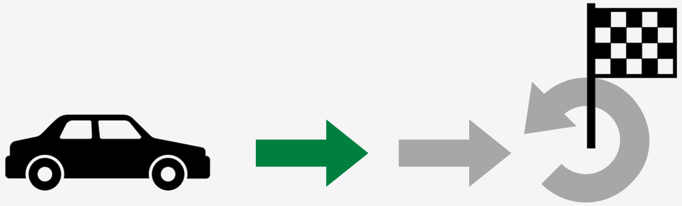
$$(straight) \mathrel{\Box\!\!\rightarrow} (\mathbf{F}\, goal)$$

Counterfactuals = Variably Strict Conditionals

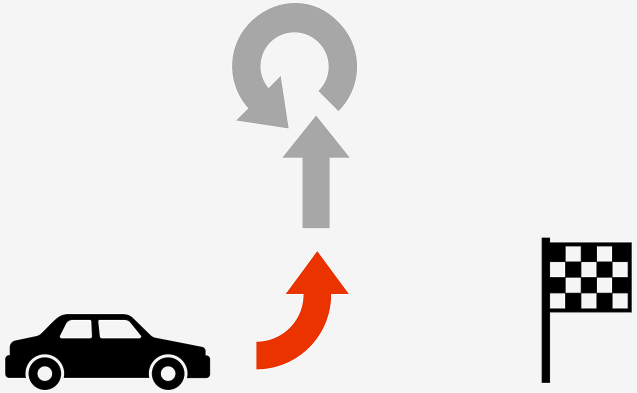


$$\models (straight) \;\Box\!\!\rightarrow (\mathtt{F}\; goal)$$

because the closest counterfactual world is:

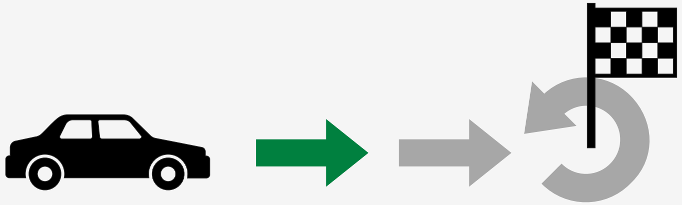


Counterfactuals = Variably Strict Conditionals

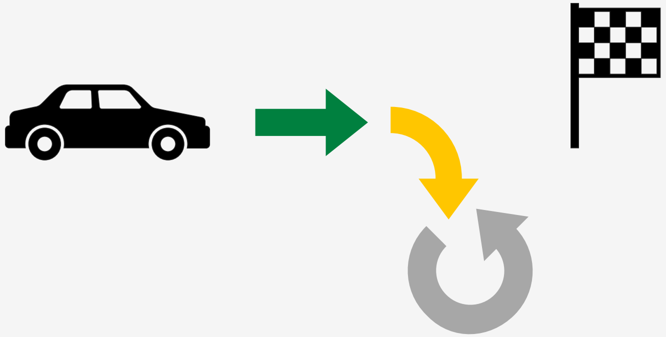


$$\models (straight) \mathrel{\Box\!\!\rightarrow} (\mathbf{F}\ goal)$$

counterfactual worlds further away do not matter:



$$<_{far}$$



Applications of Counterfactual Reasoning

- Analyzing **causality** [Halpern '15], [Leitner-Fischer '15], [Coenen et al. '22]
- Generating **explanations** for, e.g., model checkers [Beer et al. '09], [Wachter et al. '18]
- Counterfactual **fairness** [Kusner et al. '17]

Applications of Counterfactual Reasoning

Analyzing **causality** [Halpern '15], [Leitner-Fischer '15], [Coenen et al. '22]

Definition 5 (Property Causality). *Let $\mathcal{T}$ be a system, $\pi \in traces(\mathcal{T})$ a trace, $\mathsf{C} \subseteq (2^I)^\omega$ a cause property, and $\mathsf{E} \subseteq (2^O)^\omega$ an effect property. We say that* C *is a* cause *of* E *on π in $\mathcal{T}$ if the following three conditions hold:*

PC1: *$\pi \models \mathsf{C}$ and $\pi \models \mathsf{E}$, i.e., cause property and effect property are satisfied by the actual trace.*

PC2: *For every counterfactual input sequence $\sigma \in V_\pi^{\mathsf{C}}$, there is some contingency $\pi' \in C_\pi^\sigma$ s.t. $\pi' \not\models \mathsf{E}$, i.e., the counterfactual trace under contingency does not satisfy the effect property.*

PC3: *There is no C' s.t. $\mathsf{C}' \subset \mathsf{C}$ and C' satisfies PC1 and PC2.*

[Coenen et al. '22]

Definition 3.1: (Actual cause) $\vec{X} = \vec{x}$ is an *actual cause of* φ *in* $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$. (That is, both $\vec{X} = \vec{x}$ and φ are true in the actual world.)

AC2. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}) \models Z = z^*$ for $Z \in \vec{Z}$, then

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\varphi$. In words, changing $(\vec{X}, \vec{W})$ from $(\vec{x}, \vec{w})$ to $(\vec{x}', \vec{w}')$ changes φ from true to false,

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$. In words, setting $\vec{W}$ to $\vec{w}'$ should have no effect on φ as long as $\vec{X}$ is kept at its current value $\vec{x}$, even if all the variables in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2. Minimality ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing φ in AC2(a) are considered part of a cause; inessential elements are pruned. ∎

[Halpern&Pearl '05]

Applications of Counterfactual Reasoning

- Analyzing **causality** [Halpern '15], [Leitner-Fischer '15], [Coenen et al. '22]

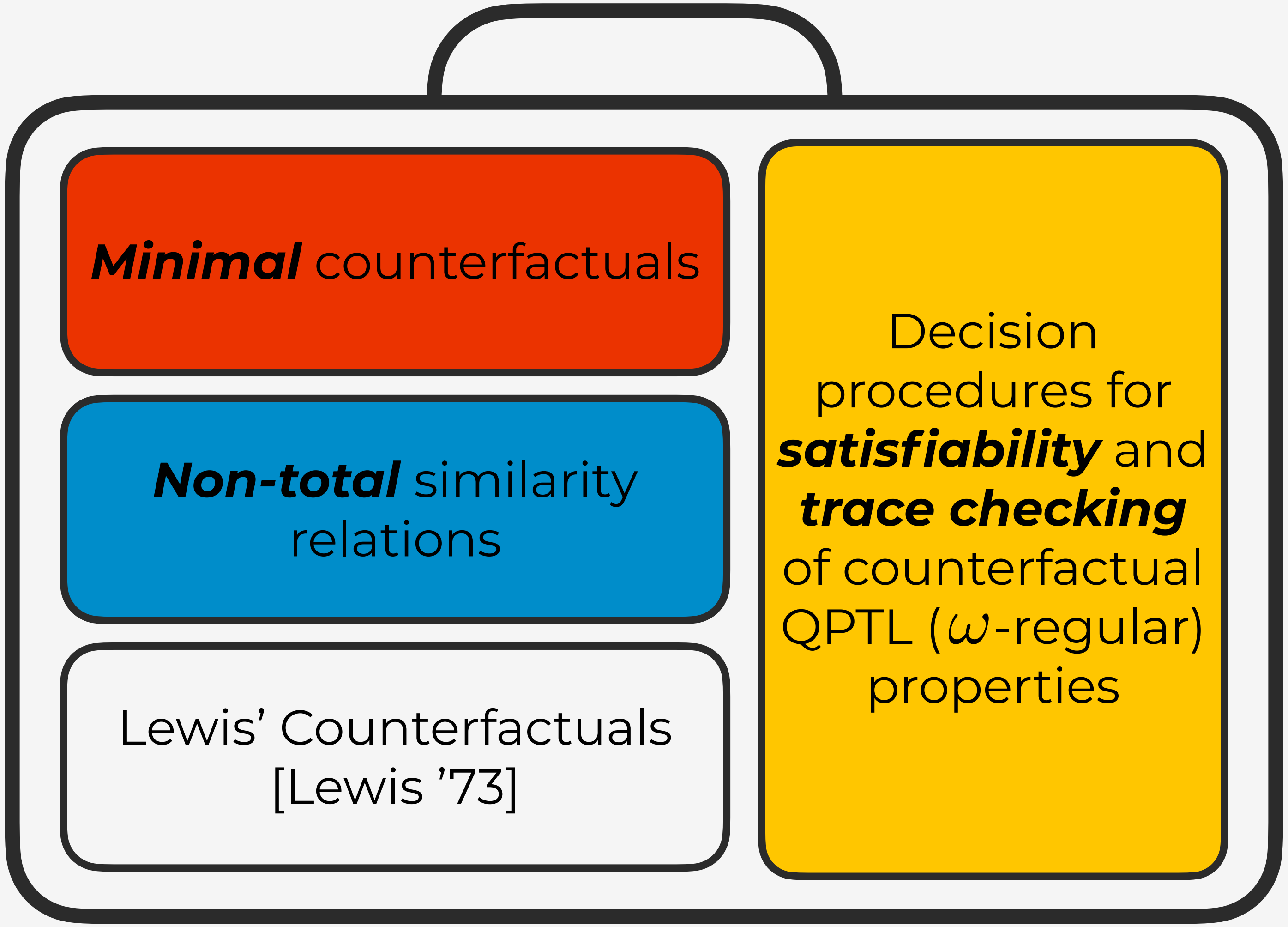
$$\varphi \wedge \psi \wedge \big((\neg\varphi \boxdot\!\!\rightarrow_{min} \neg\psi) \vee (\neg\varphi \diamond\!\!\rightarrow_{min} \neg\psi)\big)$$

[Coenen et al. '22]

$$\varphi \wedge \psi \wedge (\neg\varphi \diamond\!\!\rightarrow_{min} \neg\psi)$$

[Halpern '15]

Overview



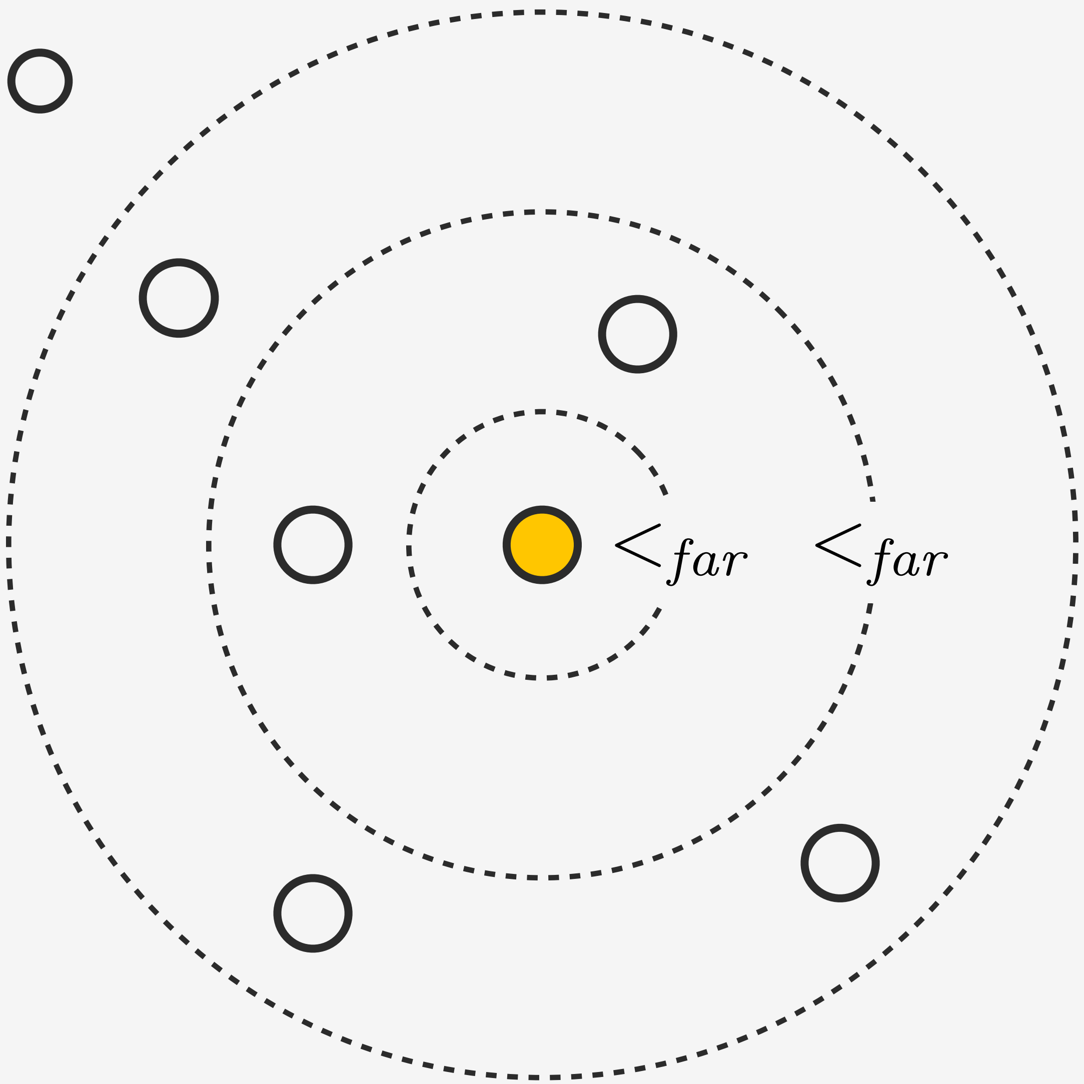
Specification toolbox for, e.g., temporal causality

Similarity Relation $\leq_{far}(\ )$

Encodes the distance of a world from the reference world .

Needs to be a total preorder.

is the unique minimum.

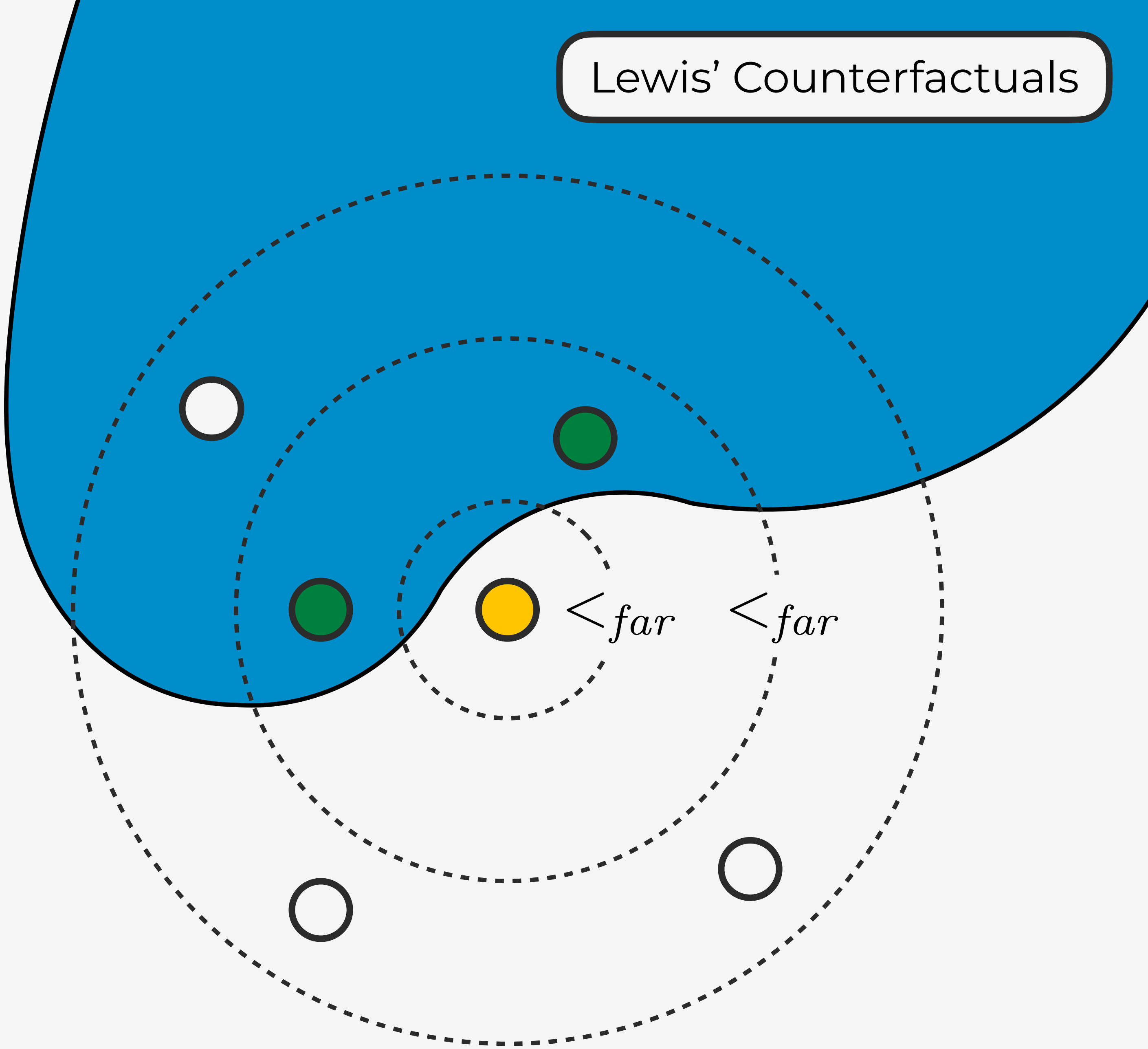


‘Would’ Operator

$$\varphi \; \Box\!\!\rightarrow \psi$$

All closest worlds satisfying φ have to satisfy ψ .

Worlds in spheres further away do not matter.

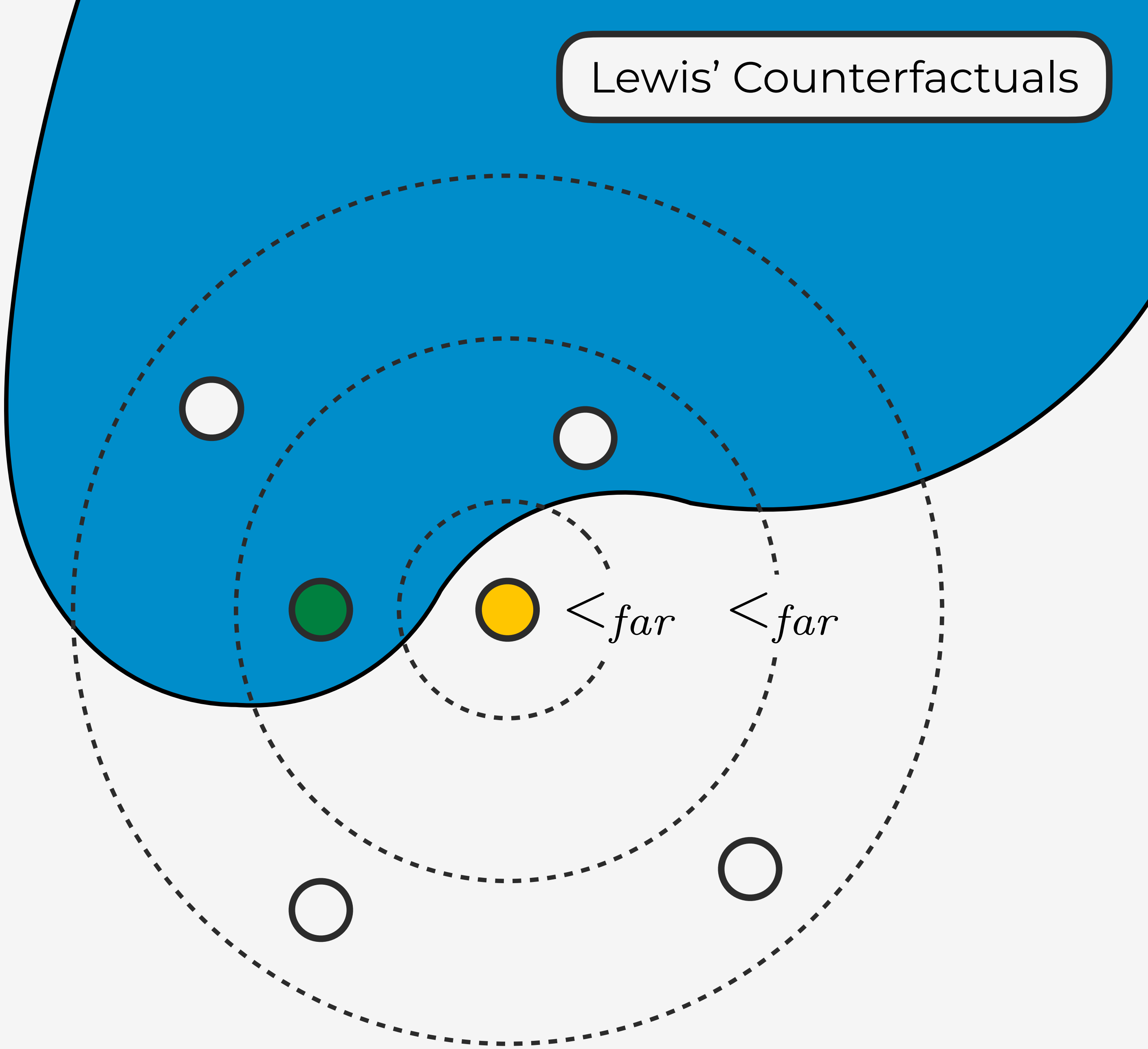


'Might' Operator

$$\varphi \Diamond\!\!\rightarrow \psi$$

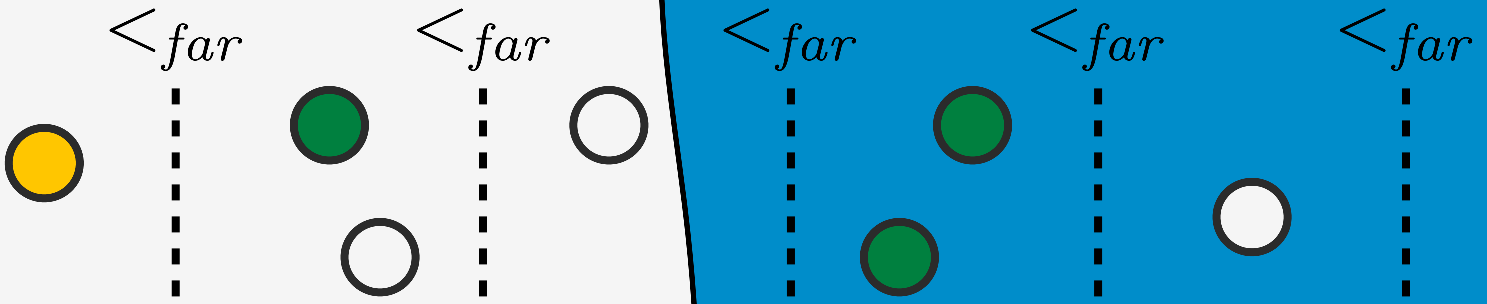
Some closest world satisfying φ has to satisfy ψ .

Again, worlds in spheres further away do not matter.



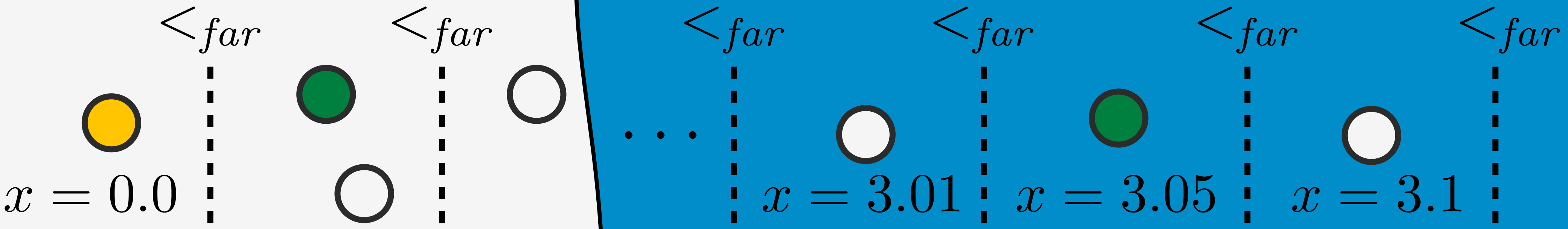
The Limit Assumption

"There always exist well-defined closest worlds satisfying φ *."*



The Limit Assumption

"There always exist well-defined closest worlds satisfying φ *."*



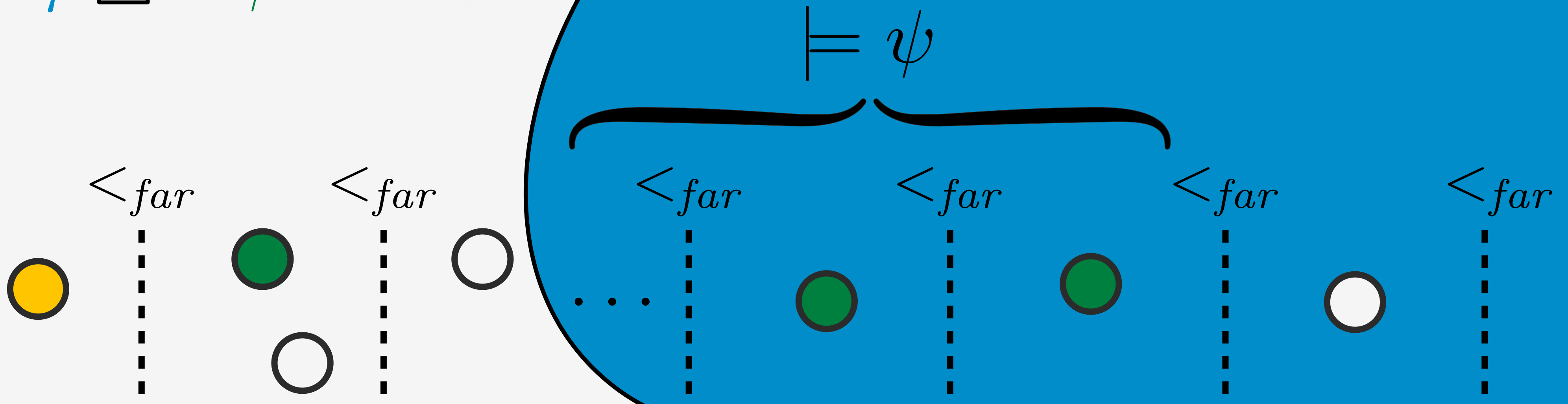
Generally not satisfied and already **rejected by Lewis.**

$x \leq 3.0$ $\quad$ $x > 3.0$

Semantics of ‘Would’

Lewis’ Counterfactuals

$$\bigcirc \models \varphi \Box\!\!\rightarrow \psi \qquad \text{iff:}$$

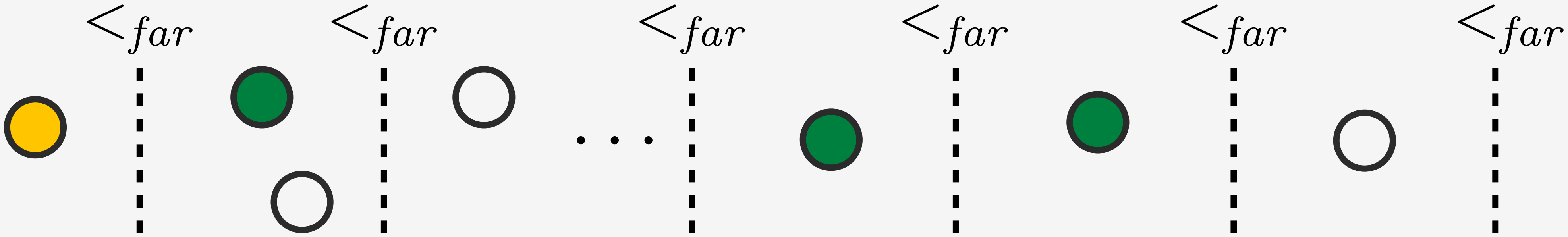


There is a threshold world after which all closer φ-worlds satisfy ψ .

(1) $\exists \bullet : \bullet \models \varphi \wedge \forall \bigcirc : \bigcirc \leq_{far} \bullet \Rightarrow (\bigcirc \models \varphi \rightarrow \psi)$

Semantics of 'Would'

● $\models \varphi \,\Box\!\!\rightarrow \psi$ iff:



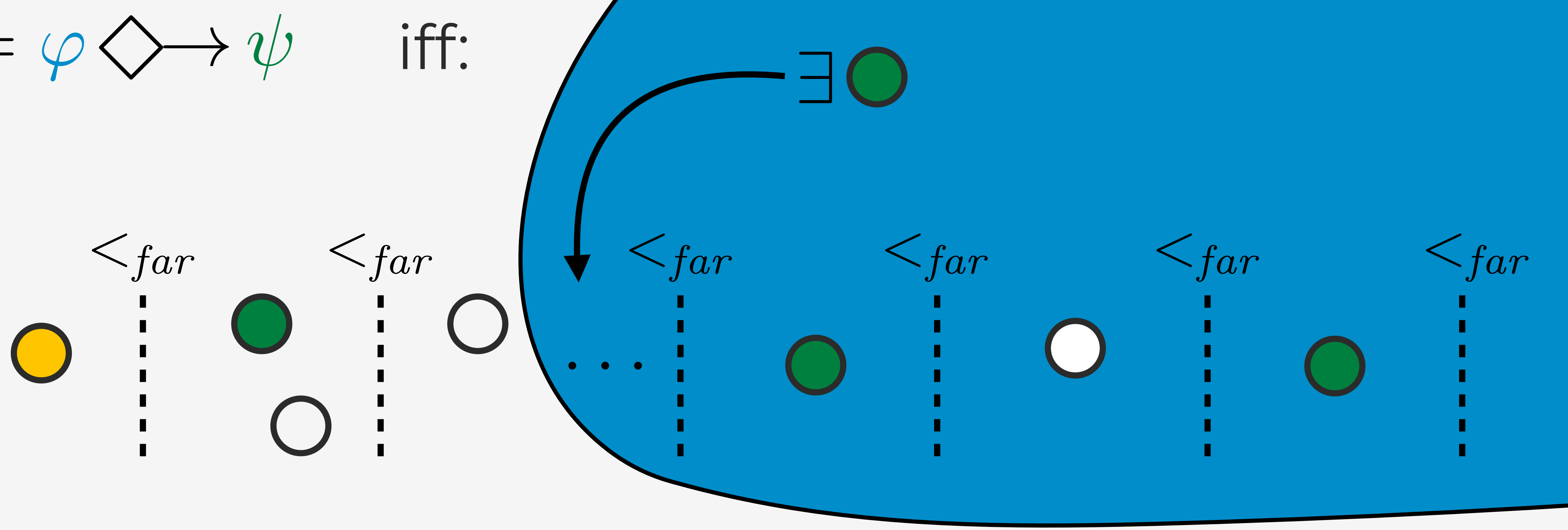
Or: There are no φ-worlds (vacuous case).

(1) or (2): $\forall$○ : ○ $\not\models \varphi$

Semantics of ‘Might’

Lewis’ Counterfactuals

$\bigcirc \models \varphi \Diamond\!\!\rightarrow \psi$ iff:

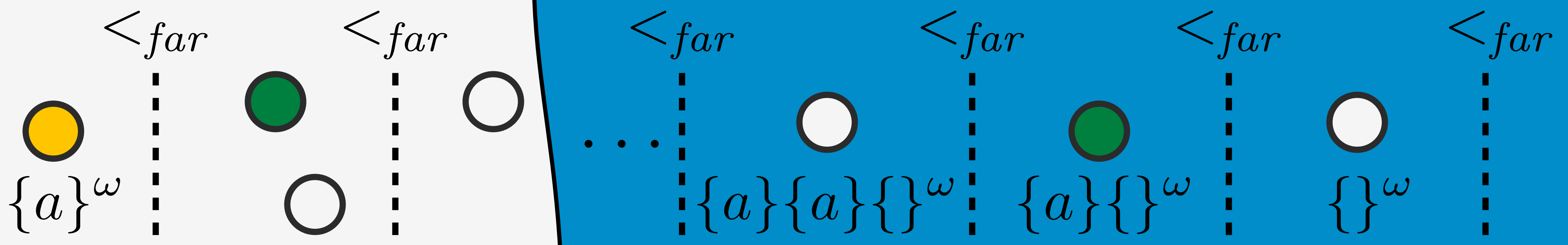


For any φ -world (and there is at least one) there exists a closer world satisfying ψ .

$$\exists \bigcirc : \bigcirc \models \varphi \wedge \forall \bigcirc : \bigcirc \models \varphi \Rightarrow \exists \bigcirc : \bigcirc \leq_{far} \bigcirc \wedge \bigcirc \models \varphi$$

The Limit Assumption

"There always exist well-defined closest worlds satisfying φ *."*

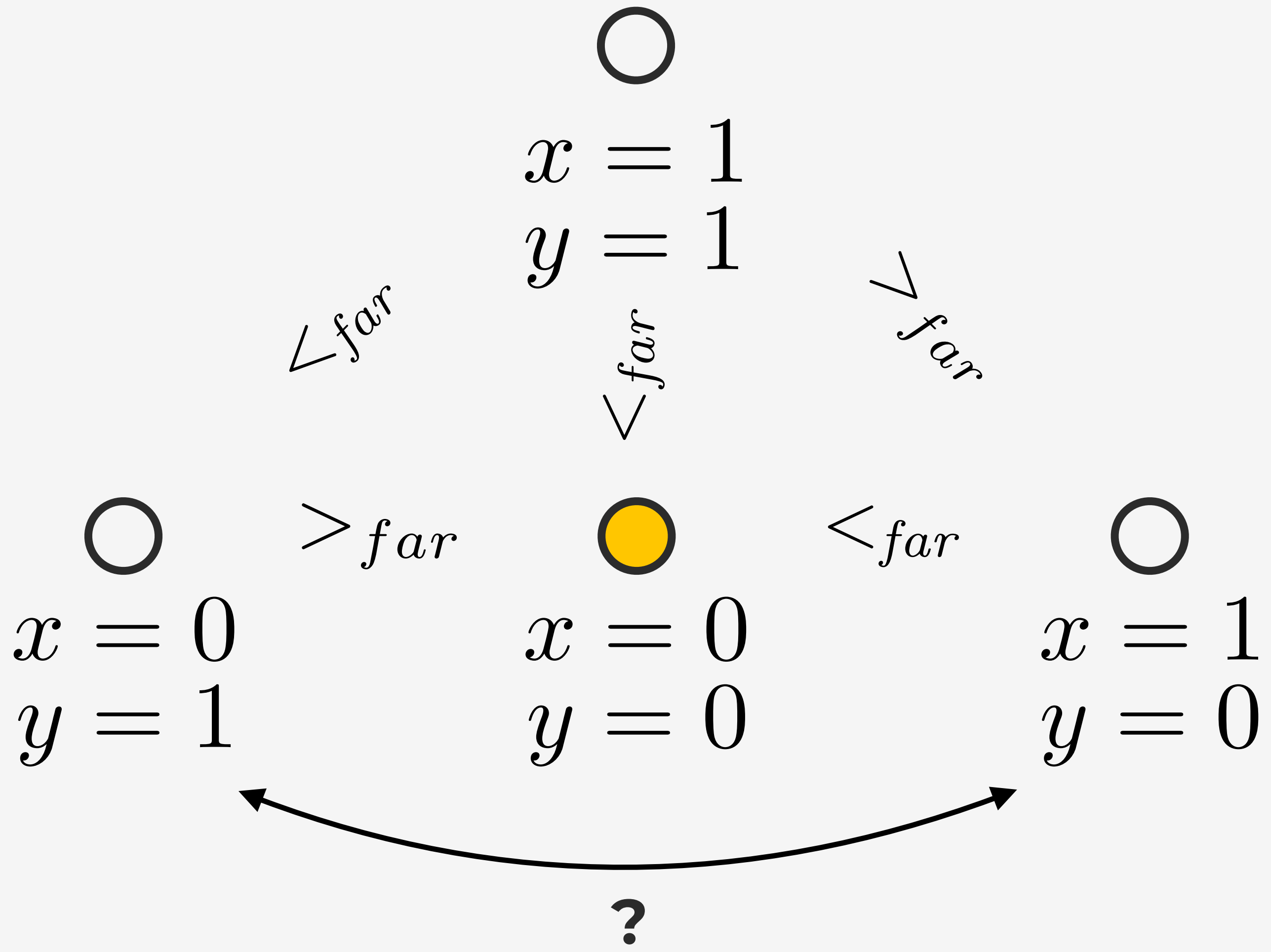


Also does not hold for temporal properties.

$\mathbf{G}\,\mathbf{F}\,a$ $\quad$ $\mathbf{F}\,\mathbf{G}\,\neg a$

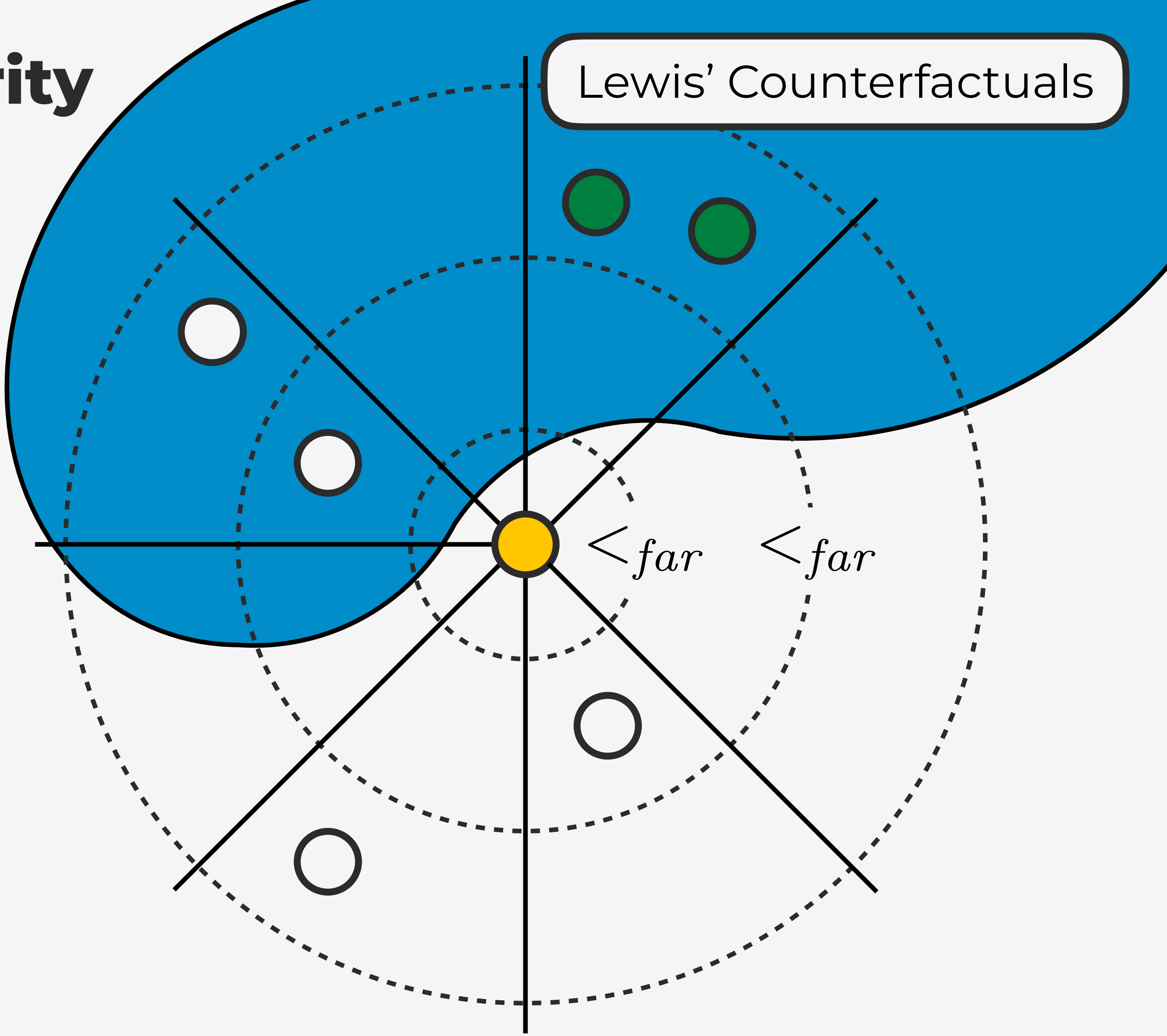
The Problem with Linearity

Lewis' Counterfactuals



In practice, counterfactual worlds are often incomparable.

The Problem with Linearity

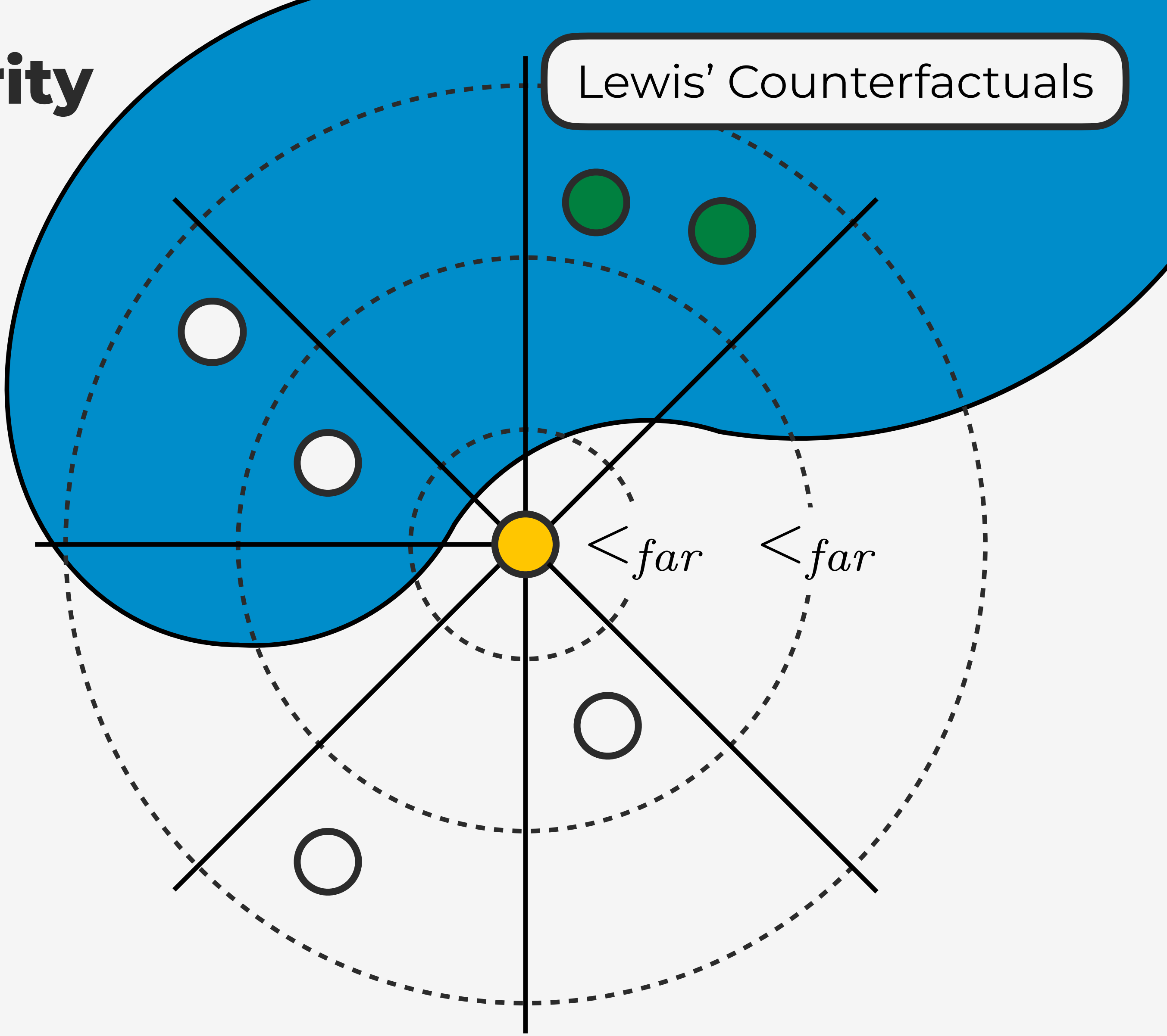


Applying Lewis' semantics to non-total orders leads to unintuitive judgements.

$\bullet \models \varphi \;\Box\!\!\rightarrow \psi$ ✔ iff:

(1) $\exists \bullet : \bullet \models \varphi \land \forall \circ : \circ \leq_{far} \bullet \Rightarrow (\circ \models \varphi \rightarrow \psi)$

The Problem with Linearity



Applying Lewis' semantics to non-total orders leads to unintuitive judgements.

$\bullet \not\models \varphi \mathrel{\Diamond\!\!\rightarrow} \psi$ iff:

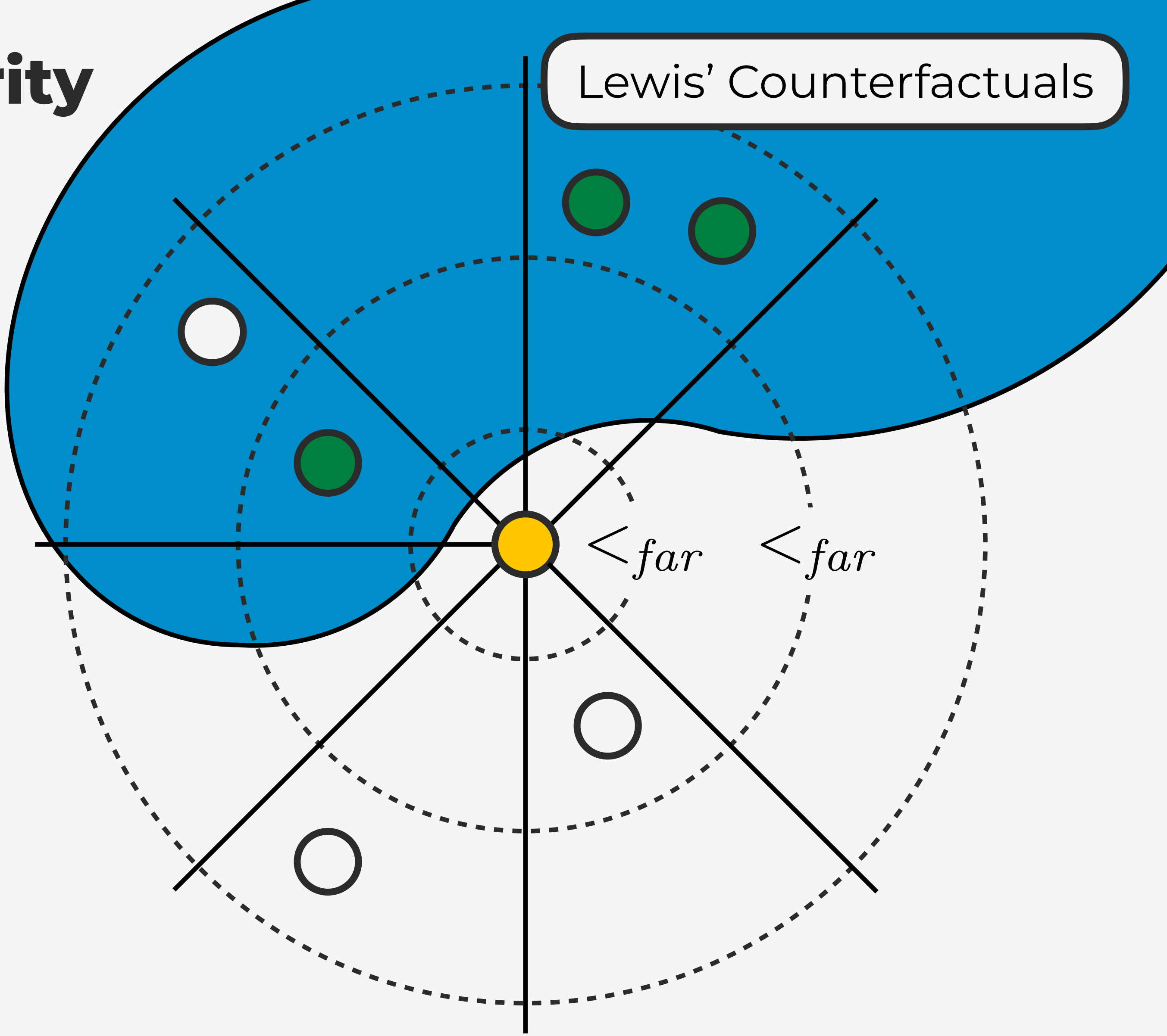
$$\exists \bigcirc : \bigcirc \models \varphi \land \forall \bullet : \bullet \models \varphi \Rightarrow \exists \bullet : \bullet \leq_{far} \bullet \land \bullet \models \varphi$$

The Problem with Linearity

Applying Lewis' semantics to non-total orders leads to unintuitive judgements.

$\bigcirc \models \varphi \mathrel{\Box\!\!\rightarrow} \psi$ iff:

$$\exists \bigcirc : \bigcirc \models \varphi \land \forall \bigcirc : \bigcirc \models \varphi \Rightarrow \exists \bigcirc : \bigcirc \leq_{far} \bigcirc \land \bigcirc \models \varphi$$



Fixing Lewis' Semantics

Non-total similarity relations

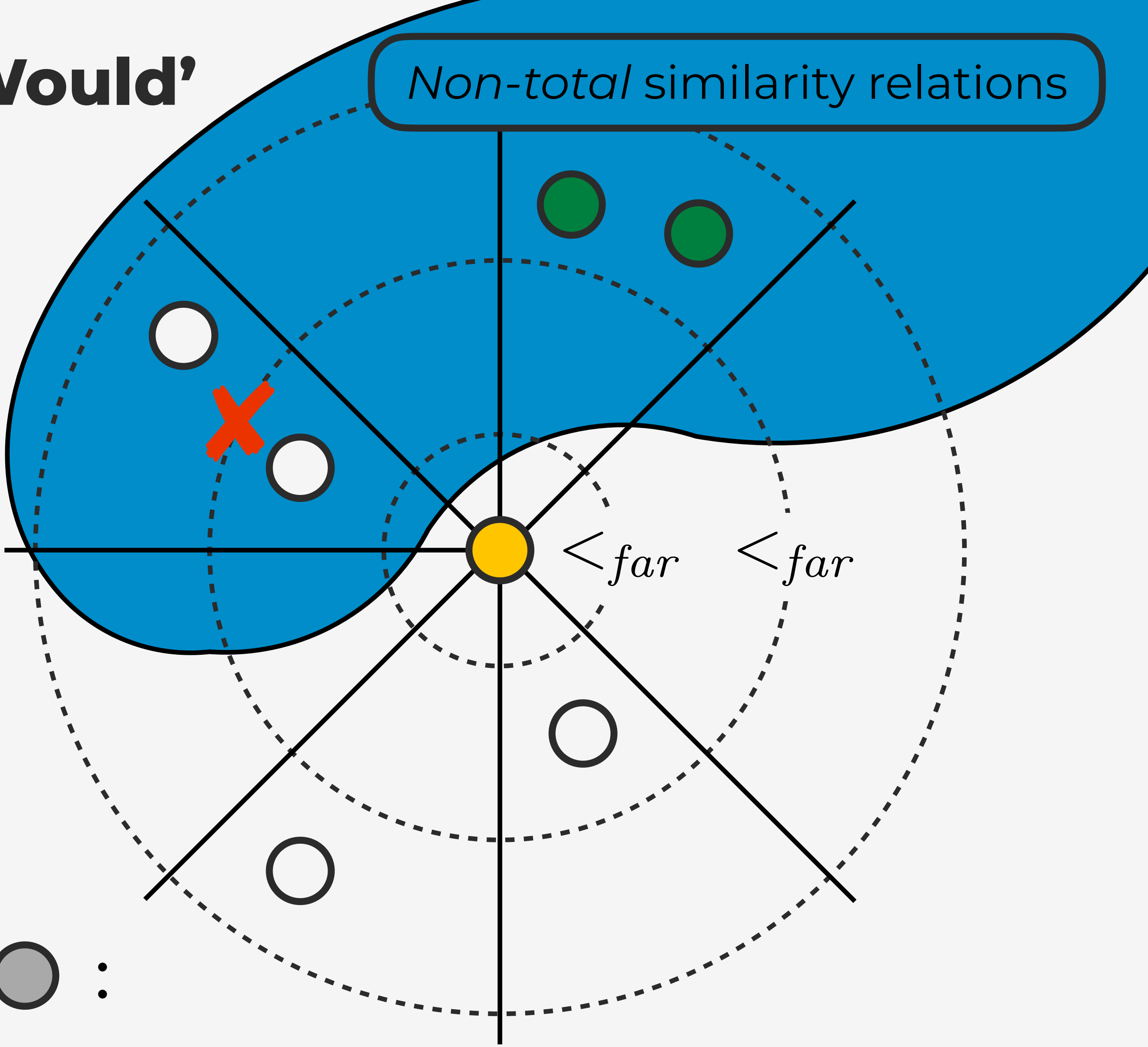
The semantics of $\Box\!\!\rightarrow$ is too weak and of $\Diamond\!\!\rightarrow$ too strong to capture the intended meaning on non-total relations.

We introduce operators with an additional level of quantification:

$\boxdot\!\!\rightarrow$ 'Universal Would'
*"If [...], **under all circumstances**, ψ would have been true as well."*

$\Diamond\!\!\!\cdot\!\rightarrow$ 'Existential Might'
*"If [...], **under some circumstance**, ψ might have been true, too."*

Semantics of ‘Universal Would’



We quantify universally over φ-worlds and require a closer threshold world ${\color{green}\bullet}$ for all of them.

${\color{orange}\bullet} \models \varphi \boxdot\!\!\rightarrow \psi$ iff:

$$\forall \bigcirc : \bigcirc \models \varphi \Rightarrow \exists {\color{green}\bullet} :$$

$${\color{green}\bullet} \leq_{far} \bigcirc \wedge {\color{green}\bullet} \models \varphi \wedge \forall {\color{gray}\bullet} :$$

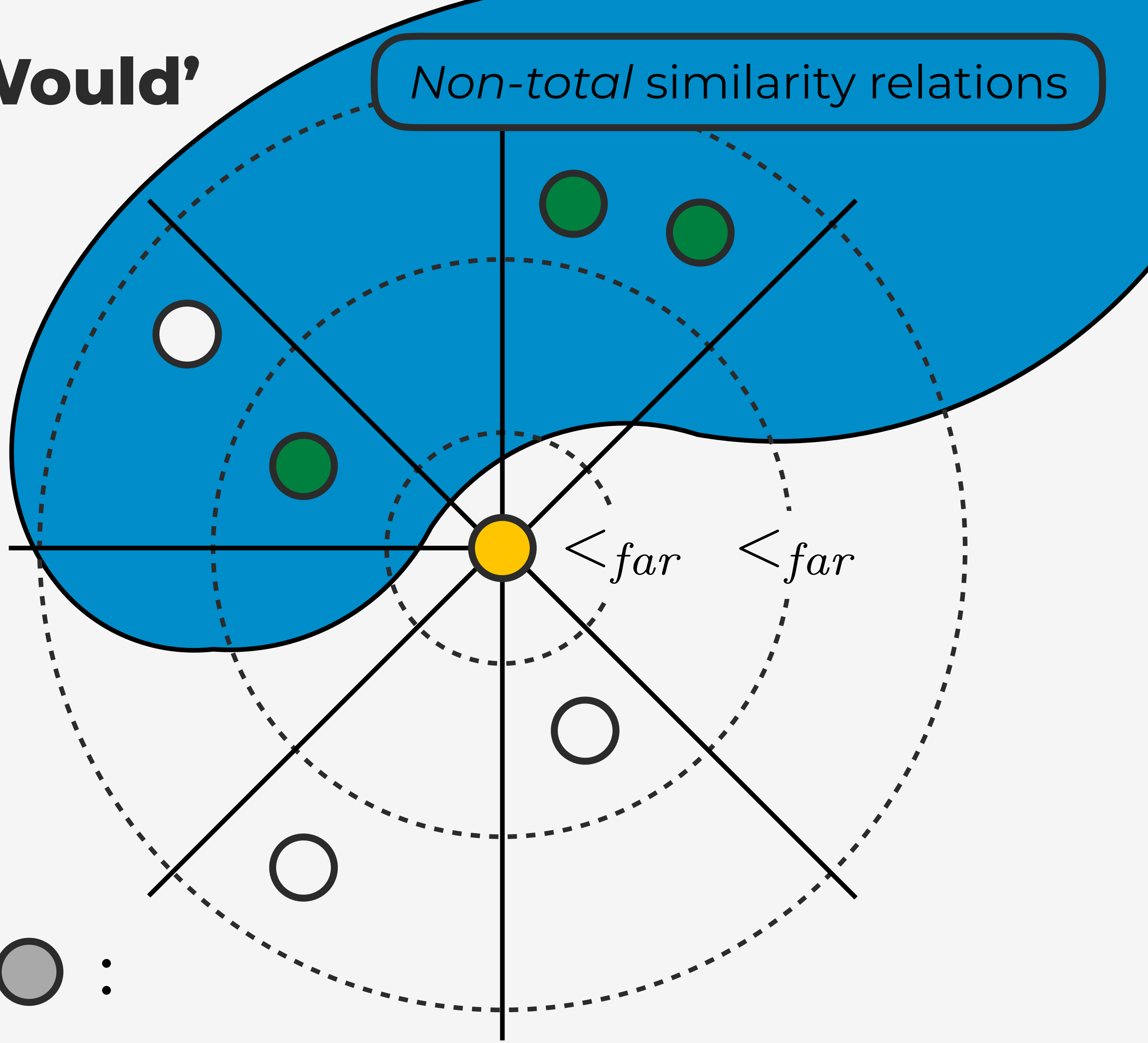
$${\color{gray}\bullet} \leq_{far} {\color{green}\bullet} \Rightarrow ({\color{gray}\bullet} \models \varphi \rightarrow \psi)$$

Semantics of ‘Universal Would’

We quantify universally over φ-worlds and require a closer threshold world $\textcolor{green}{\bullet}$ for all of them.

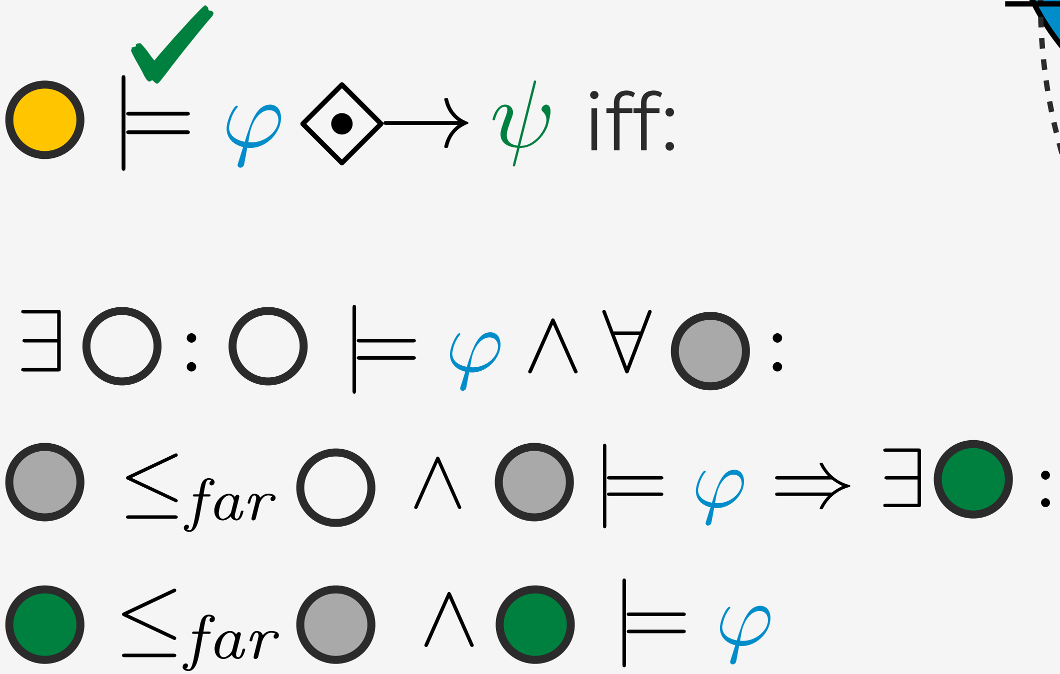
$$\textcolor{yellow}{\bullet} \models \varphi \boxdot\!\!\rightarrow \psi \text{ iff:}$$

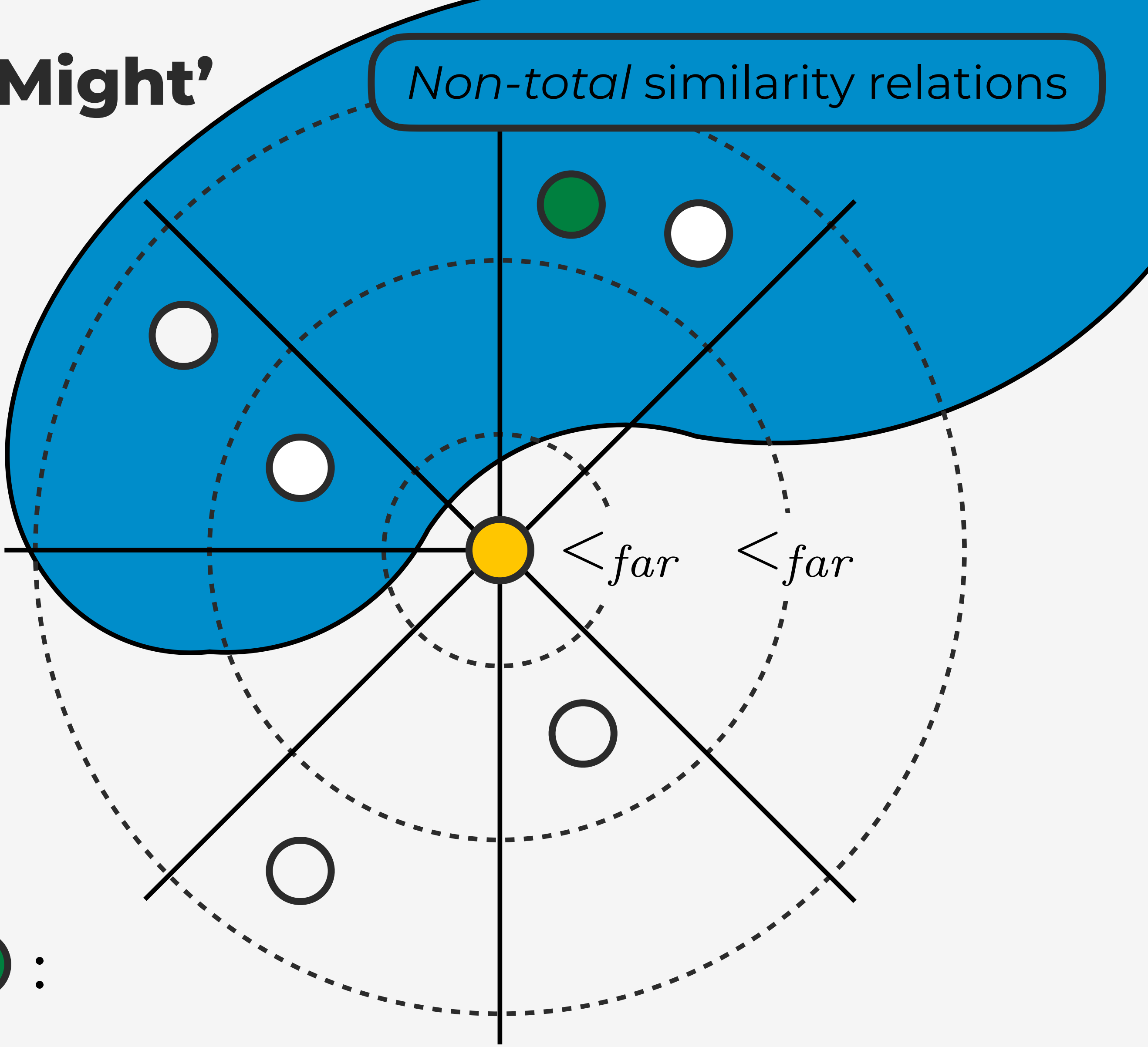
$$\forall \bigcirc : \bigcirc \models \varphi \Rightarrow \exists \textcolor{green}{\bullet} :$$
$$\textcolor{green}{\bullet} \leq_{far} \bigcirc \wedge \textcolor{green}{\bullet} \models \varphi \wedge \forall \textcolor{gray}{\bullet} :$$
$$\textcolor{gray}{\bullet} \leq_{far} \textcolor{green}{\bullet} \Rightarrow (\textcolor{gray}{\bullet} \models \varphi \rightarrow \psi)$$



Semantics of ‘Existential Might’

We quantify existentially over φ-worlds and require ever-closer worlds for one world only.





Minimality

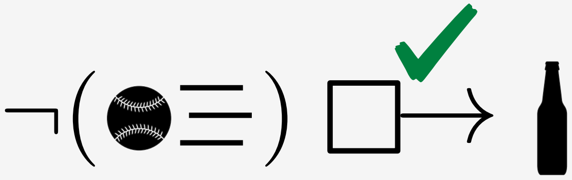
Minimal counterfactuals

For causality, causes are counterfactuals that describe the *minimal changes necessary* to avoid the effect.

Ball is thrown and bottle breaks.

Ball is not thrown.

Ball is not thrown and is red.



Antecedent is not *minimal*.

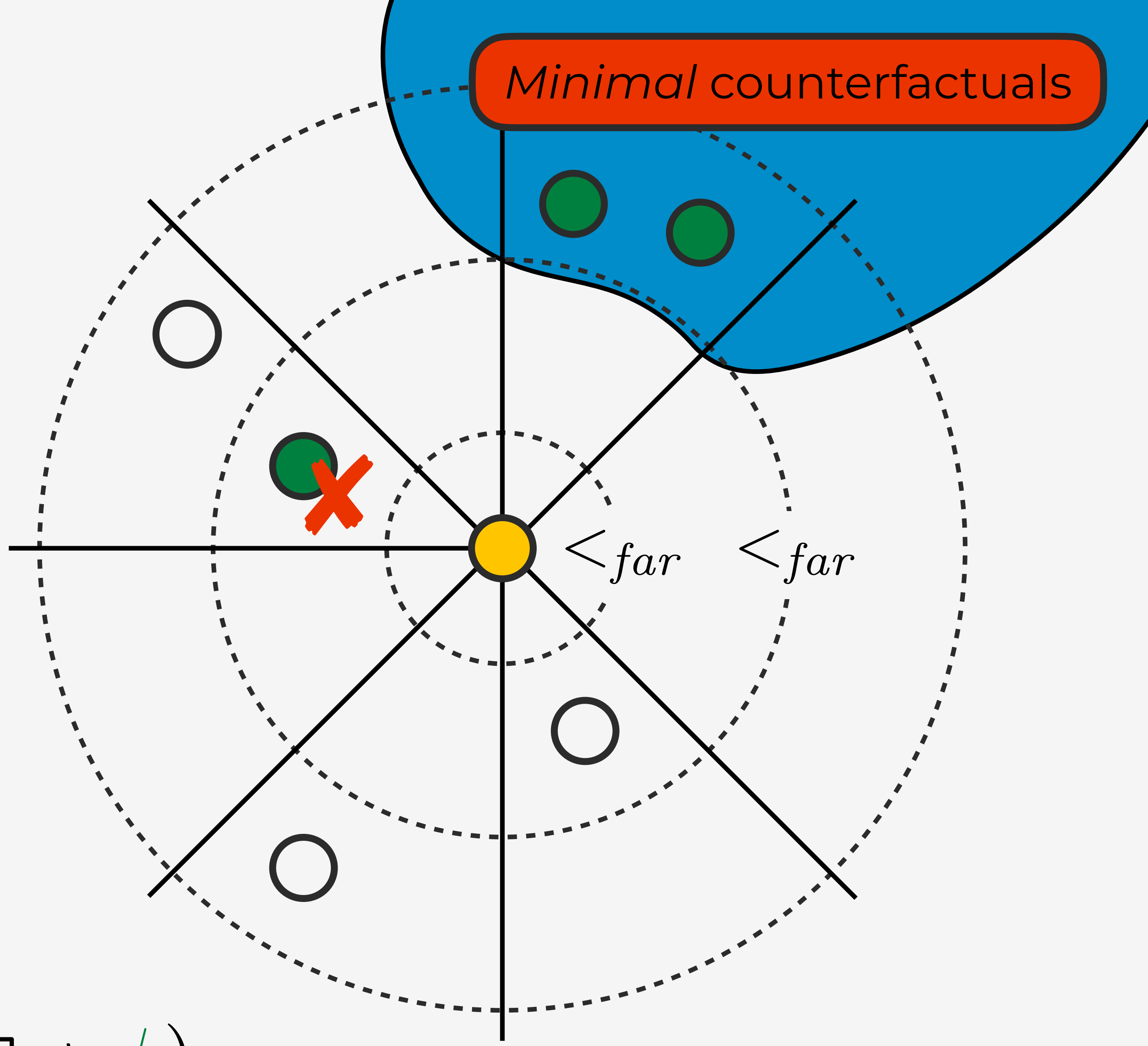
Minimal Counterfactuals

A counterfactual is minimal if its antecedent describes the largest set that qualifies, e.g.,

$$\bigcirc \models \varphi \mathrel{\boxdot\!\!\rightarrow}_{min} \psi \text{ iff:}$$

$$\bigcirc \models \varphi \mathrel{\boxdot\!\!\rightarrow} \psi \wedge$$

$$\neg\exists\theta : \theta \supset \varphi \wedge (\bigcirc \models \theta \mathrel{\boxdot\!\!\rightarrow} \psi)$$



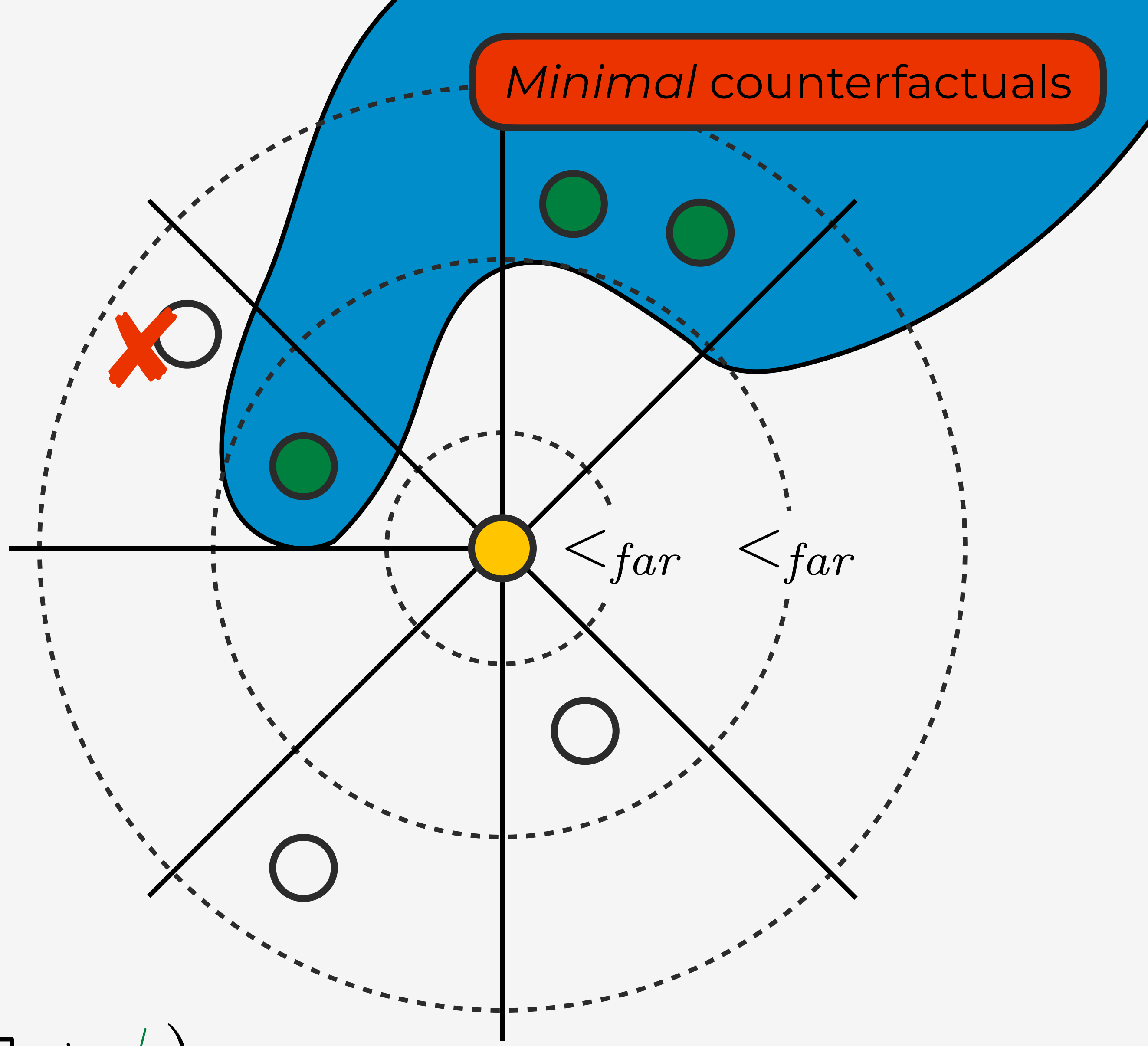
Minimal Counterfactuals

A counterfactual is minimal if its antecedent describes the largest set that qualifies, e.g.,

$$\bullet \models \varphi \boxdot\!\!\rightarrow_{min} \psi \quad \text{iff:}$$

$$\bullet \models \varphi \boxdot\!\!\rightarrow \psi \; \wedge$$

$$\neg \exists \theta : \theta \supset \varphi \wedge (\bullet \models \theta \boxdot\!\!\rightarrow \psi)$$



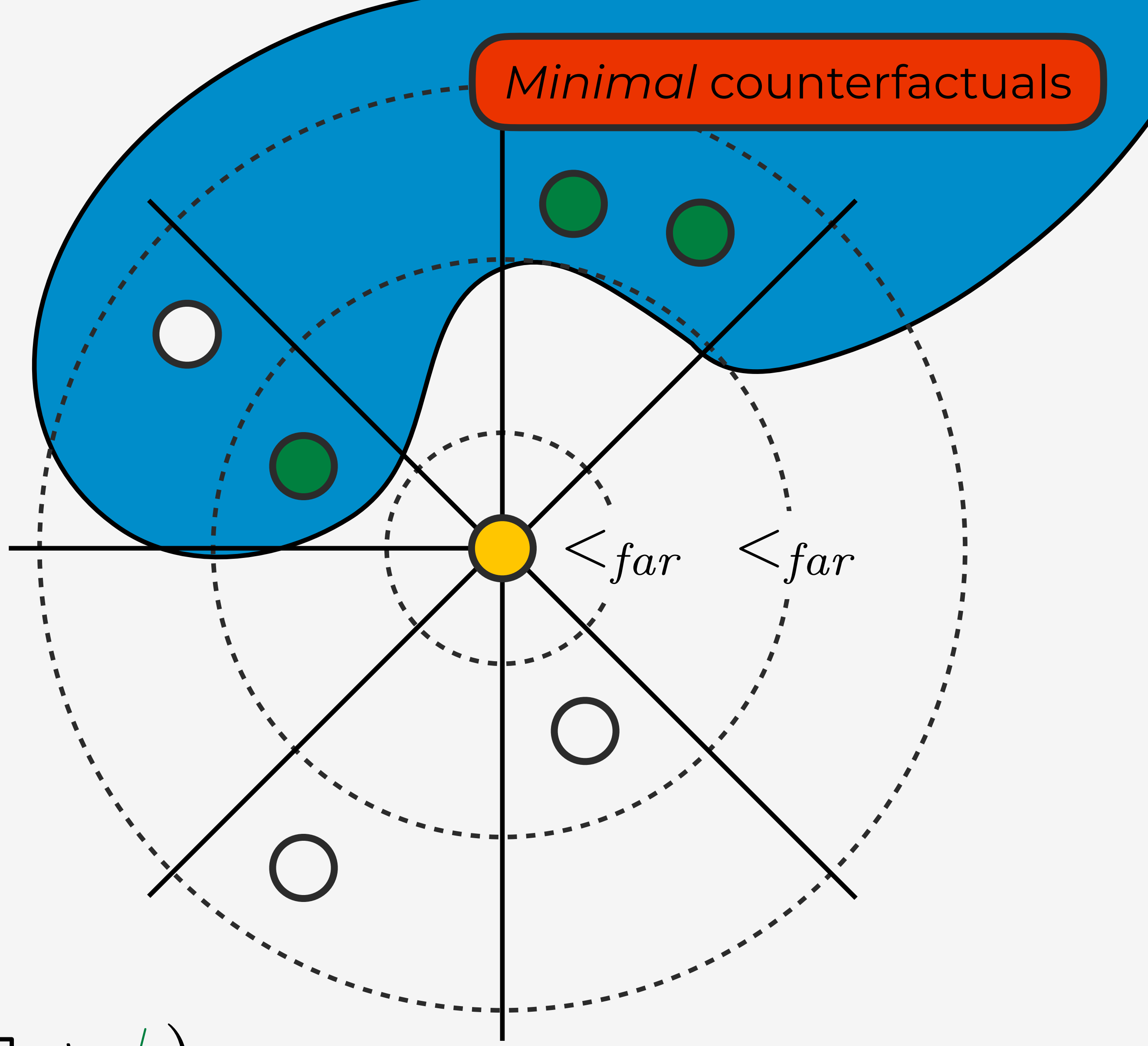
Minimal Counterfactuals

A counterfactual is minimal if its antecedent describes the largest set that qualifies, e.g.,

$\bigcirc \models \varphi \;\square\!\!\rightarrow_{min}\; \psi$ iff:

$\bigcirc \models \varphi \;\square\!\!\rightarrow\; \psi \;\wedge$

$\neg \exists \theta : \theta \supset \varphi \wedge (\bigcirc \models \theta \;\square\!\!\rightarrow\; \psi)$

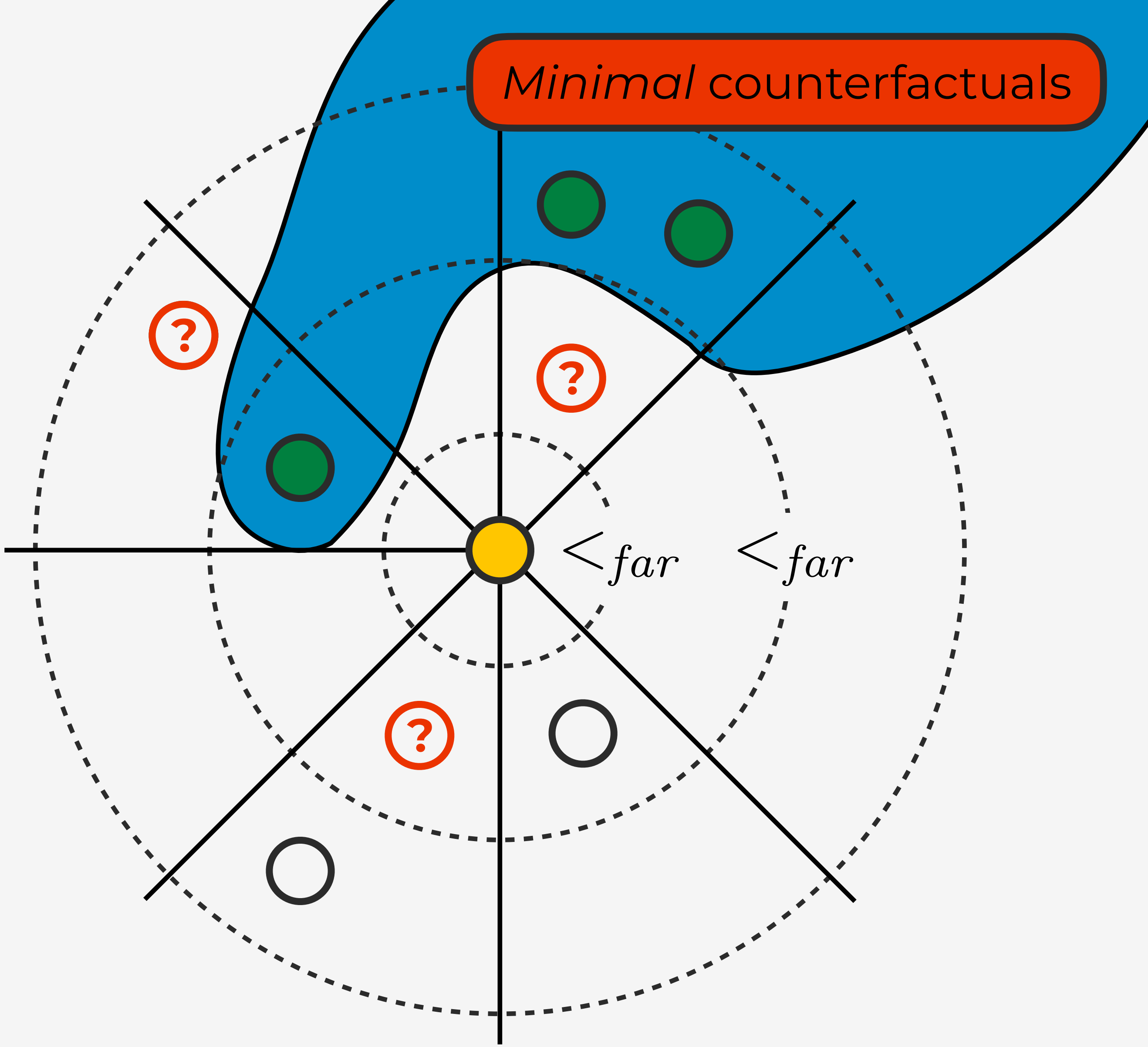


Minimal Counterfactuals

The SO quantification can be avoided by searching for worlds ? that could be added to φ.

$\bullet \models \varphi \boxdot\!\!\rightarrow_{min} \psi$ iff:

Large FO-formula ($\forall / \exists \bigcirc$)
See our paper!



Moving to Temporal Logics

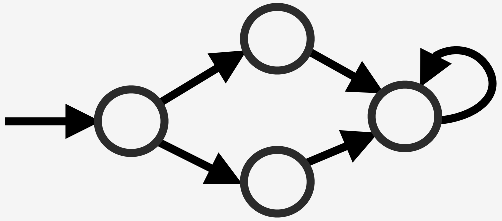
Counterfactual TL

$\varphi \mathrel{\Box\!\!\rightarrow} \psi$ ➡ $(straight) \mathrel{\Box\!\!\rightarrow} (\mathbf{F}\ goal)$

● ➡ lasso trace, e.g. $\pi = \{a,b\}\{\}(\{a,b\})^\omega$

○ ➡ infinite traces, e.g. $\pi' = \{a,b\}\{a,b\}\ldots$

$\sigma \leq_{far}(\pi)\ \rho$ ➡ QPTL formula, e.g. $\bigwedge_{a \in AP} (a_\sigma \not\leftrightarrow a_\pi) \rightarrow (a_\rho \not\leftrightarrow a_\pi)$

◌ ➡ 

$\forall/\exists$ ○ ➡ $\forall/\exists\ \pi$ (hyperproperty)

Satisfiability

Counterfactual TL

$(\mathtt{X}\, a \,\Box\!\!\rightarrow_{min} \mathtt{G}\, b) \wedge \dots$

Counterfactuals Modulo QPTL

$\forall/\exists \bigcirc$

FO formula

$\forall/\exists\, \pi$

Prenex HyperQPTL

HyperQPTL Model Checking
[Beutner&Finkbeiner, LPAR ’23]

Trace Checking

Counterfactual TL

$$(\mathtt{X}\, a \mathrel{\Box\!\!\rightarrow_{min}} \mathtt{G}\, b) \wedge \dots$$

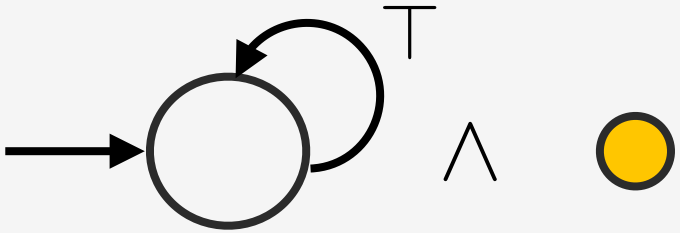
Counterfactuals Modulo QPTL

$$\forall/\exists \bigcirc$$

FO formula

$$\forall/\exists\ \pi$$

Prenex HyperQPTL



HyperQPTL Model Checking
[Beutner&Finkbeiner, LPAR ’23]

Conclusion

Uniform specification language for counterfactual reasoning in, e.g., causality, fairness etc.

Automatic decision procedures for the resulting theory modulo QPTL.

System-Level Counterfactuals
@LPAR: Counterfactuals Modulo Theories?