6(b). Beyond Identifying τ

ISS5096 || ECI
Jaewon (“Jay-one”) Yoo
National Tsing Hua University

1/ Beyond Identifying τ

Cf. Shmueli, Martens, Yoo, & Greene (2025), *From What Ifs to Insights: Counterfactuals in CI and XAI*; Liao, Yoo, Yang, & Chen, *Structure-Aware Robust Counterfactual Explanations*.

Recap: DAG as a Causal Language

- Course-long question: how to **identify** a causal parameter

$$\tau = f\left(Y_i(1), Y_i(0)\right) \quad \text{(e.g., ATE, ATT, CATE)}$$

  - *Identification* = rewrite a counterfactual quantity as a function of **observable** quantities.

- Two languages give us machinery to do this:
  - **Potential outcomes** (Rubin): write $Y_i(d)$; eliminate via *ignorability* + *consistency*.
  - **DAGs** (Pearl, today): write $\mathbb{P}(Y \mid do(D))$; eliminate $do(\cdot)$ via *backdoor* / *frontdoor* / *do-calculus*.

- In both, identification is a *symbol-manipulation game*: counterfactual operator $\rightsquigarrow$ observable expression.

What Else Can a (Learned) DAG Buy Us?

- Rubin CM is *tight* around $(Y(1), Y(0), D, \mathbf{X})$, focused on one $D \longrightarrow Y$ channel.
- A DAG encodes the **entire causal model** around that channel:
  - feature-to-feature dependencies, mediators, multiple outcomes, multiple intervention points.
- Given a (causally valid) **learned DAG**, we can do more than report $\widehat{\tau}$:
  - Decompose direct / indirect effects (mediation, §14).
  - Evaluate hypothetical policies.
  - **Counterfactual recourse / explanation**, "what must change for the outcome to flip?" at the *individual* level.
- *Warning*: recourse is **not** an estimation problem, it's a *search* problem on a fitted predictor. But its **validity** depends on the causal structure.

Two "What Ifs": Population vs. Individual

Same word, different objects, different targets.

- **CI's counterfactual**: $Y_i(1)$ vs. $Y_i(0)$ on the same unit $\rightsquigarrow$ population-level $\widehat{\tau}$.
- **XAI's counterfactual explanation** (Wachter, Mittelstadt, & Russell, 2017): for *this* individual, what is the smallest input change $\mathbf{x}'$ that flips $\widehat{f}(\mathbf{x})$?
  - Concrete: *"If your income were $50K (vs. $45K) and debt $10K (vs. $15K), you would be approved."*
  - A *search* problem on a fitted predictor, not an estimation problem.

The validity problem: $\widehat{f}$ rides on correlation. A minimal perturbation that flips the prediction may be *causally invalid*, "reduce your age to get approved."

Why this lives in a DAG lecture: the diagnosis is *structural*, whether a prescribed change is actionable depends on the DAG, not on $\widehat{f}$.

The Validity Gap, Concretely

- Karimi, Schölkopf, & Valera (2021): recourse should be an **intervention on a causal model**, not a perturbation of features. Structure-free recourse (Wachter et al., 2017) can prescribe actions incompatible with the data-generating process.
- **The gap is sharpest for *sparsity*-optimizing methods** (e.g., DICE, Mothilal, Sharma, & Tan, 2020): by design, they minimize the *number* of features changed.
  - Example: *"raise income by $10K, leave education, occupation, age unchanged."*
  - But income co-varies with education and occupation in the DAG $\rightsquigarrow$ the prescription is structurally impossible as an intervention.
- Empirical-CI framing (Shmueli, Martens, Yoo, & Greene, 2025): XAI's "what-ifs" become actionable *insights* only when grounded in the identification machinery used for $\widehat{\tau}$.

Structural Recourse

- One concrete instance, Liao, Yoo, Yang, & Chen: embed the DAG *inside* the classifier (**Conditional Gaussian Network Classifier**). Feature dependencies become constraints on the recourse search:

$$\arg\min_{\mathbf{x}'} \; d(\mathbf{x}^{\text{fac}}, \mathbf{x}') \quad \text{s.t.} \quad \widehat{H}(\mathbf{x}') \geq \tau', \quad \mathbf{x}' \text{ respects the DAG.}$$

- Evaluating any recourse method against its DAG, a proposed **structural violation (SV) metric** (Liao et al.): for each modified feature, at least one of its neighbors in the dependency graph must also move; otherwise, count the modification as a violation.
  - Directly penalizes the sparsity-by-ignorance pathology; model-agnostic (applies to NN-based and CGNC-based CEs alike).

- **Payoff of the DAG, restated**: the same structure that identifies $\widehat{\tau}$ (*description*, population-level) also disciplines recourse (*prescription*, individual-level), and gives us a yardstick for holding CE methods accountable to that structure.