ICLR 2024 SPOTLIGHT

# Procedural Fairness Through Decoupling Objectionable Data Generating Components

**Zeyu Tang[1], Jialu Wang[2], Yang Liu[2 4], Peter Spirtes[1], Kun Zhang[1 3]**

zeyutang@cmu.edu {faldict, yangliu}@ucsc.edu {ps7z@andrew., kunz1@}cmu.edu

[1]Carnegie Mellon University
[2]University of California, Santa Cruz
[3]MBZUAI
[4]ByteDance Research

# In a nutshell

- We reveal the frequently over-looked issue of ***disguised procedural unfairness***

- We propose a framework to address procedural fairness
  - **value instantiation rule**
  - **reference point**

- We determine the value of reference point per Rawls's ***pure procedural justice*** principles

# What is procedural fairness or justice?

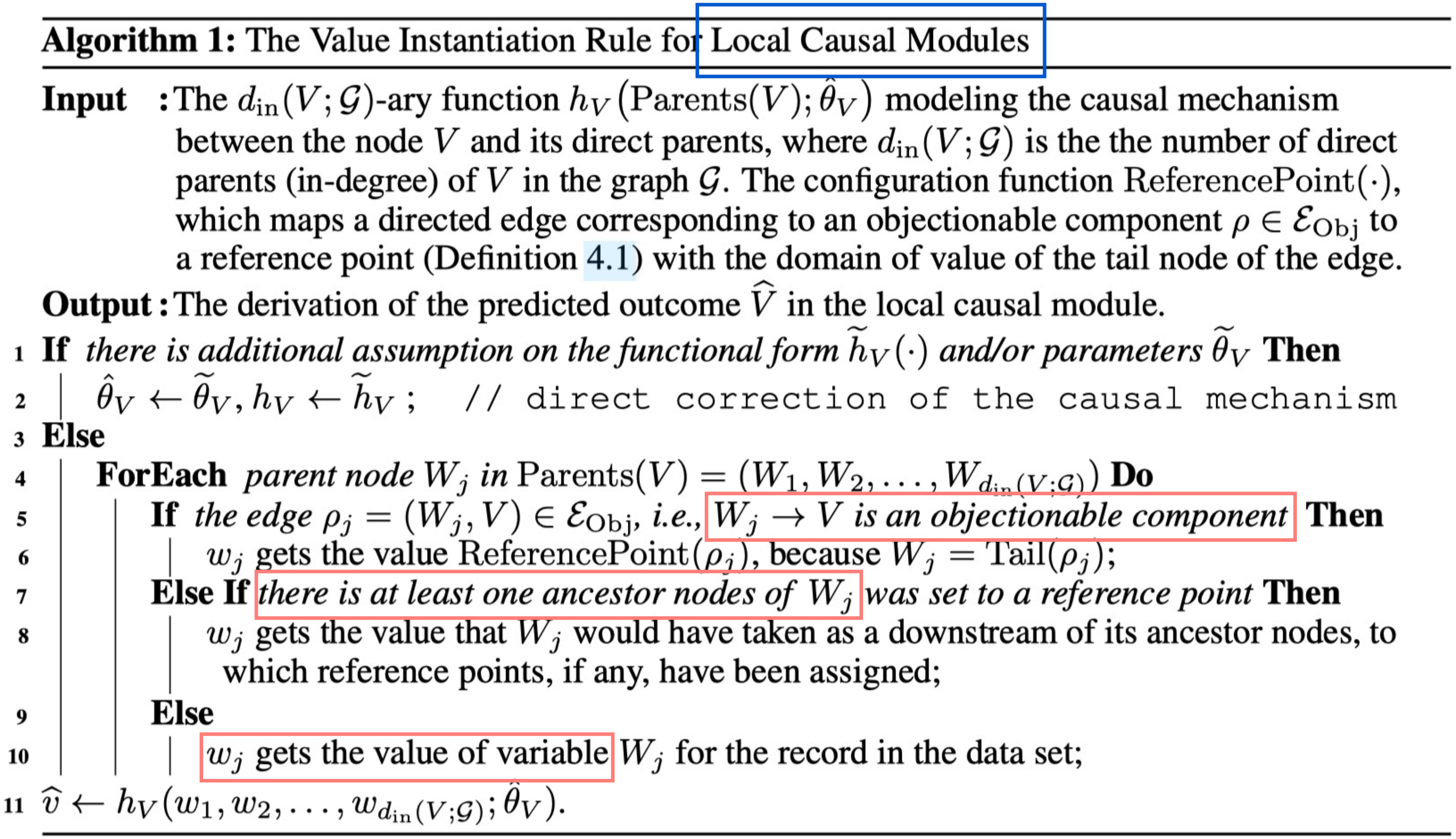
- Three kinds of procedural justice (Rawls 1971;2001)
  - perfect procedural justice
  - imperfect procedural justice
  - pure procedural justice

# Principles of *pure procedural justice*

- **Requirement I: Fair Equality of Opportunity**
  - *The opportunity should be open and attainable, with the same prospects of success, for those who are at the same level of talent and ability, and have the same willingness to use them. Such equality of opportunity should not be influenced by arbitrary contingencies.*

- **Requirement II: The Difference Principle**
  - *The (social and economic) inequalities are to be arranged so that they are to the greatest benefit to the least advantaged members of the society.*

# Our framework (part 1) value instantiation rule



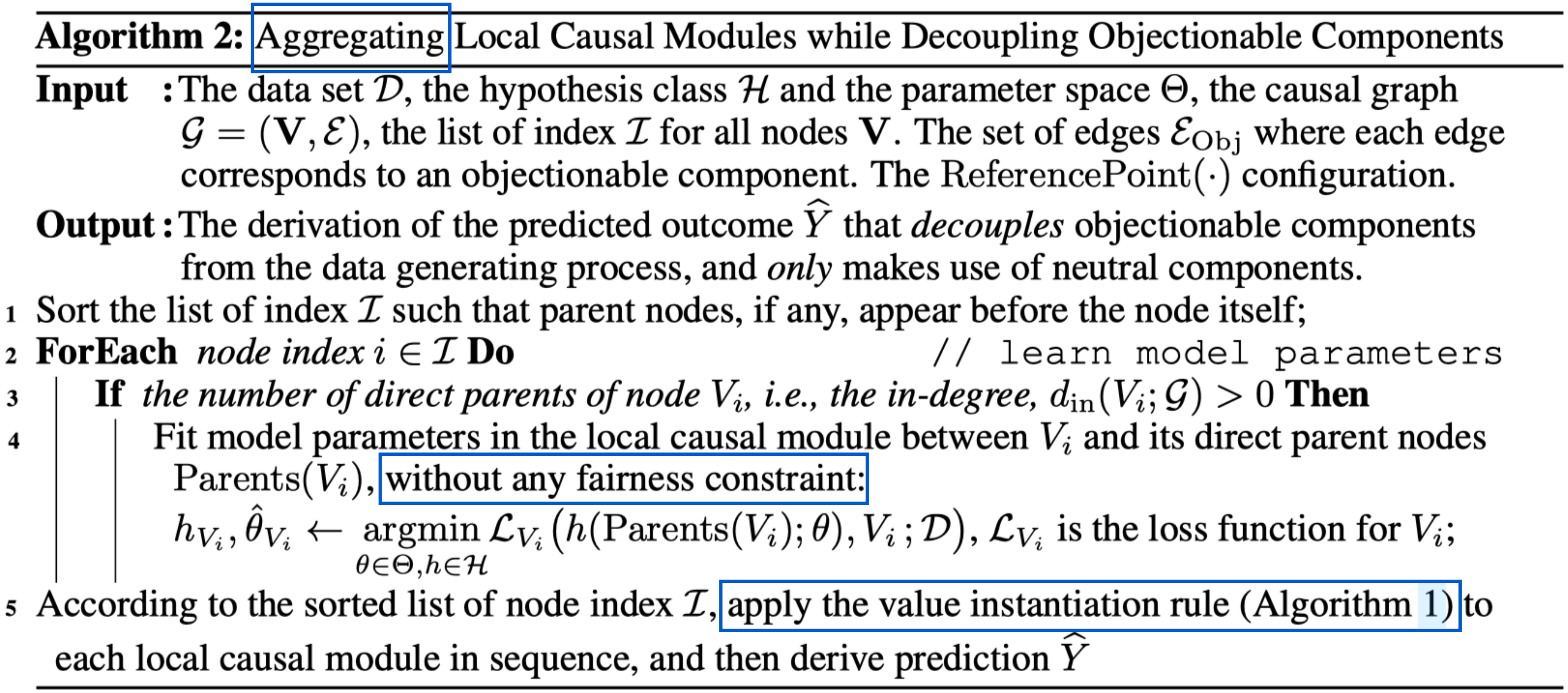
**Algorithm 1:** The Value Instantiation Rule for Local Causal Modules

**Input :** The $d_{\text{in}}(V;\mathcal{G})$-ary function $h_V\left(\text{Parents}(V);\hat{\theta}_V\right)$ modeling the causal mechanism between the node $V$ and its direct parents, where $d_{\text{in}}(V;\mathcal{G})$ is the the number of direct parents (in-degree) of $V$ in the graph $\mathcal{G}$. The configuration function $\text{ReferencePoint}(\cdot)$, which maps a directed edge corresponding to an objectionable component $\rho \in \mathcal{E}_{\text{Obj}}$ to a reference point (Definition 4.1) with the domain of value of the tail node of the edge.

**Output :** The derivation of the predicted outcome $\widehat{V}$ in the local causal module.

```
If there is additional assumption on the functional form $\widetilde{h}_V(\cdot)$ and/or parameters $\widetilde{\theta}_V$ Then
    $\hat{\theta}_V \leftarrow \widetilde{\theta}_V, h_V \leftarrow \widetilde{h}_V$ ;   // direct correction of the causal mechanism
Else
    ForEach parent node $W_j$ in $\text{Parents}(V) = (W_1, W_2, \ldots, W_{d_{\text{in}}(V;\mathcal{G})})$ Do
        If the edge $\rho_j = (W_j, V) \in \mathcal{E}_{\text{Obj}}$, i.e., $W_j \rightarrow V$ is an objectionable component Then
            $w_j$ gets the value $\text{ReferencePoint}(\rho_j)$, because $W_j = \text{Tail}(\rho_j)$;
        Else If there is at least one ancestor nodes of $W_j$ was set to a reference point Then
            $w_j$ gets the value that $W_j$ would have taken as a downstream of its ancestor nodes, to
              which reference points, if any, have been assigned;
        Else
            $w_j$ gets the value of variable $W_j$ for the record in the data set;
$\widehat{v} \leftarrow h_V(w_1, w_2, \ldots, w_{d_{\text{in}}(V;\mathcal{G})}; \hat{\theta}_V)$.
```

# Our framework (part 2) reference point values



```
Algorithm 2: Aggregating Local Causal Modules while Decoupling Objectionable Components
Input  : The data set $\mathcal{D}$, the hypothesis class $\mathcal{H}$ and the parameter space $\Theta$, the causal graph
         $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, the list of index $\mathcal{I}$ for all nodes $\mathbf{V}$. The set of edges $\mathcal{E}_{\mathtt{Obj}}$ where each edge
         corresponds to an objectionable component. The $\mathrm{ReferencePoint}(\cdot)$ configuration.
Output : The derivation of the predicted outcome $\widehat{Y}$ that *decouples* objectionable components
         from the data generating process, and *only* makes use of neutral components.
Sort the list of index $\mathcal{I}$ such that parent nodes, if any, appear before the node itself;
ForEach node index $i \in \mathcal{I}$ Do                                   // learn model parameters
    If the number of direct parents of node $V_i$, i.e., the in-degree, $d_{\text{in}}(V_i; \mathcal{G}) > 0$ Then
        Fit model parameters in the local causal module between $V_i$ and its direct parent nodes
         $\mathrm{Parents}(V_i)$, without any fairness constraint:
         $h_{V_i}, \hat{\theta}_{V_i} \leftarrow \operatorname*{argmin}_{\theta \in \Theta, h \in \mathcal{H}} \mathcal{L}_{V_i}\big(h(\mathrm{Parents}(V_i); \theta), V_i\,; \mathcal{D}\big)$, $\mathcal{L}_{V_i}$ is the loss function for $V_i$;
According to the sorted list of node index $\mathcal{I}$, apply the value instantiation rule (Algorithm 1) to
 each local causal module in sequence, and then derive prediction $\widehat{Y}$
```

$$\widehat{y} = \mathop{\circ}_{i \in \mathcal{I}} \big(h_{V_i} \circ \mathrm{ReferencePoint}\big)\big(\mathbf{z}; \mathcal{E}_{\mathtt{Obj}}, \hat{\theta}_{V_1}, \ldots, \hat{\theta}_{V_{|\mathcal{I}|}}\big)$$

# Thank you!