Heterogeneous Treatment Effects and Causal Mechanisms
Corrigendum

Jiawei Fu* Tara Slough†

August 27, 2026

This document corrects one error and one imprecision for "Heterogeneous Treatment Effects and Causal Mechanisms" that were detected by an AI review of the manuscript and appendix after online publication.[1] In order of importance:

Proposition 4

On page 10 we write: "We will denote $\mathbf{X}^R = \{X_1, X_2, ..., X_K\}$, as the set of all possible pre-treatment covariates with non-zero effects on the outcome, Y." In Footnote 15, we formalize this by writing: "Formally, if $X_k \in \mathbf{X}^R$, then there exist $x \neq x' \in X_k$ such that $Y(Z, M(Z; X_k = x, X_{\neg k}); X_k = x, X_{\neg k}) \neq Y(Z, M(Z; X_k = x', X_{\neg k}); X_k = x', X_{\neg k})$." Note that this is stronger than our subsequent informal description "Covariates in $\mathbf{X}^R$ can be thought of as 'relevant for predicting outcome Y" (p. 15) because a proxy variable of any variable in $\mathbf{X}^R$ (as defined above) may also predict Y. The less formal notion is used in the proof of Proposition 4 when we write: "By definition of $\mathbf{X}^R$, we know $X \notin \mathbf{X}^R$ implies that X must be independent of Y."

The corrected version of this sentence should be: "We will denote $\mathbf{X}^R = \{X_1, X_2, ..., X_K\}$, as the set of all possible pre-treatment covariates that are statistically associated with the outcome, Y." And Footnote 15 should read: "Formally, if $X_k \in \mathbf{X}^R$, then $X_k \not\perp Y \mid Z$." The proof follows from this corrected definition.

Proposition 1

The existing proof establishes sufficiency of Assumptions 1 and 2. However, the claim of generic necessity and the conditions required for it need to be stated more explicitly.

Meaning of "generically." Here, "generically" refers to variation over exclusion patterns and over the effect contrasts left unrestricted under each pattern, rather than to sampling uncertainty. Each possible exclusion pattern is assigned positive probability and, conditional on that pattern, the unrestricted contrasts have a joint density that is absolutely continuous with respect to Lebesgue measure. Consequently, contrasts set equal to zero by an exclusion restriction occur with positive probability, whereas accidental zeros and exact cancellations among unrestricted contrasts occur only on a null set.

Proof. Define:

$$\delta_0 \equiv \Delta ADE^Y = ADE^Y(z, z'; X_k = x) - ADE^Y(z, z'; X_k = x'),$$
$$\delta_j \equiv \Delta AIE_j^Y = AIE_j^Y(z, z'; X_k = x) - AIE_j^Y(z, z'; X_k = x'), \qquad j = 2, \ldots, J.$$

*Assistant Professor, Duke University jiawei.fu@duke.edu
†Associate Professor, New York University. taraslough@nyu.edu

[1]These errors were detected by GPT-5.6 Sol in response to the prompt: "Check this for serious errors: https://www.cambridge.org/core/journals/american-political-science-review/article/heterogeneous-treatment-effects-and-causal-mechanisms/A4FAFB58D3E9F1197F18E913590CC512." We thank P.M. Aronow for bringing this to our attention. We subsequently used GPT-5.6 Sol to edit and refine this correction and clarification.

The decomposition of the conditional average treatment effect gives

$$CATE^Y(X_k = x) - CATE^Y(X_k = x') - L[\delta_1] = L[\delta_0 + \sum_{j=2}^{J} \delta_j]$$

Let $\delta = (\delta_0, \delta_2, \delta_3, \ldots, \delta_J)' \in \mathbb{R}^J$, which excludes δ_1, the contrast in the focal mechanism's indirect effect. Now let $S = (S_0, S_2, \ldots, S_J) \in \{0,1\}^J$ index all possible exclusion patterns where:

$$S_j = \begin{cases} 0 & \text{the corresponding non-target contrast is excluded} \\ 1 & \text{the corresponding non-target contrast is unrestricted.} \end{cases}$$

Thus $S = 0_J$ corresponds to Assumption 1 and the generalized version of Assumption 2 (Assumption 3 on page 10 of the SI).

Let Π be a reference probability measure for the pair (S, δ), defined as follows. First, draw an exclusion pattern s with probability $\pi_s > 0$, $\sum_{s \in \{0,1\}^J} \pi_s = 1$. Conditional on $S = s$, every coordinate of δ corresponding to a zero coordinate of s is equal to zero. For every $s \neq 0_J$, assume that the coordinates left unrestricted by s have a joint distribution that is absolutely continuous with respect to Lebesgue measure on the space of those unrestricted contrasts.

Let $\mathcal{E}$ denote the equality in Proposition 1: $\mathcal{E} = \Big\{CATE^Y(X_k = x) - CATE^Y(X_k = x') = L[\delta_1]\Big\}$. For any exclusion pattern $s \neq 0_J$, the event $\mathcal{E}$ requires $\sum_{j:s_j=1} \delta_j = 0$. This is a nontrivial linear restriction on the unrestricted coordinates of δ. Its solution set is therefore a proper hyperplane in the space of those coordinates and has Lebesgue measure zero. Absolute continuity implies:

$$\Pi(\mathcal{E} \mid S = s) = 0 \qquad \text{for every } s \neq 0_J.$$

Because there are only finitely many exclusion patterns,

$$\Pi\big(\mathcal{E} \cap \{S \neq 0_J\}\big) = \sum_{s \neq 0_J} \pi_s \Pi(\mathcal{E} \mid S = s) = 0$$

Thus, outside a Π-null set, the equality in Proposition 1 cannot hold under any exclusion pattern that leaves at least one non-target effect contrast unrestricted. It therefore implies $S = 0_J$ Π-almost surely, which is precisely Assumption 1 together with Assumption 3. Hence, the exclusion assumptions are sufficient and generically necessary (as defined above). □

We regret the error and imprecision.