Heterogeneous Treatment Effects and Causal Mechanisms

Jiawei Fu[*] Tara Slough[†]

June 21, 2023

Contents

[*] Ph.D. Candidate, New York University jf3739@nyu.edu
[†] Assistant Professor, New York University. taraslough@nyu.edu

Appendix A Motivating Example

A1.1 Incorrect DAG

Figure A1 depicts the DAG that is evaluated by the HTE analysis in Remarks 1-2. Note that the dashed lines do not correspond to the theoretical model. In the model, only the learning mechanism is active. This mechanism is evauated by examining heterogeneity in treatment effects with respect to voters' prior beliefs.

We note that this graph does not directly correspond to the model in the paper. Here, the the valence shock, v_i is measured pre-treatment (*ex-ante*), and the researcher (wrongly) believes that it moderates treatment effects. We represent this in the graph with $\widetilde{v}_i$, a measure of "*ex-post*" valence.

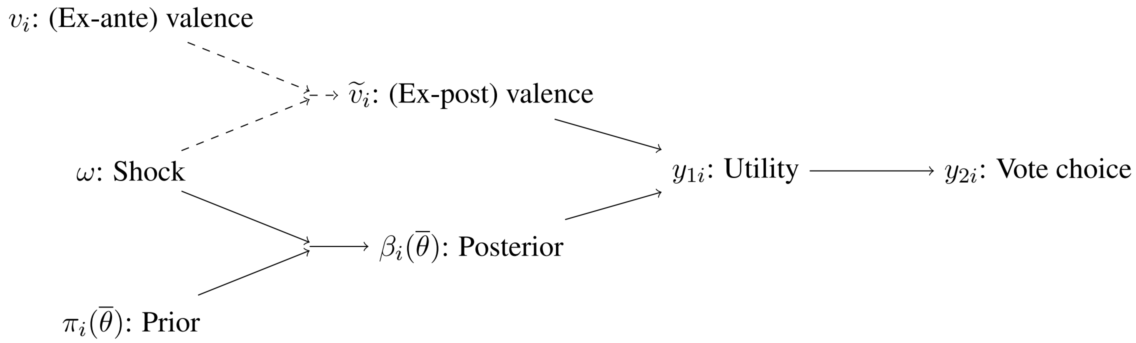

v_i: (Ex-ante) valence
$\widetilde{v}_i$: (Ex-post) valence
ω: Shock
y_{1i}: Utility
y_{2i}: Vote choice
$\beta_i(\overline{\theta})$: Posterior
$\pi_i(\overline{\theta})$: Prior

Figure A1: Incorrect directed acyclic graph representation (relative to the model). The dashed arrows are not implied by the model but correspond to a test of the valence mechanism.

A1.2 Proofs of Remarks 1-2

Remark 1(a) :

Proof.

$$\begin{aligned} CATE(y_1, X = \pi) &= E[y_1|\omega'', \pi] - E[y_1|\omega', \pi] && \text{(A1)} \\ &= E[\beta(\overline{\theta}|\pi, \omega'') + v_i|\omega'', \pi] - E[\beta(\overline{\theta}|\pi, \omega') + v_i|\omega', \pi] && \text{(A2)} \\ &= E[\beta(\overline{\theta}|\pi, \omega'')] - E[\beta(\overline{\theta}|\pi, \omega')] + E[v_i] - E[v_i] && \text{(A3)} \\ &= E[\beta(\overline{\theta}|\pi, \omega'')] - E[\beta(\overline{\theta}|\pi, \omega')] && \text{(A4)} \end{aligned}$$

where equation (A2) to (A3) follows from the linearity of expectations. Similarly, we have:

$$CATE(y_1, X = \pi') = E\beta(\overline{\theta}|\pi', \omega'') - E\beta(\overline{\theta}|\pi', \omega')$$

Recall that $\beta(\overline{\theta} \mid \pi_i, \omega)$ is given by:

$$\beta(\overline{\theta}|\pi_i, \omega) = \frac{\pi_i \phi(g - f(\overline{\theta}, \omega))}{\pi_i \phi(g - f(\overline{\theta}, \omega)) + (1 - \pi_i)\phi(g - f(\underline{\theta}, \omega))} = \frac{1}{1 + \frac{1-\pi_i}{\pi_i} \frac{\phi(g - f(\underline{\theta}, \omega))}{\phi(g - f(\overline{\theta}, \omega))}}$$

Given function f and pdf ϕ, we conclude $CATE(y_1, X = \pi) \neq CATE(y_1, X = \pi')$.

□

Remark 1(b) :

Proof.

$$
\begin{aligned}
CATE(y_1, v) &= E[y_1|\omega'', v] - E[y_1|\omega', v] && \text{(A5)}\\
&= E[\beta(\overline{\theta}|\pi, \omega) + v|\omega'', v] - E[\beta(\overline{\theta}|\pi, \omega) + v|\omega', v] && \text{(A6)}\\
&= E[\beta(\overline{\theta}|\pi, \omega'')] - E[\beta(\overline{\theta}|\pi, \omega')] && \text{(A7)}\\
&= E[\beta(\overline{\theta}|\pi, \omega'') + v'|\omega'', v'] - E[\beta(\overline{\theta}|\pi, \omega') + v'|\omega', v'] && \text{(A8)}\\
&= E[y_1|\omega'', v'] - E[y_1|\omega', v'] && \text{(A9)}\\
&= CATE(y_1, v') && \text{(A10)}
\end{aligned}
$$

Note that v is independent of $\beta(\overline{\theta}|\pi, \omega)$. As a result, equality holds from (A7) to (A8) when we add v' to both expectations. □

Remark 1(c) :

Follows directly from Remark 1(a) and 1(b).

Remark 2(a) :

Proof. Recall that y_2 is given by:

$$
y_2 = \begin{cases} 1 & \text{if } \beta(\overline{\theta}|\pi_i, \omega) + v_i - \pi^C \geq 0 \\ 0 & \text{else} \end{cases} \tag{A11}
$$

$CATE(y_2, X = \pi)$ is therefore given by:

$$
\begin{aligned}
CATE(y_2, m = \pi) &= E[y_2|\pi, \omega''] - E[y_2|\pi, \omega']\\
&= \Pr[y_2 = 1|\pi, \omega''] - \Pr[y_2 = 1|\pi, \omega']\\
&= \Pr[\beta(\overline{\theta}|\omega'', \pi) + v_i - \pi^C \geq 0] - \Pr[\beta(\overline{\theta}|\omega', \pi) + v_i - \pi^C \geq 0]\\
&= \Pr[v_i \geq \gamma(\pi, \omega'')] - \Pr[v_i \geq \gamma(\pi, \omega')],
\end{aligned}
$$

where $\gamma(\pi, \omega) = \pi^c - \beta(\overline{\theta}|\omega, \pi)$. Note that $\gamma(\pi, \omega) \in [-1, 1]$. Because $v_i \sim U(-1, 1)$, and the posterior $\beta(\overline{\theta}|\omega, \pi)$ is continuous in π, $CATE(y_2, \pi) - CATE(y_2, \pi') \neq 0$ almost everywhere. □

Remark 2(b) :

Proof.

$$
\begin{aligned}
CATE(y_2, v) &= E[y_2|v, \omega''] - E[y_2|v, \omega']\\
&= \Pr[y_2 = 1|v, \omega''] - \Pr[y_2 = 1|v, \omega']\\
&= \Pr[\beta(\overline{\theta}|\omega'', \pi_i) + v - \pi^C \geq 0] - \Pr[\beta(\overline{\theta}|\omega', \pi_i) + v - \pi^C \geq 0]
\end{aligned}
$$

To calculate the above probability, the randomness comes from π_i. It is useful to rewrite $\beta(\overline{\theta}|\omega, \pi_i) + v_i - \pi^c \geq 0$ so that we can separate π_i and other non-random components:

$$
\frac{\pi_i}{1 - \pi_i} \geq \frac{\phi(g - f(\underline{\theta}, \omega))}{\phi(g - f(\overline{\theta}, \omega))} \frac{\pi^c - v}{1 - \pi^c + v} \tag{A12}
$$

Note that $\frac{\pi_i}{1-\pi_i}$ is monotone in π_i, which has distribution F_π. We use $\alpha(\omega, v_i)$ to denote the RHS of A12. We can then express $\Pr(\beta(\overline{\theta}|\omega, \pi_i) + v - \pi^C)$ as $F_\pi[\frac{\alpha(\omega,v)}{1+\alpha(\omega,v)}]$, so the CATE is given by:

$$CATE(y_2, v) = F_\pi[\frac{\alpha(\omega'', v)}{1 + \alpha(\omega'', v)}] - F_\pi[\frac{\alpha(\omega', v)}{1 + \alpha(\omega', v)}]$$

It is clear that $CATE(y_2, v)$ depends on the values of v and α. If there exists at least one $\alpha > 0$ so that $\frac{\alpha}{1+\alpha} \in (0, 1)$, then we can easily find $CATE(y_2, v) \neq CATE(y_2, v')$. A sufficient condition for $\alpha \in (0, 1)$ is $\min\{v, v'\} < \pi^C$.

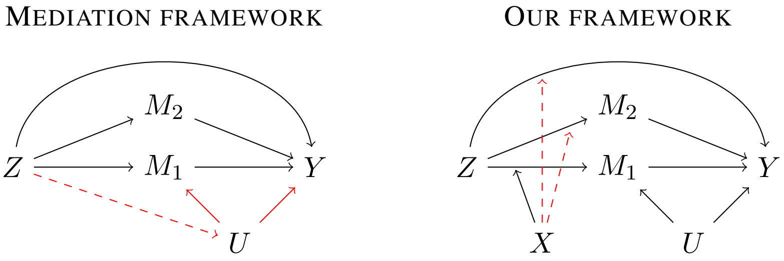
□

Remark 2(c) :

Follows directly from Remarks 2(a) and 2(b).

Appendix B Comparison to Mediation Analysis

In this section, we compare our framework connecting HTEs and mechanisms to mediation analysis. It is important to note that the two frameworks are built on different principles and objects. In mediation analysis, the main purpose is to identify and estimate the average causal mediation effects (ACMEs). Identification of these effects relies on the assumption of sequential ignorability (Imai and Yamamoto, 2013). On the other hand, our methodology aims to infer the activation of a mechanism by using heterogeneous treatment effects, which instead, relies on exclusion assumptions that we propose (Assumptions 1-2). The following DAGs facilitate our discussion of the differences in these approaches.


MEDIATION FRAMEWORK

OUR FRAMEWORK

Figure A2: DAGs when mechanisms M_1 and M_2 are independent. Note that the red arrows are ruled out by assumption of each respective framework. The left DAG, representing the assumptions of the causal mediation framework, highlights that all variables U must be in the adjustment set and also cannot be affected by the treatment Z in the mediation analysis. The right DAG, representing our framework, emphasizes covariate X cannot moderate other channels.

Consider first the case with multiple independent causal mechanisms in Figure A2. In the left DAG, treatment Z indirectly affects outcome Y through two channels M_1 and M_2, and may also directly affect Y. To nonparametrically identify average indirect effect mediated by M_1, the key part of the sequential ignorability is $Y_i(z', m_1, M_{2i}(z')) \perp\!\!\!\perp M_{1i} | Z_i = z$. The assumption is challenging to interpret and cannot be satisfied by any experimental design because it involves cross-world independence, from $z-$worlds to z'-worlds (Pearl, 2014). Graphically, all variables U should be observed and controlled by assumptions. Another important implication of the assumption is that U cannot be affected by the treatment Z, i.e., the dashed line is not allowed. When these assumptions hold, mediator M_1 must be measured in order to estimate the AMCE.

In the right DAG, treatment Z again affects outcome Y through two channels M_1 and M_2, and may also directly affect Y. Mechanism M_1 is activated if its average indirect effect is non-zero for some unit. HTEs may provide a sufficient condition for this activation. In our framework, variables U may or may not be measured or included in the adjustment set. Further, U can be a child of treatment, Z (though this introduces a third mechanism). For HTE to (ever) be informative of mechanism activation, we need to observe another pretreatment variable X. It is assumed not to moderate (average) direct effect and (average) indirect effect mediated by M_2. That is, two dashed lines are excluded in the right DAG in Figure A2. From these DAGs, it is clear that there is no logical ordering of the two types of assumptions.

There are many other differences between the two methods. For example, in the mediation analysis, mediators must be measurable and measured while these measurements are not required by our framework.

However, in our framework, researchers must have a measured candidate MIV X that is believed to satisfy Assumptions 1-2. Also, when using HTE to detect mechanisms, researchers need to pay more attention to whether Y is directly affected outcome. Even though we have emphasized their differences, two frameworks also have shared features. For example, both require that the causal effect of Z on Y is identified. This is explicitly assumed by sequential ignorability $\{Y_i(z, m_{1i}, m_{2i}), M_{1i}, M_{2i}\} \perp\!\!\!\perp Z_i$ and implicitly assumed in our framework.

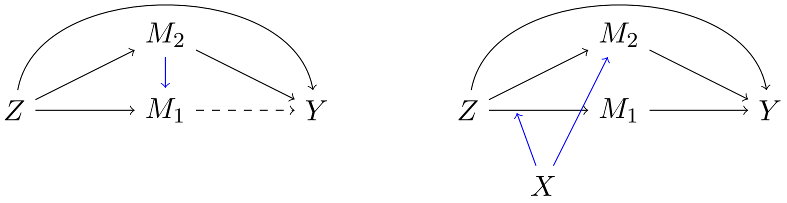


Figure A3: Two DAGs under related mechanisms (correlated mediators). The blue arrows represent correlation structures that can be accommodated using HTE analysis within our framework.

Because extension to related mechanisms (correlated mediators) is not our main focus in this study, we only make some brief comments. We consider two different correlation structures. In the left DAG of Figure A3, mediator M_2 directly affects M_1. As mentioned in the Imai and Yamamoto (2013), two assumptions are required to identify the ACME with respect to M_1. The first one is the modified sequential ignorability assumption. Unfortunately, with causally dependent multiple mediators, an assumption of no treatment-mediator interaction effects is also required. For the HTE-mechanism framework, we can simply treat correlated mechanisms as one (molar) mechanism. Then, as long as exclusion assumptions hold for the average direct effect and other indirect effects, our results in the main text still hold. One caveat is that if the M_1 mechanism does not exist, for example, the dashed line in the figure disappears, then the HTE-mechanism framework may not give the correct inference.

Multiple mechanisms can also be correlated due to other common covariates. This correlated structure can be easily accommodated to the HTE-mechanism framework. For example, in the right DAG of Figure A3, variable X affects two indirect channels with respect to M_1 and M_2. However, X does not moderate the M_2 channel and the direct channel, and thus exclusion assumptions hold. In this case, our results can be directly applied without any modification. Note that in the mediation analysis, this related structure is still classified as having independent causal mechanisms.

Appendix C Proofs of Propositions

A3.1 Proof of Proposition 1

We prove a stronger version of Proposition 1 for any non-zero $L(Y)$ where L is a non-zero affine transformation. By non-zero affine transformation, we mean that there exists a non-zero constant matrix A such that $L(Y) = AY$.

Proof. By definition of $CATE$:

$$\begin{aligned} CATE_{L(Y)}(X_j = x) &= E_{X_{\neg k}}[L(Y)|Z = z, X_k = x] - E_{X_{\neg k}}[L(Y)|Z = z', X_k = x] && \text{(A13)} \\ &= E_{X_{\neg k}}[L(DE(z, z'; X_k = x) + \sum_{j=1}^{J} IE_j(z, z'; X_k = x))] && \text{(A14)} \\ &= L\{E_{X_{\neg k}}[DE(z, z'; X_k = x) + \sum_{j=1}^{J} IE_j(z, z'; X_k = x)]\} && \text{(A15)} \end{aligned}$$

Equation (A14) follows from the linearity of expectations and the decomposition of the total effect in 11. Equation (A15) is guaranteed by the linearity of L.

Then, under exclusion assumptions 1 and 2, we can express:

$$CATE_{L(Y)}(X_j = x) - CATE_{L(Y)}(X_j = x') = L\{E_{X_{\neg k}}[IE_j(z, z'; X_k = x) - IE_j(z, z'; X_k = x']\} \tag{A16}$$

HTE exist with respect to X_k if (A16) is non-zero. In this case, then $X_k \in \mathbf{X}^{MIV}$ by the definition of MIV. □

A3.2 Proof of Propostions 2

We prove a stronger version of Proposition 2 for any non-zero $L(Y)$ where L is a non-zero affine transformation. By non-zero affine transformation, we mean that there exists a non-zero constant matrix A such that $L(Y) = AY$.

Proof. Prove by contrapositive. Suppose not, which means $\mathbf{X}^{MIV}$ is non-empty and for some $x, x' \in X_k \in \mathbf{X}^{MIV}$, $IE_j(X_k = x) \neq IE_j(X_k = x')$. Then:

$$L\{E_{X_{\neg k}}[IE_j(z, z'; X_k = x) - IE_j(z, z'; X_k = x']\} \neq 0 \tag{A17}$$

We then can reconstruct $CATE_{L(Y)}(X_j = x) - CATE_{L(Y)}(X_j = x')$ from A17:

$$L\{E_{X_{\neg k}}[IE_j(z, z'; X_k = x)]\} \neq L\{E_{X_{\neg k}}[IE_j(z, z'; X_k = x')]\} \tag{A18}$$

$$E_{X_{\neg k}}[L(DE(X_k = x) + IE_j(X_k = x) + IE_{i \neq j}(X_k = x)] \neq E_{X_{\neg k}}[L(DE(X_k = x') + \sum_{j=1}^{J} IE_j(z, z'; X_k = x)] \tag{A19}$$

$$CATE_{L(Y)}(X_j = x) \neq CATE_{L(Y)}(X_j = x') \tag{A20}$$

Equation (A19) follows because $ADE(X_k = x) = ADE(X_k = x')$ and $AIE_{i\neq j}(X_k = x) = AIE_{i\neq j}(X_k = x')$ under exclusion assumptions 1 and 2. We find HTE with respect to X_k.

So, we have shown that if two conditions do not hold, then HTE exists for some $x \neq x' \in X_k$. By contrapositive, we prove that if no HTEs exist with respect to X_k, at least one of the two conditions must be true.

□

A3.3 Proof of Proposition 3

Proof. By definition of $\mathbf{X}^R$, we can write $Y(Z, \mathbf{X}^R)$ as a function of treatment, Z, and relevant covariates $\mathbf{X}^R$. In other words, $X \notin \mathbf{X}^R$ implies that X must be independent of Y. We prove this proposition by contrapositive. Let $\mathbb{P}(\tilde{y}|Z, X_k)$ be the conditional distribution of $h(Y)$. Suppose $X_k \notin \mathbf{X}^R$, then:

$$\begin{aligned}
CATE(X_k = x) &= \int \tilde{y} d[\mathbb{P}(\tilde{y}|Z = z, X_k = x) - \mathbb{P}(\tilde{y}|Z = z', X_k = x)] && \text{(A21)}\\
&= \int \tilde{y} d[\mathbb{P}(\tilde{y}|Z = z) - \mathbb{P}(\tilde{y}|Z = z')] && \text{(A22)}\\
&= \int \tilde{y} d[\mathbb{P}(\tilde{y}|Z = z, X_k = x') - \mathbb{P}(\tilde{y}|Z = z', X_k = x')] && \text{(A23)}\\
&= CATE(X_k = x') && \text{(A24)}
\end{aligned}$$

Equations (A22) and (A23) follow from the fact that X_k is independent of Y if $X_k \notin \mathbf{X}^R$.
Therefore, equivalently, we have shown if HTEs exist with respect to X_k, then $X_k \in \mathbf{X}^R$ by contrapositive.

□

For additional intuition about how non-linear transformations of Y affect HTE, we use the following Lemma.

Lemma A1. *Suppose outcome variable $Y = g(\boldsymbol{X}^R, Z)$. Some $\{X_1, ..., X_m\} \subset \boldsymbol{X}$ are MIVs (denote them as the set $\boldsymbol{X}^{MIV}$ and the remaining as the set $\boldsymbol{X}^{non-MIV}$) if and only if there exits function $g_1(\cdot)$ and non-additively separable function $g_2(\cdot)$, and Y satisfies:*

$$Y = g_1(X^{non-MIV}, X^{MIV}) + g_2(X^{MIV}, Z) \quad \text{(A25)}$$

A function, $F(X_1, X_2)$, will be called additively separable if it can written as $f_1(X_1) + f_2(X_2)$ for some functions $f_1(X_1)$ and $f_2(X_2)$. Note further that:

1. The equation A25 in the above theorem should be understood as

$$Y = g_1(X_1, X_2, ..., X_n) + g_2(X_1, ..., X_m, Z).$$

2. The non-additively separable function $g_2(X^{MIV}, Z)$ can take the form $g_3(T) + g_4(X^{MIV}, T)$ for some function $g_3(\cdot)$ and non-additively separable function $g_4(\cdot)$.

For any non-zero linear transformation of Y, $h(Y)$, calculation of conditional expectations yields:

$$CATE(X = x) - CATE(X = x') = E[h_2(X^{MIV}, Z)|X = x] - E[h_2(X^{MIV}, Z')|X = x'] \quad \text{(A26)}$$

Equation (A26) is only a function $\mathbf{X}^{MIV}$ because $h_1(\cdot)$ cancels out.
However, for nonlinear transformed $h(Y)$, we cannot cancel $g_1(\cdot)$ in the absence of additional assumptions restricting the function form of $h(Y)$.

A3.4 Proof of Proposition 4

Proof. The result follows simply because if $X_k \notin \mathbf{X}$, then $X_k = \emptyset$. □

In the main text, Proposition 4 indicates that if there exist no HTE for the indirectly affect outcome, X_k can be any relevant or non-relevant covariate. Now we provide a stronger version of Proposition 4 by imposing assumptions about the directly-affected outcome, Y, and the form of the non-linear transformation $h(Y)$. These assumptions permit additional learning from the lack of HTE in this case.

In practice, most indirectly affected outcomes are discrete variables, such as voting behavior, survey responses, or choices. Let us consider the following non-linear transformation of the directed affected outcome Y:

$$h(Y) = \begin{cases} y_1 & Y \in (-\infty, c_1] \\ y_2 & Y \in (c_1, c_2] \\ ... & \\ y_q & Y \in (c_{q+1}, \infty) \end{cases} \tag{A27}$$

Here, will assume $y_i \in \mathbb{R}$ in (A27) has no substantive interpretation. In practice, values of y_i are typically normalizations, that are arbitrarily determined by the researcher. As such, the value is independent of model parameters.

To simplify some notation, we define:

$$p_i(x; z) \equiv \Pr[y \in (c_{i-1}, c_i] | X = x, Z = z] \tag{A28}$$

$$p_i(x; z, z') \equiv Pr[y \in (c_{i-1}, c_i] | X = x, Z = z] - \Pr[y \in (c_{i-1}, c_i] | X = x, Z = z']. \tag{A29}$$

In order to calculate CATEs, we need at least two possible values of the treatment Z and two distinct values of the covariate X_k. We define a covariate as *effective* as follows:

Definition A1. *$X_k \in \boldsymbol{X}$ is effective if $\exists i \in \{1, 2, ..., q\}$ and $x, x' \in X_k$ such that $p_i(x; z, z') \neq p_i(x'; z, z')$.*

Effectiveness means that as X_i changes, it can induce a different probability of $h(Y) = y_i$. It should be clear that if X_k is effective, then it must the case that $X_k \in \mathbf{X}^R$. In general, if X_k is not effective, then $X_k \notin \mathbf{X}^R$.

Proposition A1. *Suppose that observed outcome $h(Y)$ is a discrete non-linear mapping of directly-affected outcome Y in equation A27 and Assumptions 1 and 2 hold. Assume further that Y has an absolutely continuous distribution. If HTEs do not exist with respect to X_k, then X_k is almost surely not effective.*

Proof. Given $x, x' \in X$, CATEs are given by:

$$CATE(X_i = x) = \sum_{i=1}^{q} y_i [p_i(x; z) - p_i(x; z')] = \sum_{i=1}^{q} y_i p_i(x; z, z') \tag{A30}$$

$$CATE(X_i = x') = \sum_{i=1}^{q} y_i [p_i(x'; z) - p_i(x'; z')] = \sum_{i=1}^{q} y_i p_i(x'; z, z'). \tag{A31}$$

We now will prove the proposition by contrapositive. Suppose that X_k is effective. If so, then there exists an index set, D, with at least two elements such that $CATE(x) - CATE(x') = \sum_{i \in D} y_j[p_i(x; z, z') - p_i(x'; z, z')]$ and any $p_i(x; z, z') = 0$ for all $i \notin D$. Because y_i is arbitrarily set and is independent of p_i, and Y has absolutely continuous distribution, the probability that $\sum_{j \in D} y_j[p_j(x; z, z') - p_j(x'; z, z')] = 0$ is zero.

□

We use the following example to illustrate the above proposition.

Example A1. *Suppose $h(Y)$ has the following form:*

$$h(Y) = \begin{cases} y_1 & Y \in (-\infty, c_1] \\ y_2 & Y \in (c_1, \infty) \end{cases}$$

where $Y = h(X_1, X_2, Z)$.

Then, let us calculate the CATE X_2 , given $z, z' \in Z$:

$$\begin{aligned} CATE(X_2 = x) &= y_1[p_1(x; z) - p_1(x; z')] + y_2[p_2(x; z) - p_2(x; z')] \\ &= y_1 p_1(x; z, z') + y_2 p_2(x; z, z') \end{aligned}$$

and

$$\begin{aligned} CATE(X_2 = x') &= y_1[p_1(x'; z) - p_1(x'; z')] + y_2[p_2(x'; z) - p_2(x'; z')] \\ &= y_1 p_1(x'; z, z') + y_2 p_2(x'; z, z') \end{aligned}$$

If X_k is not effective, then $CATE(X_2 = x) = CATE(X_2 = x')$, therefore there exist no HTE. If X_k is effective, then non-existance of HTE requires that

$$y_1 p_1(x; t, t') + y_2 p_2(x; z, z') = y_1 p_1(x'; z, z') + y_2 p_2(x'; z, z')$$

$$\frac{y_1}{y_2} = \frac{p_2(x'; z, z') - p_2(x; z, z')}{p_1(x; z, z') - p_1(x'; z, z')} \tag{A32}$$

For arbitrarily chosen $y_1 \in \mathbb{R}$ and $y_2 \in \mathbb{R}$, the above equality holds with probability zero if p_1 or p_2 can take value in a set with Lebesgue measure larger than 0.

Appendix D Discrete Outcomes under Monotonicity Assumptions

In this section, we illustrates when and how we use HTE to learn about mechanisms when outcomes are indirectly affectd by mechanisms. To be specific, consider two DGPs in figure A4. We will index these DGPs by $s \in \{1, 2\}$ where $s = 1$ corresponds to the left DAG and $s = 2$ corresponds to the right DAG.

$$\begin{tikzcd} & X^{MIV} \arrow[d] & & & \\ Z \arrow[r, dash] & {} \arrow[r] & M_1 \arrow[r] & Y_1 \arrow[r] & h(Y_1) \end{tikzcd} \qquad \begin{tikzcd} & X^{non-MIV} \arrow[d] & & \\ Z \arrow[r] & M_1 \arrow[r] & Y_2 \arrow[r] & h(Y_2) \end{tikzcd}$$

Figure A4: Two different DGPs. On the left, in DGP 1, X is a MIV. On the right, in DGP 2, X is not a MIV.

The left panel assumes X is a MIV. X is not a MIV in the right panel. In the figure, there are no other mediators. Therefore, both graphs satisfy exclusion assumptions 1 and 2 by construction.

We will assume that Y_s is a latent directly-affected outcome and $h(Y_s)$ is the observed binary variable:

$$h(Y_s) = \begin{cases} 0 & Y_s \in (-\infty, c] \\ 1 & Y_s \in (c, \infty], \end{cases} \tag{A33}$$

for some $c \in (-\infty, \infty)$. Propositions 3 and 4 show that we cannot differentiate between the left and right on the basis of the existance or non-existance of HTE for indirectly-affected outcome $h(Y_s)$.

We will consider what can be gained by imposing a monotonicity assumption of the form: $\frac{\partial^2 Y}{\partial Z \partial X} > (<) 0$ (note that the inequalities are strict).[1] Clearly montonicity can hold in the left panel (where X is a MIV) but $\frac{\partial^2 Y}{\partial Z \partial X} = 0$ in the right panel in which X is not a MIV. To explore the implications of monotonicity, we consider following DGPs for $i = \{1, 2\}$:

$$Y_i = g_i(Z, X) + e \tag{A34}$$

For $X \in X^{MIV}$, e.g., the left DGP in Figure A4, suppose that montonicity holds such that $\frac{\partial^2 Y_1}{\partial X \partial Z} := \beta(x, z)$, where $\beta(x, z)$ is either strictly positive or negative. For $X \notin X^{MIV}$, e.g., the right DGP in Figure A4, by definition, we have $\frac{\partial^2 Y_2}{\partial x \partial z} = 0$.

We ask whether researchers can differentiate these two cases when motonicity holds for the first DGP (e.g., an assumption of monotonicity). First, given (A33) and (A34), note that:

$$E[h(Y_s)|Z, X] = \Pr(e \geq c - g_s(Z, X)) \tag{A35}$$

Let f_e be the densify of e denote its derivative by f_e'. We can then express HTE for $h(Y_1)$ as:

$$-f_e'(c - g_1)\frac{\partial g_1}{\partial X}\frac{\partial g_1}{\partial Z} + f_e(c - g_1)\beta(x, z) \tag{A36}$$

[1] Writing the monotonicity assumption in this way assumues that this derivative exists.

and the HTE for $h(Y_2)$ is

$$-f_e'(c-g_2)\frac{\partial g_2}{\partial X}\frac{\partial g_2}{\partial Z} \tag{A37}$$

The additional term $-f_e(c-g_1)\beta(x,z)$ in equation A36 may help us to differentiate two DGPs by generating a differently-signed HTE. For example, if under certain x and z, (A37) and the first term of (A36) have the same sign and the second term of (A36) has the opposite sign and is sufficiently large, (A37) and (A36) will have different signs.

A second possibility to identify two DGPs comes from the $f_e'(\cdot)$. If e is uniformly distributed, then $f_e' = 0$ and thus equation A37 is equal to 0 while equation A36 is non-zero.

We summarize the discussion in the following proposition.

Proposition A2. *Consider the indirectly affected outcome $h(Y_s)$ satisfying equation A33 in which moderation effect $\beta(x,z)$ is monotonic.*

(1) Suppose that e *is uniformly distributed, then HTE for $h(Y_2)$ is 0.*

(2) Suppose e *is not uniformly distributed, then HTE for $h(Y_1)$ and $h(Y_2)$ have different signs under two cases:*

(2a) $f_e'(c-g_2)\frac{\partial g}{\partial X}\frac{\partial g}{\partial Z} < 0$ *and* $\beta(x,z) < \frac{f_e'(c-g_1)\frac{\partial g_1}{\partial X}\frac{\partial g_1}{\partial Z}}{f_e(c-g_1)}$*; or*

(2b) $f_e'(c-g_2)\frac{\partial g}{\partial X}\frac{\partial g}{\partial Z} > 0$ *and* $\beta(x,z) > \frac{f_e'(c-g_1)\frac{\partial g_1}{\partial X}\frac{\partial g_1}{\partial Z}}{f_e(c-g_1)}$

In practice, however, it is difficult to verify conditions in (2). Corollary A1 provides additional assumptions on $g(\cdot)$ and/or the tail behavior of e that are sufficient to satisfy these conditions.

Corollary A1. *Suppose conditions in proposition A2 holds. Assume that:*

(a) g is increasing in X and Z,

(b) the distribution of e *is unimodal,*

(c) β is increasing in X and Z,

then

(1) small values of X and Z satisfy condition (2a), if any such x, z exists;

(2) large values of X and Z satisfy condition (2b), if any such x, z exists.

Proof. It is straightforward to prove the corollary. If we pick small values of x and z in the data, then by condition (1) g is small and $\frac{\partial g}{\partial X}\frac{\partial g}{\partial Z} > 0$, and by (2) $f_e'(c-g) < 0$, by (3) β is small enough as well. These together imply 2(a) in proposition A2 is satisfied. The same logic holds is for (2). □

Appendix E Simulation

Illustration: The distinction between directly-affected and indirectly-affected outcomes is novel to this paper. To illustrate the logic and implications of learning about mechanisms from HTE in the case of an indirectly-affected outcome, we provide a short simulation that incorporates real attitudinal data. Specifically, we consider a hypothetical persuasion experiment that aims to shift support for greenhouse gas regulation among partisans in the US. Consistent with approaches used by scholars of persuasion, we will examine heterogeneity in partisan affiliation (here, simplified to Democrats and Republicans) (see Coppock, 2022, etc.). We use data on (1) partisan affiliation; (2) support for greenhouse gas regulation, coded as a binary outcome where 1 designates support for increased regulation; and (3) demographic covariates from the 2020 American National Election Study. It is useful to note that partisans' opinions are relatively polarized on this issue: while 82.2% (95% CI: [80.5%, 84.0%]) of Democrats favor increasing regulations, just 38.3% (95% CI: [35.7%, 40.9%]) of Republicans favor such regulations. This suggests that partisanship is strongly prognostic of support for greenhouse gas regulation.

Various theories of learning or additudinal change incorporate mechanisms that imply that partisans may react differently to information about greenhouse gas regulation. Little, Schnakenberg, and Turner (2022) classify two mechanisms for belief formation and attitude change: accuracy and directional motives. Within their model, ideology (partisanship) is posited as a moderator of directional motives but not accuracy motives, meaning that partisanship is a candidate MIV for directional motives. To this end, we simulate different processes of attitudinal change to examine when we observe HTE in partisanship. Our simulation proceeds as follows:

1. Estimate latent untreated potential outcomes ($Y_i(0)$) from observed data, using gender, education, ideology and partisanship.
2. Simulate a (latent) treatment effect of the form:

$$Y_i(1) = Y_i(0) + \tau \mathbb{I}(\text{Partisanship}_i = P)$$

   We consider three different indicator functions for partisanship. $P \in \{$Democrat, Republican, Democrat $\cup$ Republican$\}$. The latter case includes the full sample since the sample is conditioned on either of the two parties.
3. Randomly assign treatment, $Z \in \{0, 1\}$ to half of the sample to reveal (latent) potential outcomes $Y_i(Z)$.
4. Reveal observed potential outcomes $L(Y_i) = Bernoulli(logit^{-1}(Y_i(Z)))$.
5. Estimate $CATE(P = Democrat) - CATE(P = Republican)$ for the binary outcome $L(Y_i)$.

We vary $\tau \in [-1.5, 1.5]$, which are treatment effects on a logistic scale.[2] Figure A5 reports the results of our simulation. In the left panel, we see that for non-zero treatment effects (e.g., for any $\tau \neq 0$), we

[2] The assumption of a constant τ is clearly a simplification for the purposes of illustration. It does not generally follow from the Little, Schnakenberg, and Turner (2022) model.

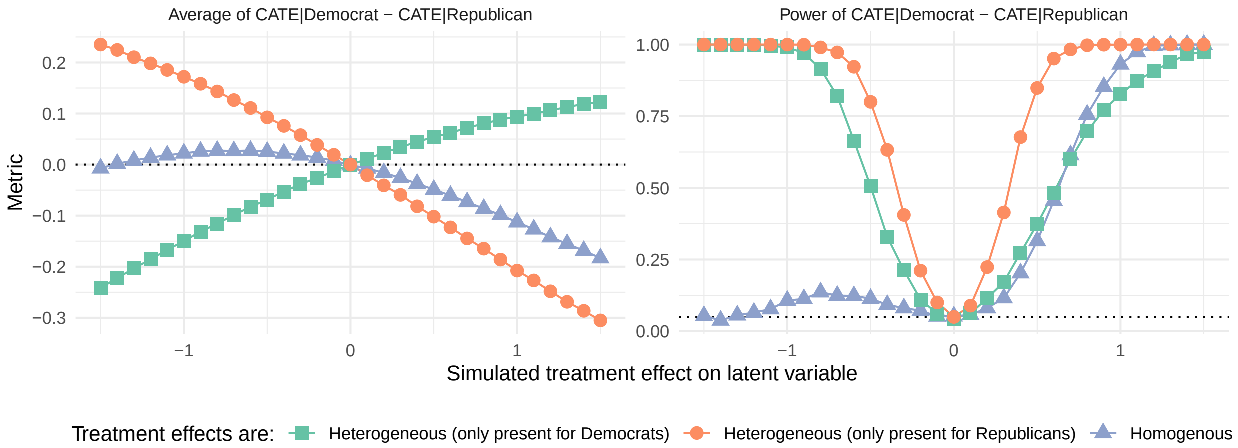


Figure A5: Note that $N = 2883$ partisans. We assess power at the $\alpha = 0.05$ level.

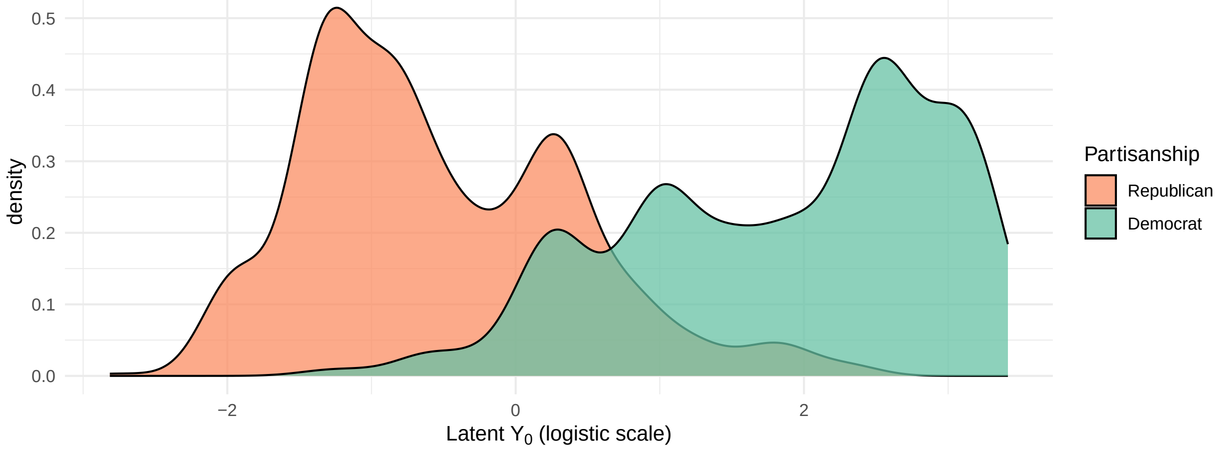


Figure A6: Distribution of latent $Y_i(0)$'s, by party.

always observe HTE in partisanship, even when effects on the latent scale are *homogeous*, e.g., the degree of atitudinal change is not moderated by partisanship. We observe different treatment effects for Democrats and Republicans on the binary outcome even with homogenous treatment effects on the latent attitude because of different densities of respondents about the relevant cutpoint in the latent variable (see Figure A6).

Supplementary Appendix: References

Coppock, Alexander. 2022. *Persuasion in Parallel*. Chicago, IL: University of Chicago Press.

Imai, Kosuke, and Teppei Yamamoto. 2013. “Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments.” *Political Analysis* 21 (2): 141–171.

Little, Andrew T., Keith E. Schnakenberg, and Ian R. Turner. 2022. “Motivated Reasoning and Democratic Accountability.” *American Political Science Review* 116 (2): 751–767.

Pearl, Judea. 2014. “Interpretation and identification of causal mediation.” *Psychological methods* 19 (4): 459.