# Tracing Causal Paths from Experimental and Observational Data[*]

Xiang Zhou Teppei Yamamoto

Harvard University MIT

July 17, 2019

[Preliminary draft; Comments are welcome]

## Abstract

Despite a growing interest in the study of causal mechanisms in political science, existing methods for causal mediation analysis face an important limitation when the effect of the treatment on the outcome involves multiple, potentially overlapping, causal pathways. To circumvent this limitation, empirical studies often assume that different pathways are causally independent, an assumption that is strong, untestable, and unrealistic in many applications. In this article, we relax this assumption using a framework for tracing causal paths in the presence of multiple causally dependent mediators. In this framework, the total effect of the treatment on the outcome is decomposed into a set of path-specific effects, which are identified under standard assumptions of causal mediation analysis. We then describe an imputation approach for estimating path-specific causal effects from experimental and observational data. In contrast to existing methods for analyzing causal mediation, this approach does not require any model for the mediators of interest. All we need is to model the expected outcome given treatment, pretreatment confounders, and varying sets of mediators, which can be implemented via highly nonparametric methods. We illustrate this approach by estimating the path-specific effects of democracy on public opposition to war and the path-specific effects of political violence on descendants' political attitudes.

[*] Direct all correspondence to Xiang Zhou, Department of Government, Harvard University, 1737 Cambridge Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu. We thank Michael Tomz, Jessica Weeks, Noam Lupu, and Leonid Peisakhin for making available the replication data of their studies.

# 1 Introduction

Political scientists, no longer content with establishing the mere presence of causal effects, are increasingly interested in uncovering the pathways through which one variable affects another. For example, Brader, Valentino and Suhay (2008) examine whether the effect of negative media framing on public opposition to immigration is mediated by respondent anxiety rather than beliefs about the harms of immigration. Similarly, in studying the democratic peace, Tomz and Weeks (2013) investigate whether the effect of democracy on public opposition to war is mediated by moral qualms over military strikes against another democracy rather than beliefs about the costs and benefits of war. Over the past decade, studies of causal mediation have grown rapidly across different subfields of political science (e.g., Abramson and Carter 2016; Hall 2017; Holbein 2017; Knutsen et al. 2017; Lupu and Peisakhin 2017; Reese, Ruby and Pape 2017; Zhu 2017) because empirical evaluation of the mechanisms hypothesized to transmit causal effects is central for testing and refining theoretical models of social and political processes (Imai et al. 2011; Acharya, Blackwell and Sen 2016).

A common approach to assessing causal mediation involves decomposing the total effect of a treatment on an outcome into two components: an indirect effect operating through a mediator of interest and a direct effect operating through alternative pathways. This is typically accomplished via an additive decomposition in which the average total effect of treatment is partitioned into so-called natural direct and indirect effects (Pearl 2001; Robins 2003; VanderWeele 2015), which are also known as the average direct effect (ADE) and average causal mediation effect (ACME), respectively (Imai et al. 2010, 2011).

Despite its conceptual simplicity, this approach faces an important limitation when the effect of the treatment on the outcome involves multiple, potentially overlapping, causal pathways — a common scenario in social science applications. In particular, the ADE and ACME can only be identified under a set of potentially strong assumptions: (i) no unobserved treatment-outcome confounding, (ii) no unobserved treatment-mediator confounding, (iii) no unobserved mediator-outcome confounding, and (iv) no treatment-induced mediator-outcome confounding (VanderWeele 2009*a*; Imai et al. 2010). Of these assumptions, the last assumption is especially restrictive because it requires that there must not be any posttreatment variables that affect both the mediator and outcome, whether they are observed or not.

Therefore, if two mediators are present and one mediator affects both the other mediator and the outcome, the ACME for the second mediator cannot be nonparametrically identified (Imai and Yamamoto 2013). For example, when assessing the degree to which respondent anxiety mediates the effect of negative media framing on attitudes toward immigration, the ACME for anxiety is not identified if beliefs about the harms of immigration influence both respondent anxiety and attitudes toward immigration. Likewise, when assessing the degree to which moral concerns transmit the effect of democracy on attitudes toward war, the ACME for morality is not identified if beliefs about the costs and benefits of war affect both perceptions of morality and attitudes toward war. To circumvent this problem, empirical studies have often assumed that different mediators are causally independent (i.e., they do not affect each other), whether implicitly or explicitly. Unfortunately, this assumption is strong, untestable, and unrealistic in many applications.

Moreover, from a substantive point of view, when the effect of the treatment on the outcome involves multiple mediators that are causally dependent, the mediating effects of these variables cannot be neatly separated from each other. In the democratic peace study, for example, the ACME for perceived costs and benefits of war involves the causal path *democracy→perceived costs and benefits→opposition to war*, and the ACME for morality involves the causal path *democracy→moral concerns→opposition to war*. However, if moral concerns about war are partly influenced (perhaps subconsciously) by perceived costs and benefits of war, both of the ACMEs will also involve the causal path *democracy→perceived costs and benefits→moral concerns→opposition to war*. Thus, in such cases, the ACMEs for different mediators may reflect overlapping causal mechanisms, making it difficult to adjudicate between competing theories of the underlying processes.

In sum, when the effect of the treatment on the outcome involves multiple mediators that are causally dependent, an exclusive attention to the ACMEs for different mediators could be unproductive both methodologically and substantively. In this paper, we argue that in the presence of multiple causally dependent mediators, a more fruitful approach to the study of causal mechanisms is to trace the various causal paths directly. In the democratic peace study, for example, the researcher might be interested in the strength of the causal path *democracy→moral concerns→opposition to war*, i.e., the amount of treatment effect operating via moral concerns above and beyond that operating via other mediators. Unlike the ACME for morality, this *path-specific effect* is identified even if perceptions about the costs and benefits of war influence both morality and attitudes toward war. This quantity is also substantively important because it gauges the degree to which morality plays an independent role in transmitting the effect of democracy on public opposition to war.

In general, when multiple causally dependent mediators lie on the causal paths from a treatment to an outcome, the total effect of the treatment can be decomposed into a set of path-specific effects. These path-specific effects are nonparametrically identified as long as all observed variables can be arranged in a directed acyclic graph (DAG) and, in this DAG, no unobserved confounding exists for any of the causal relationships (Avin, Shpitser and Pearl 2005) — a standard assumption in empirical applications of causal mediation analysis. Despite this general identification result, however, few practical methods have been proposed to implement this decomposition and estimate the corresponding path-specific effects. The lack of off-the-shelf estimation methods has prevented empirical analyses of causal mediation from tracing causal paths directly in the presence of multiple mediators. Instead, as noted earlier, researchers have often resorted to the strong and restrictive assumption that different mediators do not affect each other, which enables identification of the ACMEs but would lead to inaccurate assessments of causal mechanisms if the mediators are in fact causally dependent.

This article bridges this gap. In what follows, we first introduce a framework for defining and identifying path-specific causal effects in the presence of multiple causally dependent mediators. We then briefly review an inverse-probability-weighting-based method for estimating path-specific causal effects proposed by VanderWeele, Vansteelandt and Robins (2014), which, as we show, has several practical limitations that make it challenging to implement beyond highly stylized applications. We then describe a new imputation-based approach for estimating path-specific causal effects that can be flexibly applied to both experimental and observational data. In contrast to existing methods for analyzing causal mediation (e.g. Imai et al. 2011; VanderWeele 2015), this approach does not require any model for the mediators of interest. All we need is to model the expected outcome given treatment, pretreatment confounders, and varying sets of mediators, which can be implemented via any method of the analyst's choice, be it linear models, generalized additive models, or highly nonparametric methods such as Bayesian Additive Regression Trees (BART; Chipman, George and McCulloch 2010; Hill 2011). We illustrate this approach by estimating the path-specific effects of shared democracy on public opposition to war and the path-specific effects of political violence on descendants' political attitudes.

# 2 Path-Specific Causal Effects: A Review

## 2.1 Definition of Path-Specific Effects

We use $A$ to denote a binary treatment, $Y$ to denote the observed outcome, and $X$ to denote a vector of observed pretreatment confounders. To simplify the exposition, let us first consider the case where two (sets of) mediators, $L$ and $M$, lie on the causal paths from $A$ to $Y$. Without loss of generality, we assume that $L$ precedes $M$, such that no variable in $M$ can causally affect any variable in $L$.[1] A causal diagram that is consistent with the relationships between these variables is shown in the top panel of Figure 1. In Tomz and Week's (2013) study on the democratic peace, for example, $A$ denotes whether the potential opponent is a democracy, $Y$ denotes the respondent's attitude toward war, $L$ denotes the respondent's beliefs about the costs and benefits of war, and $M$ denotes the respondent's perceived morality of war.

In this causal diagram, there exist four possible paths from the treatment to the outcome, as shown in the panels below: (a) $A \to Y$; (b) $A \to M \to Y$; (c) $A \to L \to Y$; and (d) $A \to L \to M \to Y$. If the mediators $L$ and $M$ are causally independent, i.e., if they do not affect each other, the last path does not exist. In this case, the total effect of $A$ on $Y$ can be partitioned into the effect operating through $L$ ($A \to L \to Y$), the effect operating through $M$ ($A \to M \to Y$), and a "direct" effect not operating through $L$ or $M$ $(A \to Y)$ (Imai and Yamamoto, 2013). However, in the general case where $L$ and $M$ are causally dependent, it is not possible to partition the mediating effects of $L$ and $M$ into their respective components even conceptually, since some of the total effect of $A$ on $Y$ goes through both $L$ and $M$.

To define the path-specific effects formally, let us use the potential outcomes notation for both the outcome and the mediators. Specifically, we use $Y(a, l, m)$ to denote the potential outcome under treatment status $a$ and mediator values $L = l$ and $M = m$, $M(a, l)$ to denote the potential value of the mediator $M$ under treatment status $a$ and mediator value $L = l$, and $L(a)$ to denote the potential value of the mediator $L$ under treatment status $a$.[2] This notation allows us to define nested coun-

[1] This is not a binding assumption as long as the full causal structure can be represented by a DAG, in which case the mediators can always be partitioned into two ordered sets.

[2] We omit the usual unit index (often denoted by a subscript $i$) from our notation for the sake of notational brevity, with the implicit assumption that the potential outcomes and mediators refer to unit-level counterfactuals.

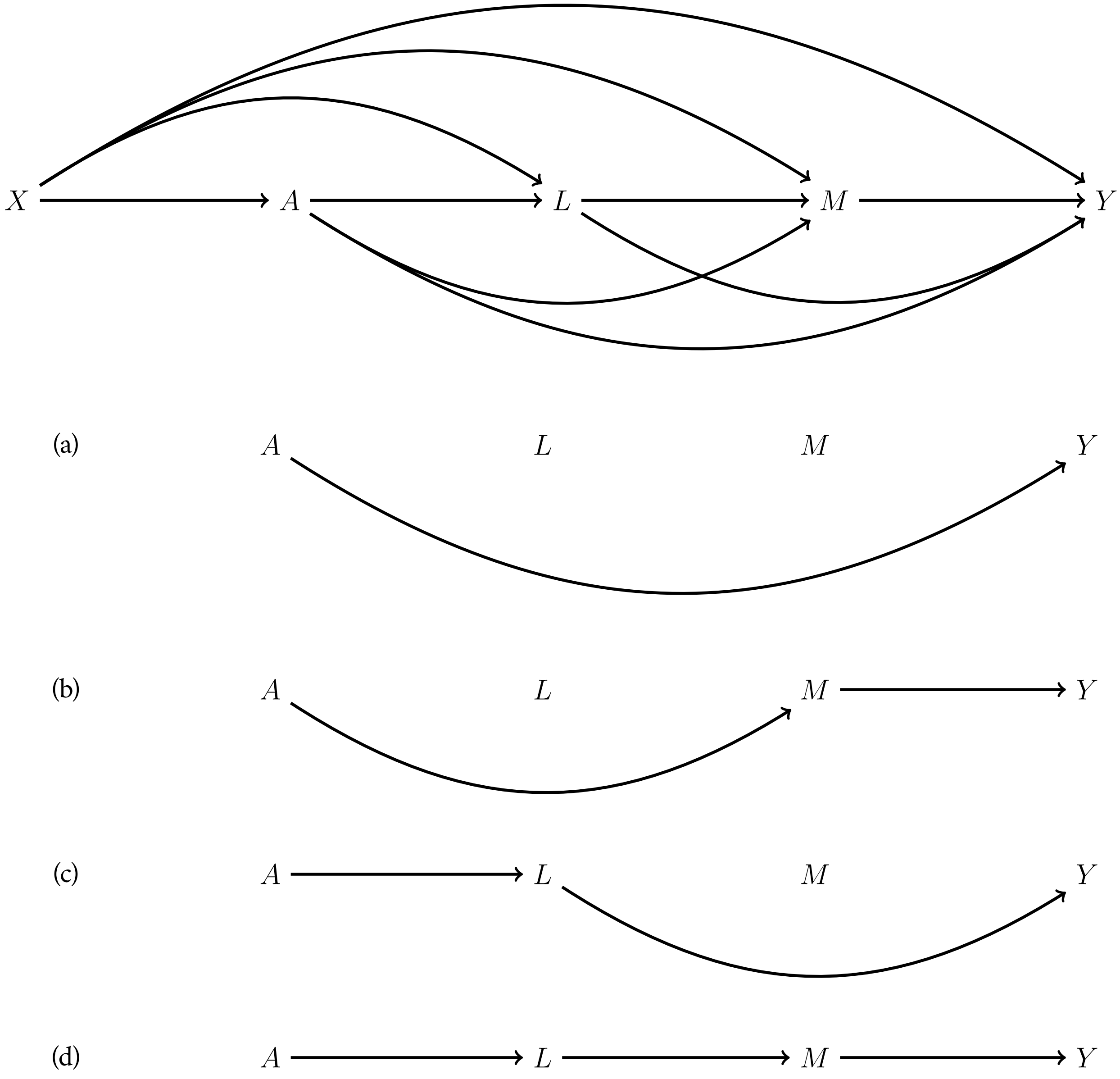

Figure 1: Causal Relationships with Two (Sets of) Causally Dependent Mediators Shown in Directed Acyclic Graph (DAG).

Note: $A$ denotes the treatment, $Y$ denotes the outcome, $X$ denotes pretreatment confounders, and $L$ and $M$ denote two (sets of) causally dependent mediators.

terfactuals.[3] For example, $Y\big(1, L(0), M(0, L(0))\big)$ represents the potential outcome in the counterfactual scenario where the subject was treated but the mediators $L$ and $M$ were set to values they would haven taken if the subject had not been treated. Further, if we let $Y(a)$ denote the potential outcome when treatment status is set to $a$ and the mediators $L$ and $M$ take on their "natural" values under treatment status $a$ (i.e., $L(a)$ and $M(a, L(a))$), we have $Y(a) = Y\big(a, L(a), M(a, L(a))\big)$ by definition.

Using the above notation, the average total effect of $A$ on $Y$ can be written as a telescoping sum (VanderWeele, Vansteelandt and Robins 2014):

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[Y\big(1, L(1), M(1, L(1))\big) - Y\big(0, L(0), M(0, L(0))\big)] \\
&= \underbrace{\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big) - Y\big(0, L(0), M(0, L(0))]}_{A\to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big) - Y\big(1, L(0), M(0, L(0))\big)]}_{A\to M\to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(1, L(1), M(1, L(1))\big) - Y\big(1, L(0), M(1, L(0))\big)]}_{A\to L\to Y; A\to L\to M\to Y} \\
&\equiv \tau_{A\to Y}(0) + \tau_{A\to M\to Y}(1) + \tau_{A\to L\rightsquigarrow Y}(1).
\end{aligned} \tag{1}$$

In this decomposition, $\tau_{A\to Y}(0)$ (line 2) represents the amount of treatment effect if the mediators $L$ and $M$ were set to values they would haven taken under treatment status $A = 0$. It reflects the causal path $A \to Y$. $\tau_{A\to M\ \to Y}(1)$ (line 3) represents the amount of treatment effect operating through the mediator $M$ under treatment status $A = 1$ and mediator status $L = L(0)$. It reflects the causal path $A \to M \to Y$. Finally, $\tau_{A\to L\rightsquigarrow Y}(1)$ (line 4) represents the amount of treatment effect operating through the mediator $L$ under treatment status $A = 1$. It reflects the combination of the causal paths $A \to L \to Y$ and $A \to L \to M \to Y$.

Thus, in the democratic peace example, $\tau_{A\to Y}(0)$ reflects the direct effect of democracy on public opposition to war, i.e., the fraction of the total effect operating through neither perceived costs and benefits of war nor perceived morality of war; $\tau_{A\to M\ \to Y}(1)$ reflects the effect of democracy operating through morality only; and $\tau_{A\to L\rightsquigarrow Y}(1)$ reflects the effect of democracy operating through perceived costs and benefits of war, whether it further operates through morality or not.

[3] We maintain the stable unit treatment value assumption, or consistency, for all of the potential outcomes and mediators, such that $L = L(a)$ if $A = a$, etc.

The decomposition represented by equation (1) is not unique. Switching the order in which the causal paths $A \to Y$, $A \to M \to Y$, and $A \to L \rightsquigarrow Y$ are introduced, the total effect of $A$ on $Y$ can also be partitioned as

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y\big(1, L(1), M(1, L(1))\big) - Y\big(0, L(1), M(1, L(1))]}_{A \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(0, L(1), M(1, L(1))\big) - Y\big(0, L(1), M(0, L(1))\big)]}_{A \to M \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(0, L(1), M(0, L(1))\big) - Y\big(0, L(0), M(0, L(0))\big)]}_{A \to L \to Y; A \to L \to M \to Y} \\
&\equiv \tau_{A \to Y}(1) + \tau_{A \to M \to Y}(0) + \tau_{A \to L \rightsquigarrow Y}(0). \qquad (2)
\end{aligned}$$

In what follows, we refer to equations (1) and (2) as Type I and Type II decompositions, respectively. In general, when there is an interaction effect between the treatment and the mediators on the outcome, the path-specific effects defined by the two decompositions are different. We illustrate both Type I and Type II decompositions in Section 4.

## 2.2 Comparison to Existing Approaches

Existing work on causal mediation analysis for multiple mediators typically focuses on the ACME for each of the mediators, instead of path-specific effects. For example, Imai and Yamamoto (2013) consider the following decomposition of the average total effect:

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y(1, L(1), M(0, L(0)))] - \mathbb{E}[Y(0, L(0), M(0, L(0)))]}_{A \to Y; A \to L \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y(1, L(1), M(1, L(1)))] - \mathbb{E}[Y(1, L(1), M(0, L(0)))]}_{A \to M \to Y; A \to L \to M \to Y} \\
&\equiv \zeta_M(0) + \delta_M(1). \qquad (3)
\end{aligned}$$

In this decomposition, $\delta_M(1)$ represents the ACME of $A$ on $Y$ with respect to $M$, or the amount of treatment effect operating through $M$ under treatment status $A = 1$, regardless of whether the effect also operates through $L$ or not. By contrast, $\zeta_M(0)$ represents the amount of treatment effect that does not operate through $M$, regardless of $L$. It can also be shown that $\mathbb{E}[Y(1) - Y(0)] = \zeta_M(1) + \delta_M(0)$, where $\zeta_M(1)$ and $\delta_M(0)$ are defined analogously.

This decomposition is attractive when the researcher's substantive interest lies solely in the mediator $M$ and the other mediator $L$ is purely a nuisance that needs to be accounted for due to the confounding it causes between $M$ and $Y$. A drawback of this approach, however, is that neither $\zeta_M(1)$ nor $\delta_M(0)$ can be nonparametrically identified under standard ignorability assumptions, such as the assumptions encoded in the causal diagram in the top panel of Figure 1.[4] Moreover, empirical researchers are often in a situation where both $L$ and $M$ are of substantive interest.

In contrast, our proposed approach can be understood with respect to the following alternative representation:

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big) - Y\big(0, L(0), M(0, L(0))]}_{A \to Y; A \to M \to Y} \\ &\quad + \underbrace{\mathbb{E}[Y\big(1, L(1), M(1, L(1))\big) - Y\big(1, L(0), M(1, L(0))\big)]}_{A \to L \to Y; A \to L \to M \to Y} \\ &\equiv \zeta_L(0) + \delta_L(1), \qquad (4)\end{aligned}$$

where $\delta_L(1)$ and $\zeta_L(0)$ represent the ACME of $A$ on $Y$ with respect to $L$ under treatment status $A = 1$ and the ADE of $A$ on $Y$ with respect to $L$ under treatment status $A = 0$, respectively. Equation (4) makes it clear that $\delta_L(1) = \tau_{A \to L \rightsquigarrow Y}(1)$ and $\zeta_L(0) = \tau_{A \to Y}(0) + \tau_{A \to M \to Y}(1)$. Thus, the proposed approach allows us to estimate the amount of treatment effect that operates through $L$ under a standard set of ignorability assumptions, and also to further decompose the ADE ($\zeta_L(0)$) into the effect operating through $M$ but not through $L$ ($\tau_{A \to M \to Y}(1)$) and the effect that operates through neither $L$ nor $M$ ($\tau_{A \to Y}(0)$).

## 2.3 Identification Results

Following Pearl (2009), we use a DAG to denote a nonparametric structural equation model with independent errors. In this framework, the top panel of Figure 1 corresponds to the following non-

[4] Imai and Yamamoto (2013) circumvent this problem by adding weak functional form assumptions — specifically, by modeling the outcome and mediators semiparametrically and making a no-interaction assumption.

parametric structural equations:

$$
\begin{aligned}
A &= f_A(X, \epsilon_A) \\
L &= f_L(X, A, \epsilon_L) \\
M &= f_M(X, A, L, \epsilon_M) \\
Y &= f_Y(X, A, L, M, \epsilon_Y),
\end{aligned}
$$

where the error terms $\epsilon_A$, $\epsilon_L$, $\epsilon_M$, and $\epsilon_Y$ are jointly independent but otherwise arbitrarily distributed. The joint independence of the error terms means that no unobserved confounding exists for any of the treatment-mediator, treatment-outcome, mediator-mediator, and mediator-outcome relationships represented in Figure 1. This assumption implies (but is not implied by) the sequential ignorability assumption that Robins (2003) invoked in interpreting causal diagrams.[5]

As mentioned in the introduction, in the presence of posttreatment confounding of the mediator-outcome relationship, the ACME and ADE are generally not nonparametrically identified. This is true even if all posttreatment confounders of the mediator-outcome relationship are observed. Thus, if we think of the mediator $L$ as a posttreatment confounder of the relationship between $M$ and $Y$, it means that the ACME for the mediator $M$ is not identified. However, under the assumptions outlined above, the path-specific effects defined by equations (1) and (2) are nonparametrically identified (Avin, Shpitser and Pearl 2005; VanderWeele, Vansteelandt and Robins 2014). In fact, to identify the components of equations (1) and (2), it suffices to identify the counterfactual expectation $\mathbb{E}[Y\big(a, L(a^*), M(a^{**}, L(a^*))\big)]$ for any combination of $a$, $a^*$, $a^{**}$. This latter quantity can be expressed as a function of observed variables:

$$\mathbb{E}[Y\big(a, L(a^*), M(a^{**}, L(a^*))\big)] = \int_{x,l,m} \mathbb{E}[Y|x, a, l, m] f(m|x, a^{**}, l) f(l|x, a^*) f(x) dx dl dm. \quad (5)$$

A proof of this result is given in Appendix A. This equation can be seen as a generalization of Pearl's (2001) mediation formula to the case of two causally dependent (sets of) mediators. Note that the last components in equations (1) and (2), i.e., $\tau_{A \to L \rightsquigarrow Y}(0)$ and $\tau_{A \to L \rightsquigarrow Y}(1)$, reflect the combination of the

[5] In particular, our assumption implies the independence between the so-called cross-world counterfactuals, whereas the sequential ignorability assumption of Robins (2003) does not. See Robins and Richardson (2010) for a detailed discussion for different interpretations of causal diagrams.

causal paths $A \to L \to Y$ and $A \to L \to M \to Y$. Without additional assumptions, the path-specific effects for $A \to L \to Y$ and $A \to L \to M \to Y$ cannot be separately identified.[6] Thus, in the democratic peace example, we can identify the mediating effect (e.g., the ACME) of perceived costs and benefits of war, but we cannot know how much of this mediating effect further operates through morality.

## 2.4 VanderWeele et al.'s (2014) Weighting Estimator

To estimate the path-specific effects defined above, VanderWeele, Vansteelandt and Robins (2014) proposed a weighting estimator that entails estimating the conditional densities/probabilities of the mediators $L$ and $M$ given their antecedent variables. To see how it works, we note that equation (5) can be rewritten as

$$\begin{aligned}
&\mathbb{E}[Y(a, L(a^*), M(a^{**}, L(a^*)))] \\
=&\int_{x,l,m} \mathbb{E}[Y|x,a,l,m] f(m|x,a^{**},l) f(l|x,a^*) f(x) dx dl dm \\
=&\int_{x,l,m} \mathbb{E}[Y|x,a,l,m] f(x,l,m|a) \frac{f(m|x,a^{**},l) f(l|x,a^*) \mathbb{P}(A=a)}{f(m|x,a,l) f(l|x,a) \mathbb{P}(A=a|x)} dx dl dm \\
=&\mathbb{E}[Y \frac{f(M|X,A=a^{**},L) f(L|X,A=a^*) f(a)}{f(M|X,A=a,L) f(L|X,A=a) f(a|X)} |A=a]. \qquad (6)
\end{aligned}$$

Thus, the counterfactual expectation $\mathbb{E}[Y(a, L(a^*), M(a^{**}, L(a^*))]$ is simply a weighted average of the observed outcome among units with treatment status $a$. With estimates of $f(m|x,a,l)$, $f(l|x,a)$, $f(a|x)$, and $f(a)$, the weights can be constructed as

$$\hat{W}_{a,a^*,a^{**}} = \frac{\hat{f}(M|X,A=a^{**},L) \hat{f}(L|X,A=a^*) \hat{f}(a)}{\hat{f}(M|X,A=a,L) \hat{f}(L|X,A=a) \hat{f}(a|X)}. \qquad (7)$$

This weighting approach, while conceptually appealing, has several practical limitations. First, because it depends on estimates of the conditional density/probability functions $f(m|x,a,l)$ and $f(l|x,a)$, it performs well only when both the mediators $L$ and $M$ are discrete, in which case the conditional probabilities can often be reliably estimated. When either or both of the mediators is

[6] For the additional assumptions needed to identify the paths $A \to L \to Y$ and $A \to L \to M \to Y$ separately, see Albert and Nelson (2011).

multi-dimensional or continuous, estimates of these conditional densities/probabilities tend to be unstable and highly sensitive to model misspecification (e.g., Naimi et al. 2014; Vansteelandt 2009). Moreover, even if the models for these conditional densities/probabilities are correctly specified, inverse-probability-weighted estimators such as the one given by equation (6) are generally inefficient and susceptible to large finite sample biases (Cole and Hernán 2008; Wang et al. 2006). We now turn to an imputation approach that circumvents these limitations.

# 3 Estimating Path-Specific Effects: An Imputation Approach

## 3.1 Rationale

To date, most methods for causal mediation analysis have focused on the setting where the researcher is interested in a single mediator or a single set of mediators. In this case, the key quantity for identifying the ACME and ADE is the nested counterfactual, $\mathbb{E}[Y(a, M(a^*)]$, where $M$ is the sole mediator of interest. Different methods have been proposed to estimate this quantity (e.g., Imai et al. 2011; VanderWeele 2009*b*). In particular, Vansteelandt, Bekaert and Lange (2012) introduced an imputation method, which involves (a) fitting a model of the observed outcome conditional on treatment, the mediator, and a set of pretreatment confounders, (b) using this model to impute the counterfactual outcomes $Y(a, M(a^*))$ for each unit with treatment status $a^*$, and (c) fitting a model of these imputed counterfactuals conditional on the pretreatment confounders (see also Steen et al. 2017). Albert (2012) proposed a similar method, in which the first two steps are exactly the same and the last step involves an inverse-probability-of-treatment-weighted average of the imputed counterfactuals. Below, we extend these imputation-based methods to the estimation of path-specific effects.

Without loss of generality, let us consider the Type I decomposition defined by equation (1).[7] To estimate the three components in equation (1), it suffices to estimate four counterfactual means: $\mathbb{E}[Y(0)]$, $\mathbb{E}[Y(1)]$, $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$, and $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$. Given the assumption of no unobserved confounding for the treatment-outcome relationship, the first two quantities, $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$, can be estimated via any conventional method of covariate adjustment, such as matching, weighting, or regression. Or, in experimental studies where treatment is randomly as-

[7] Our exposition applies analogously to the Type II decomposition defined by equation (2).

signed, they can be estimated using simple averages of the observed outcome within the control and treatment groups.

Using the mediation formula (5), the latter two quantities, $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$ and $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$, can be written as

$$\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L, M]|A = 0, X\big]\Big] \tag{8}$$

$$\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L]|A = 0, X\big]\Big]. \tag{9}$$

A proof of these equations is given in Appendix B. Thus, to evaluate these nested counterfactuals, we need only to estimate (a) the conditional means $\mathbb{E}[Y|X, A = 1, L, M]$ and $\mathbb{E}[Y|X, A = 1, L]$, and (b) their own conditional means given the pretreatment confounders $X$ among the untreated units ($A = 0$). After these estimates are obtained, the outermost expectations in equations (8) and (9) can be estimated using their sample analogs.

Alternatively, the nested counterfactuals above can be written as (see also Appendix B)

$$\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)] = \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L, M]\frac{\mathbb{P}[A = 0]}{\mathbb{P}[A = 0|X]}|A = 0\big] \tag{10}$$

$$\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)] = \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L]\frac{\mathbb{P}[A = 0]}{\mathbb{P}[A = 0|X]}|A = 0\big]. \tag{11}$$

These equations suggest that to evaluate the nested counterfactuals, we need only to estimate $\mathbb{E}[Y|X, A = 1, L, M]$, $\mathbb{E}[Y|X, A = 1, L]$, and the probability ratio $\mathbb{P}[A = 0]/\mathbb{P}[A = 0|X]$. After these estimates are obtained, the outer expectation in equations (10) and (11) can be estimated using their sample analogs.

Hence, the equations (8-9) and (10-11) point to two different routes to evaluating the nested counterfactuals $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$ and $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$. They can be seen as extensions of Vansteelandt et al.'s (2012) and Albert's (2012) imputation-based estimators for the ACME, respectively, to the estimation of path-specific effects. Since the first procedure involves only model-based imputation and the second procedure involves both imputation and inverse-probability-weighting, we call them a "pure imputation estimator" and an "imputation-based weighting estimator," respectively. Unlike VanderWeele et al.'s (2014) weighting approach, neither of these estimators requires estimating the conditional densities/probabilities of the mediators. They are therefore highly amenable to the setting where the mediators $L$ and $M$ are multivariate and/or continuous. Below, we provide a practical guide on the implementation of these estimators in experimental and observational studies.

## 3.2 Implementation

First, consider the experimental setting where treatment status is randomly assigned. In this case, because treatment status $A$ is independent of the pretreatment confounders $X$, both the equations (8-9) and the equations (10-11) reduce to

$$\begin{aligned}\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)] &= \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L, M]|A = 0\big] \\ \mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)] &= \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L]|A = 0\big].\end{aligned}$$

Thus, in experimental studies, the imputation approach can be implemented as follows:

1. Estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ using simple averages of the observed outcome within the control and treatment groups.

2. Fit an outcome model conditional on the treatment $A$, the mediators $L$ and $M$, and the pretreatment confounders $X$. For the control units, impute their counterfactual outcome $Y\big(1, L(0), M(0, L(0))\big)$ using their predicted outcomes at $A = 1$ and their observed values of $X$, $L$, and $M$. The average of these imputed counterfactuals constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$.

3. Fit an outcome model conditional on the treatment $A$, the mediator $L$, and the pretreatment confounders $X$. For the control units, impute their counterfactual outcome $Y\big(1, L(0), M(1, L(0))\big)$ using their predicted outcomes at $A = 1$ and their observed values of $X$ and $L$. The average of these imputed counterfactuals constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$.

4. Calculate the path-specific effects as defined in equation (1).

In practice, to reduce model dependence, highly nonparametric methods such as Gradient Boosting Machines (GBM) or Bayesian Additive Regression Trees (BART) can be used to fit the outcome models in steps 2 and 3. This can be useful to reduce bias due to model misspecification, especially when treatment-mediator interaction effects are likely to exist (Glynn 2012). Standard errors and confidence intervals can be estimated by bootstrapping steps 1-4. In Section 5.1, we illustrate this approach by tracing the causal paths through which shared democracy reduces public support for war.

In observational studies, the pure imputation estimator (equations 8-9) and the imputation-based weighting estimator (equations 10-11) do not coincide. The pure imputation estimator can be implemented as follows:

1. Fit an outcome model conditional on the treatment $A$ and the pretreatment confounders $X$. Estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ by averaging the predicted values $\hat{\mathbb{E}}[Y|A=0,X]$ and $\hat{\mathbb{E}}[Y|A=1,X]$ among all units, respectively.

2. Fit an outcome model conditional on the treatment $A$, the mediators $L$ and $M$, and the pretreatment confounders $X$. For the untreated units, impute their counterfactual outcome $Y\big(1, L(0), M(0, L(0))\big)$ using their predicted outcomes at $A = 1$ and their observed values of $X$, $L$, and $M$.

3. Fit a model of the imputed counterfactual $\hat{Y}\big(1, L(0), M(0, L(0))\big)$ conditional on $X$ among untreated units, and obtain model-based predictions for all units. The average of these predictions constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$.

4. Fit an outcome model conditional on the treatment $A$, the mediators $L$, and the pretreatment confounders $X$. For the untreated units, impute their counterfactual outcome $Y\big(1, L(0), M(1, L(0))\big)$ using their predicted outcomes at $A = 1$ and their observed values of $X$ and $L$.

5. Fit a model of the imputed counterfactual $\hat{Y}\big(1, L(0), M(1, L(0))\big)$ conditional on $X$ among untreated units, and obtain model-based predictions for all units. The average of these predictions constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$.

6. Calculate the path-specific effects as defined in equation (1).

For the imputation-based weighting estimator, steps 3 and 5 are replaced by an inverse-probability weighted average:

1. Fit an outcome model conditional on the treatment $A$ and the pretreatment confounders $X$. Estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ by averaging the predicted values $\mathbb{E}[Y|A=0,X]$ and $\mathbb{E}[Y|A=1,X]$ among all units, respectively. In the meantime, estimate $\mathbb{P}[A=0]$ using its sample analog and $\mathbb{P}[A=0|X]$ using a propensity score model for the treatment.

2. Fit an outcome model conditional on the treatment $A$, the mediators $L$ and $M$, and the pretreatment confounders $X$. For the untreated units, impute their counterfactual outcome $Y\big(1, L(0), M(0, L(0))\big)$ using their predicted outcomes at $A=1$ and their observed values of $X$, $L$, and $M$.

3. Estimate $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$ using a weighted average of the imputed counterfactuals $\hat{Y}\big(1, L(0), M(0, L(0))\big)$ among the untreated units, where the weight is $\hat{\mathbb{P}}[A=0]/\hat{\mathbb{P}}[A=0|X]$.

4. Fit an outcome model conditional on the treatment $A$, the mediators $L$, and the pretreatment confounders $X$. For the untreated units, impute their counterfactual outcome $Y\big(1, L(0), M(1, L(0))\big)$ using their predicted outcomes at $A=1$ and their observed values of $X$ and $L$.

5. Estimate $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$ using a weighted average of the imputed counterfactuals $\hat{Y}\big(1, L(0), M(1, L(0))\big)$ among the untreated units, where the weight is $\hat{\mathbb{P}}[A=0]/\hat{\mathbb{P}}[A=0|X]$.

6. Calculate the path-specific effects as defined in equation (1).

Again, to reduce model dependence, highly nonparametric methods can be used to fit both the outcome models and the propensity score model for treatment assignment (for the imputation-based weighting estimator). Standard errors and confidence intervals can be estimated by bootstrapping steps 1-6. In Section 5.2, we illustrate both the pure imputation estimator and the imputation-based weighting estimator by tracing the intergenerational pathways through which exposure to political violence reduces descendants' regime support.

## 4 Generalization to $K(>2)$ Ordered (Sets of) Mediators

Throughout the previous discussion, we have assumed that two (sets of) mediators, $L$ and $M$, lie on the causal paths from $A$ to $Y$. The definition, identification, and estimation of path-specific effects can be generalized to the case where the treatment effect operates through $K$ causally dependent (sets of) mediators. In what follows, we denote these mediators as $M_1, M_2, \dots M_K$ and assume that for any $i < j$, $M_i$ precedes $M_j$, such that no variable in $M_j$ can causally affect any variable in $M_i$. In addition, let us denote $\mathcal{M}_0 = \varnothing$, $\mathcal{M}_k = \{M_1, M_2, \dots M_k\}$ and $\mathcal{M}_k(a) = \{M_1(a), M_2(a), \dots M_k(a)\}$, where $M_k(a) = M_k\big(a, M_1(a), M_2(a, M_1(a)), \dots\big)$ by consistency.

The total effect of $A$ on $Y$ can now be decomposed as

$$\begin{aligned}\mathbb{E}[Y(1)-Y(0)] &= \underbrace{\mathbb{E}[Y\big(1,\mathcal{M}_K(0)\big) - Y(0)]}_{A\to Y} + \sum_{k=1}^{K} \underbrace{\mathbb{E}[Y\big(1,\mathcal{M}_{k-1}(0)\big) - Y\big(1,\mathcal{M}_k(0)\big)]}_{A\to M_k\rightsquigarrow Y} \\ &= \tau_{A\,\to\,Y}(0) + \sum_{k=1}^{K} \tau_{A\,\to\,M_k\,\rightsquigarrow\,Y}(1) \qquad (12)\end{aligned}$$

or

$$\begin{aligned}\mathbb{E}[Y(1)-Y(0)] &= \underbrace{\mathbb{E}[Y\big(1\big) - Y\big(0,\mathcal{M}_k(1)\big)]}_{A\to Y} + \sum_{k=1}^{K} \underbrace{\mathbb{E}[Y\big(0,\mathcal{M}_k(1)\big) - Y\big(0,\mathcal{M}_{k-1}(0)\big)]}_{A\to M_k\rightsquigarrow Y} \\ &= \tau_{A\,\to\,Y}(1) + \sum_{k=1}^{K} \tau_{A\,\to\,M_k\,\rightsquigarrow\,Y}(0) \qquad (13)\end{aligned}$$

Following the earlier terminology, we refer to equations (12) and (13) as Type I and Type II decompositions, respectively.

We assume that the variables $A, M_1, \dots M_K, Y$ follow a DAG that encodes a nonparametric structural equation model with independent errors:

$$\begin{aligned} A &= f_A(X, \epsilon_A) \\ M_k &= f_{M_k}(X, A, M_1, \dots M_{k-1}, \epsilon_{M_k}), \quad 1 \le k \le K \\ Y &= f_Y(X, A, M_1, \dots M_K, \epsilon_Y), \end{aligned}$$

where the error terms $\epsilon_A$, $\epsilon_{M_k}$, and $\epsilon_Y$ are jointly independent but otherwise arbitrarily distributed. The joint independence of the error terms means that no unobserved confounding exists for any of the treatment-mediator, treatment-outcome, mediator-mediator, and mediator-outcome relationships.

To identify the components of equations (12) and (13), it suffices to identify the counterfactual expectation $\mathbb{E}[Y\big(a, \mathcal{M}_k(a^*)\big)]$ for any $k$ and any combination of $a$, $a^*$. This counterfactual expectation can be expressed as a function of observed variables:

$$\mathbb{E}[Y\big(a, \mathcal{M}_k(a^*)\big)] = \int \mathbb{E}[Y|x, a, m_1, \ldots m_k] f(m_1, \ldots m_k|x, a^*) f(x) dx da dm_1 \ldots dm_k$$

This equation is simply Pearl's (2001) mediation formula applied to the set of mediators $\mathcal{M}_k = \{M_1, M_2 \ldots M_k\}$ only.

The imputation approach outlined above can be easily generalized to estimate the path-specific effects defined in equations (12) and (13). Without loss of generality, let us consider equation (12), i.e., the Type I decomposition. The pure imputation estimator proceeds as follows:

1. Fit an outcome model conditional on the treatment $A$ and the pretreatment confounders $X$. Estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ by averaging the predicted values $\hat{\mathbb{E}}[Y|A = 0, X]$ and $\hat{\mathbb{E}}[Y|A = 1, X]$ among all units, respectively.

2. For $k = 1, 2, \ldots K$,

    (a) Fit an outcome model conditional on the treatment $A$, the mediators $\mathcal{M}_k$, and the pretreatment confounders $X$. For the control units, impute their counterfactual outcome $Y\big(1, \mathcal{M}_k(0)\big)$ using their predicted outcomes at $A = 1$ and their observed values of $X$ and $\mathcal{M}_k$.

    (b) Fit a model of the imputed counterfactual $\hat{Y}\big(1, \mathcal{M}_k(0)\big)$ conditional on $X$ among untreated units, and obtain model-based predictions for all units. The average of these predictions constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, \mathcal{M}_k(0)\big)]$.

3. Calculate the path-specific effects as defined in equation (12).

For the imputation-based weighting estimator, step 2(b) is replaced by an inverse-probability-weighted average:

1. Fit an outcome model conditional on the treatment $A$ and the pretreatment confounders $X$. Estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ by averaging the predicted values $\hat{\mathbb{E}}[Y|A=0,X]$ and $\hat{\mathbb{E}}[Y|A=1,X]$ among all units, respectively. In the meantime, estimate $\mathbb{P}[A=0]$ using its sample analog and $\mathbb{P}[A=0|X]$ using a propensity score model for the treatment.

2. For $k=1,2,\dots K$,

   (a) Fit an outcome model conditional on the treatment $A$, the mediators $\mathcal{M}_k$, and the pretreatment confounders $X$. For the control units, impute their counterfactual outcome $Y\big(1,\mathcal{M}_k(0)\big)$ using their predicted outcomes at $A=1$ and their observed values of $X$ and $\mathcal{M}_k$.

   (b) Estimate $\mathbb{E}[Y\big(1,\mathcal{M}_k(0)\big)]$ using a weighted average of the imputed counterfactuals $\hat{Y}\big(1,\mathcal{M}_k(0)\big)$ among the untreated units, where the weight is $\hat{\mathbb{P}}[A=0]/\hat{\mathbb{P}}[A=0|X]$.

3. Calculate the path-specific effects as defined in equation (12).

In experimental studies, step (1) is simplified as $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ can be estimated using simple averages of the observed outcome within the control and treatment groups. In the meantime, step 2(b) is unneeded, as $\mathbb{E}[Y\big(1,\mathcal{M}_k(0)\big)]$ can be estimated using a simple average of the imputed counterfactuals among the control units.

# 5 Empirical Illustrations

## 5.1 Democracy and Public Opposition to War

With a nationally representative sample of 1,273 US adults, Tomz and Weeks (2013) conducted a survey experiment to analyze the role of public opinion in the democratic peace, i.e., the empirical regularity that democracies almost never fight each other. In this experiment, they presented respondents with a situation in which a country was developing nuclear weapons and, when describing the situation, they randomly and independently varied three characteristics of the country: political regime (whether it was a democracy), alliance status (whether it had signed a military alliance with the United States), and economic ties (whether it had high levels of trade with the United States).

They then asked respondents about their levels of support for a preventive military strike against the country's nuclear facilities. The authors found that individuals are substantially less supportive of military action against democracies than against otherwise identical autocracies.

To investigate the causal mechanisms through which democracy reduces public support for war, Tomz and Weeks (2013) also measured each respondent's beliefs about the threat posed by the potential adversary (*threat*), the cost of military intervention (*cost*), and the likelihood of victory (*success*). In addition, the authors also assessed each respondent's moral concerns about using military force (*morality*). With these data, they conducted a causal mediation analysis and found that democracy reduces public support for war primarily by changing perceptions of the threat and morality of using military force. In this analysis, the authors examined the role of each mediator separately by assuming that they operate independently and do not influence one another. However, it is likely that one's perception of morality is partly influenced by beliefs about the threat, cost, and likelihood of success, which also affect support for war directly. In the following analysis, we allow these mediators to be causally dependent. In particular, we group *threat*, *cost*, and *success* together as mediators reflecting the respondent's beliefs about *the costs and benefits of war*, and treat them as causally prior to *morality*. Under this latter assumption, the causal mechanisms underlying the effect of democracy can be represented as a DAG akin to the top panel of Figure 1.

In this DAG, the outcome, $Y$, denotes whether the respondent opposes a preventive military strike; treatment, $A$, denotes whether the country developing nuclear weapons is presented as a democracy; the mediators $L$ include measures of the respondent's beliefs about the costs and benefits of war; the mediator $M$ is a dummy variable indicating whether the respondent thought it would be morally wrong to strike; finally, the pretreatment covariates $X$ include dummy variables for each of the two other randomized treatments (alliance status and economic ties) as well as a number of demographic and attitudinal controls. We control for a set of pretreatment covariates because, although treatment is randomly assigned, the mediator-mediator and mediator-outcome relationships may still be confounded by baseline factors in these data.

We estimate the path-specific effects for $A \to Y$, $A \to M \to Y$, and $A \to L \rightsquigarrow Y$ using the imputation approach outlined in Section 3.2. Because treatment is randomly assigned in this study, we first estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ using simple averages of the observed outcome within the control and treatment groups. We find that the proportion of respondents opposing war is 27.6%

when the country developing nuclear weapons is an autocracy and 38.8% when it is a democracy. The total effect of treatment, therefore, is 11.2%.

We then use BART to fit an outcome model conditional on democracy ($A$), perceived costs and benefits of war ($L$), perceived morality of war ($M$), and the pretreatment covariates ($X$), and, for the control units, extract their predicted outcomes at $A = 1$ and their observed values of $X$, $L$, and $M$. The average of these predicted outcomes constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$, which reflects the proportion of respondents opposing war if the country developing nuclear weapons was presented as a democracy but respondents' beliefs about the costs and benefits of war and about the morality of war were both set to the levels they would have taken if the country developing nuclear weapons had been presented as an autocracy. Thus, its deviation from the average baseline outcome $\mathbb{E}[Y(0)]$ reflects the direct effect of democracy on public opposition to war.

Next, we use BART to fit an outcome model conditional on democracy ($A$), perceived costs and benefits of war ($L$), and the pretreatment covariates ($X$), and, for the control units, extract their predicted outcomes at $A = 1$ and their observed values of $X$ and $L$. The average of these predicted outcomes constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$, which reflects the proportion of respondents opposing war if the country developing nuclear weapons was presented as a democracy but respondents' beliefs about the costs and benefits of war were set to the levels they would have taken if the country developing nuclear weapons had been presented as an autocracy. Thus, its deviation from the counterfactual mean $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$ gauges the path-specific effect for $A \to M \to Y$, and its difference from the average treated outcome $\mathbb{E}[Y(1)]$ gauges the path-specific effect for $A \to L \rightsquigarrow Y$.

By switching the roles of the treatment and control groups, we have also estimated the path-specific effects defined by the Type II decomposition (equation 2). The results are shown in Figure 2. We can see that in this example, different definitions of path-specific effects yield similar results. By both Type I and Type II decompositions, about three fifths of the total effect operates through either perceived costs and benefits of war or perceived (im)morality of war. Among the three fifths, about two fifths stem from the mediating effect of perceived costs and benefits of war ($A \to L \rightsquigarrow Y$), and the remaining one fifth can be attributed to morality alone ($A \to M \to Y$). Thus, echoing Tomz and Weeks (2013), we find that morality plays a small but independent role in transmitting the causal

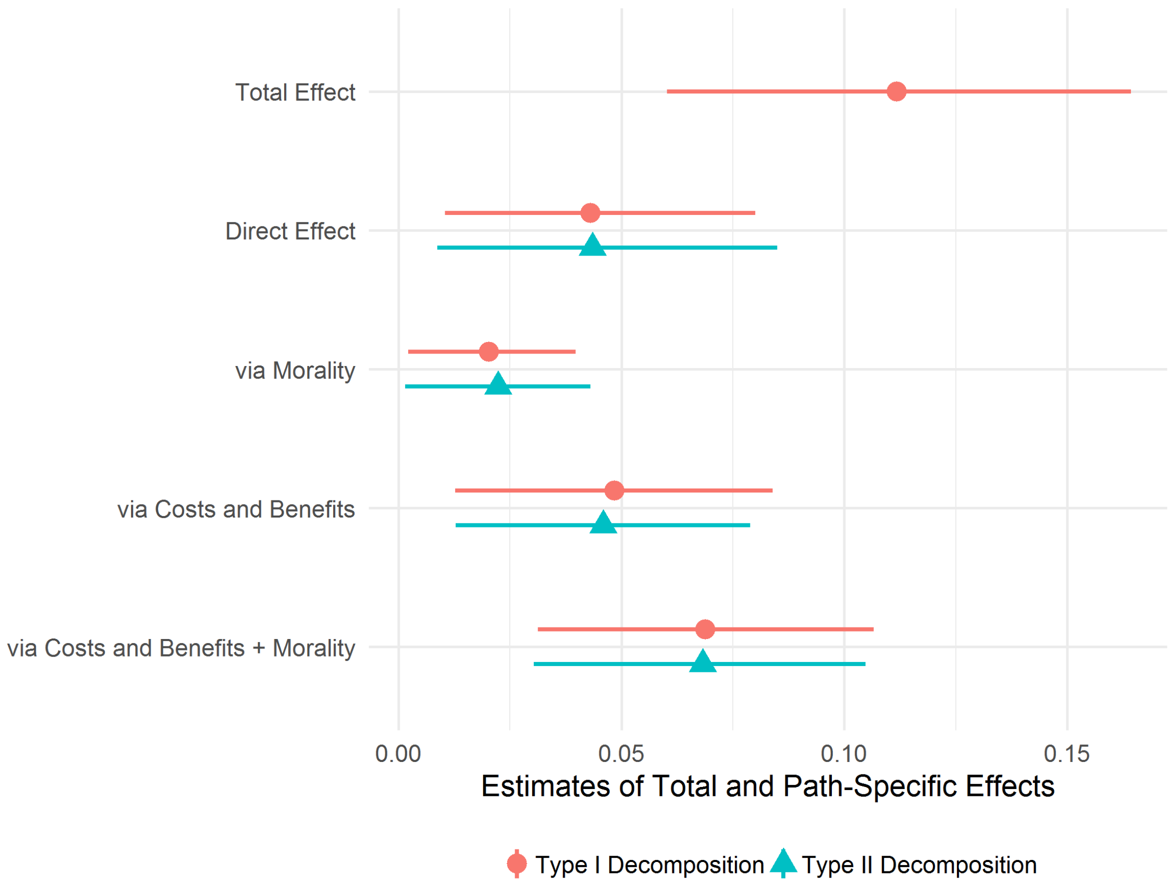

Figure 2: Estimates of Total and Path-Specific Effects of Shared Democracy on Public Opposition to War.

Note: Error ranges correspond to 95% bootstrapped confidence intervals (500 iterations).

effect of democracy on attitudes toward war. Yet, unlike the original study, the above results do not hinge on the restrictive assumption that different mediators are causally independent.

## 5.2 The Legacy of Political Violence

In this subsection, we illustrate the imputation approach for tracing causal paths from observational data. In particular, we reanalyze Lupu and Peisakhin's (2017) data to examine the intergenerational pathways through which exposure to political violence shapes descendants' political attitudes. In 2014, Lupu and Peisakhin conducted a multigenerational survey of Crimean Tatars, a minority Muslim population living in Crimea, to study the legacy of political violence that occurred during the deportation of Crimean Tatars from their homeland to Central Asia in 1944. Due to starvation and

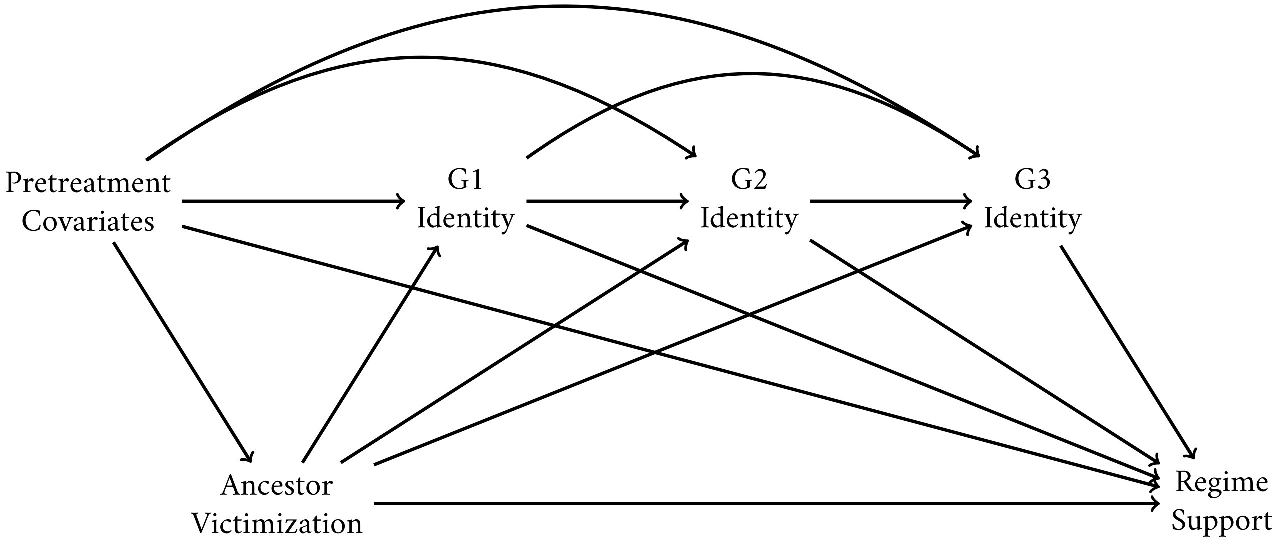

Figure 3: Causal Pathways from Ancestor Victimization to Descendants' Regime Support.

infectious diseases, a sizable portion of the deportees died during or shortly after the deportation. Yet, "[a]lthough all Crimean Tatars suffered the violence of deportation, some lost more family members along the way" (p. 837). Leveraging this variation in violent victimization, the authors found that the grandchildren of individuals who suffered more deaths of family members support more strongly the Crimean Tatar political leadership, hold more hostile attitudes toward Russia, and participate more in politics.

To investigate the causal mechanisms that transmit the legacy of political violence, Lupu and Peisakhin (2017) conducted an "implicit mediation analysis" by adding measures of the descendant's political identity into their main regression models and assessing the changes in the coefficients of ancestor victimization. This approach can be problematic, however, because descendants' political identities are likely shaped by the political identities of their parents and grandparents, which might have a direct effect on descendant political attitudes and behavior. In other words, the identities of first- and second-generation respondents may be posttreatment confounders of the mediator-outcome relationship, i.e., the relationship between descendants' identities and their political attitudes and behavior. A mediation analysis omitting these posttreatment confounders would likely lead to biased assessments of causal mechanisms. Moreover, as noted earlier, the ACME and ADE cannot be nonparametrically identified in the presence of posttreatment confounders.

Here, we treat the political identities of first-, second-, and third-generation respondents as three sets of causally dependent mediators, and focus on the effect of ancestor victimization on the respondent's attitude toward Russia's annexation of Crimea. Our analytical framework can be represented by the DAG in Figure 3. In this DAG, ancestor victimization (i.e., the treatment) denotes whether any family member of the first-generation respondent died during or shortly after the deportation due to poor conditions; the political identities of first-, second-, and third-generation respondents (i.e., the mediators) are measured by the intensity of their attachment to the Crimean Tatars as a social group, their association of that group with victimhood, and their perception of the threat posed by Russia; regime support (i.e., the outcome) denotes whether the third-generation respondent supported Russia's annexation of Crimea; finally, the pretreatment covariates include measures of the first generation respondent's family wealth, religiosity, attitudes toward the Soviet Union, and experience with persecution by state authorities prior to deportation. These covariates are used to control for potential confounding of the treatment-mediator, treatment-outcome, mediator-mediator, and mediator-outcome relationships.

We then estimate the path-specific effects as defined by equations (12) and (13), using both the pure imputation estimator and the imputation-based weighting estimator. For the pure imputation estimator, we use BART to estimate all outcome models (including the models for the imputed counterfactuals). For the imputation-based weighting estimator, we estimate all outcome models using BART and estimate the propensity score model using boosted regression trees that are calibrated to maximize covariate balance (McCaffrey, Ridgeway and Morral 2004; Ridgeway et al. 2017). The results, as shown in Figure 4, are similar between the two estimators and between the two decompositions. Consistent with the original study, we find that ancestor victimization significantly reduces the descendant's support for Russia's annexation of Crimea — by 0.2 on the linear probability scale (from 0.54 to 0.34). The "direct effect" is about -0.09, meaning that slightly more than half of the total effect operates through the political identities of first-, second-, and third-generation respondents. Yet, most of the indirect effect is transmitted through the political identities of grandparents and parents, rather than directly through the political identities of third-generation respondents. This finding supports Lupu and Peisakhin's hypothesis that exposure to political violence affects the identities of first-generation respondents and that they transmit these through the family line to shape the political attitudes of their descendants.

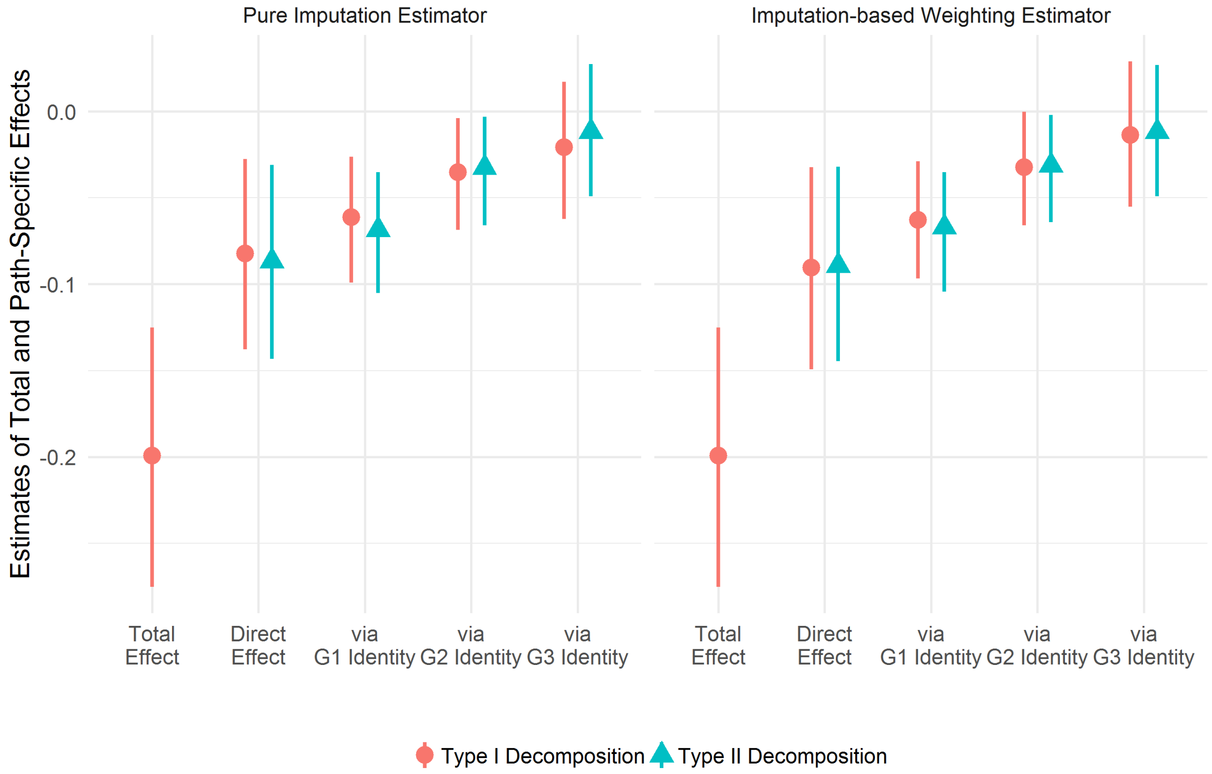

Figure 4: Estimates of Total and Path-Specific Effects of Ancestor Victimization on Support for Russia's Annexation of Crimea.
Note: Error ranges correspond to 95% bootstrapped confidence intervals (500 iterations).

## 6 Concluding Remarks

Despite a growing interest in the study of causal mechanisms in political science, existing methods for causal mediation analysis face an important limitation when the effect of the treatment on the outcome involves multiple, potentially overlapping, causal pathways. In particular, the average causal mediation effect (ACME) cannot be nonparametrically identified if the mediator-outcome relationship is confounded by other variables that are causally prior to the mediator of interest, even if these variables are observed. To circumvent this problem, empirical studies have often assumed that different mediators are causally independent, an assumption that enables identification of the ACMEs but would lead to biased assessments of causal mechanisms if the mediators are in fact causally dependent.

To confront this problem without invoking strong and restrictive assumptions, methodologists have proposed several alternative estimands, such as the controlled direct effect (Pearl 2001; Robins 2003; VanderWeele and Vansteelandt 2009; Acharya, Blackwell and Sen 2016; Zhou and Wodtke 2019) and the randomized analogs to the natural direct and indirect effects (Geneletti 2007; VanderWeele, Vansteelandt and Robins 2014; Wodtke and Zhou 2019), which can still be identified in the presence of posttreatment confounding of the mediator-outcome relationship. Yet, none of these alternative approaches allows us to directly gauge the strengths of different causal paths from the treatment to the outcome, a task essential to the evaluation of competing theories of social and political processes.

To bridge this gap, we introduced a framework for tracing causal paths in the presence of multiple causally dependent mediators. In this framework, the total effect of the treatment on the outcome is decomposed into a set of path-specific effects. These path-specific effects, unlike the ACMEs of individual mediators, are nonparametrically identified under standard assumptions of causal mediation analysis (Avin, Shpitser and Pearl 2005). We then described an imputation approach for estimating path-specific causal effects from experimental and observational data. In contrast to existing methods for analyzing causal mediation (e.g. Imai et al. 2010, 2011), this approach does not require modeling the conditional distributions of the mediators given their antecedent variables. All we need is to model the conditional means of the outcome given treatment, pretreatment confounders, and varying sets of mediators. These conditional means, unlike the conditional distributions of the mediators, can be flexibly estimated using highly nonparametric methods such as GBM and BART. Therefore, minimal modeling assumptions are needed to implement this approach, and different models of the expected outcome can be used to check the robustness of results.

For sure, the framework introduced in this article is not without limitations. In particular, the identification of path-specific effects is premised on a set of potentially strong assumptions, which require that all relevant confounders of the treatment-outcome, treatment-mediator, mediator-mediator, and mediator-outcome relationships have been observed and adjusted for. Although standard in studies of causal mediation, these assumptions must be scrutinized against the research design and subject matter knowledge in each empirical application. In experimental studies where treatment is randomly assigned, the assumptions of no unobserved treatment-outcome or treatment-mediator confounding are met by design, but the mediator-mediator and mediator-outcome relationships may still be confounded by unobserved factors. In cases where one or more of these assumptions are questionable, estimates of path-specific effects can be combined with a general-purpose sensitivity analysis (e.g., VanderWeele 2010; VanderWeele and Arah 2011) to assess their robustness to unobserved confounding.

# A Proof of the "Mediation Formula" for Path-Specific Effects (Equation 5)

Since we interpret a DAG as Pearl's (2009) nonparametric structural equation model with independent errors, the DAG in the top panel of Figure 1 implies the following conditional independence relationships: (a) $L(a^*) \perp\!\!\!\perp M(a^{**}, l)|X$; (b) $Y(a, l, m) \perp\!\!\!\perp (L(a^*), M(a^{**}, l))|X$; (c) $L(a) \perp\!\!\!\perp A|X$; (d) $M(a, l) \perp\!\!\!\perp (A, L)|X$; (e) $Y(a, l, m) \perp\!\!\!\perp (A, L, M)|X$ (see also VanderWeele, Vansteelandt and Robins 2014). Thus we have

$$
\begin{aligned}
&\mathbb{E}[Y(a, L(a^*), M(a^{**}, L(a^*)))|X = x] \\
&= \int \mathbb{E}[Y(a, l, M(a^{**}, L(a^*)))|X = x, L(a^*) = l] f_{L(a^*)|X=x}(l) dl \\
&= \int \mathbb{E}[Y(a, l, m)|X = x, L(a^*) = l, M(a^{**}, l) = m] f_{L(a^*)|X=x}(l) f_{M(a^{**},l)|X=x}(m) dldm \quad \text{by (a)} \\
&= \int \mathbb{E}[Y(a, l, m)|X = x] f_{L(a^*)|X=x}(l) f_{M(a^{**},l)|X=x}(m) dldm \quad \text{by (b)} \\
&= \int \mathbb{E}[Y(a, l, m)|X = x] f_{L(a^*)|X=x,A=a^*}(l) f_{M(a^{**},l)|X=x,A=a^{**},L=l}(m) dldm \quad \text{by (c) and (d)} \\
&= \int \mathbb{E}[Y(a, l, m)|X = x, A = a, L = l, M = m] f(l|x, a^*) f(m|x, a^{**}, l) dldm \quad \text{by (e)} \\
&= \int \mathbb{E}[Y|x, a, l, m] f(l|x, a^*) f(m|x, a^{**}, l) dldm \qquad (14)
\end{aligned}
$$

Integrating the above expression over $f(x)$ yields equation (5).

# B Proofs of the Imputation Formulas for Path-Specific Effects (Equations 8-11)

Let us first consider equations (8) and (9). By equation (14), we have

$$\begin{aligned}\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)|X = x] &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l|x, A = 0)f(m|x, A = 0, l)dldm \\ &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l, m|x, A = 0)dldm \\ &= \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L, M]|A = 0, X\big].\end{aligned}$$

Integrating the above expression over $f(x)$ yields equation (8). Similarly,

$$\begin{aligned}\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)|X = x] &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l|x, A = 0)f(m|x, A = 1, l)dldm \\ &= \int \mathbb{E}[Y|x, A = 1, l]f(l|x, A = 0)dl \\ &= \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L]|A = 0, X\big].\end{aligned}$$

Here, the second line uses the fact that $\int \mathbb{E}[Y|x, A = 1, l, m]f(m|x, A = 1, l)dm = \mathbb{E}[Y|x, A = 1, l]$. Integrating the above expression over $f(x)$ yields equation (9). Now, consider equations (10) and (11). By the mediation formula (5), we have

$$\begin{aligned}\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)] &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l|x, A = 0)f(m|x, A = 0, l)f(x)dldmdx \\ &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l, m|x, A = 0)f(x)dldmdx \\ &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l, m, x|A = 0)\frac{f(x)}{f(x|A = 0)}dldmdx \\ &= \int \mathbb{E}[Y|x, A = 1, l, m]f(l, m, x|A = 0)\frac{\mathbb{P}(A = 0)}{\mathbb{P}(A = 0|X = x)}dldmdx \\ &= \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, L, M]\frac{\mathbb{P}[A = 0]}{\mathbb{P}[A = 0|X]}|A = 0\big].\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)] &= \int \mathbb{E}[Y|x, A=1, l, m] f(l|x, A=0) f(m|x, A=1, l) f(x) dldmdx \\
&= \int \mathbb{E}[Y|x, A=1, l] f(l|x, A=0) f(x) dldx \\
&= \int \mathbb{E}[Y|x, A=1, l] f(l, x|A=0) \frac{f(x)}{f(x|A=0)} dldmdx \\
&= \int \mathbb{E}[Y|x, A=1, l] f(l, x|A=0) \frac{\mathbb{P}(A=0)}{\mathbb{P}(A=0|X=x)} dldmdx \\
&= \mathbb{E}\big[\mathbb{E}[Y|X, A=1, L] \frac{\mathbb{P}[A=0]}{\mathbb{P}[A=0|X]} |A=0\big].
\end{aligned}$$

# References

Abramson, Scott F and David B Carter. 2016. "The Historical Origins of Territorial Disputes." *American Political Science Review* 110(4):675–698.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–529.

Albert, Jeffrey M. 2012. "Mediation Analysis for Nonlinear Models with Confounding." *Epidemiology* 23(6):879.

Albert, Jeffrey M and Suchitra Nelson. 2011. "Generalized Causal Mediation Analysis." *Biometrics* 67(3):1028–1038.

Avin, Chen, Ilya Shpitser and Judea Pearl. 2005. "Identifiability of path-specific effects.".

Brader, Ted, Nicholas A Valentino and Elizabeth Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat." *American Journal of Political Science* 52(4):959–978.

Chipman, Hugh A, Edward I George and Robert E McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics* 4(1):266–298.

Cole, Stephen R and Miguel A Hernán. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168(6):656–664.

Geneletti, Sara. 2007. "Identifying Direct and Indirect Effects in a Non-counterfactual Framework." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2):199–215.

Glynn, Adam N. 2012. "The Product and Difference Fallacies for Indirect Effects." *American Journal of Political Science* 56(1):257–269.

Hall, Matthew EK. 2017. "Macro Implementation: Testing the Causal Paths from US Macro Policy to Federal Incarceration." *American Journal of Political Science* 61(2):438–455.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Holbein, John B. 2017. "Childhood Skill Development and Adult Political Participation." *American Political Science Review* 111(3):572–583.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765–789.

Imai, Kosuke, Luke Keele, Teppei Yamamoto et al. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical science* 25(1):51–71.

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.

Knutsen, Carl Henrik, Andreas Kotsadam, Eivind Hammersmark Olsen and Tore Wig. 2017. "Mining and Local Corruption in Africa." *American Journal of Political Science* 61(2):320–334.

Lupu, Noam and Leonid Peisakhin. 2017. "The Legacy of Political Violence across Generations." *American Journal of Political Science* 61(4):836–851.

McCaffrey, Daniel F, Greg Ridgeway and Andrew R Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9(4):403.

Naimi, Ashley I, Erica EM Moodie, Nathalie Auger and Jay S Kaufman. 2014. "Constructing Inverse Probability Weights for Continuous Exposures: a Comparison of Methods." *Epidemiology* 25(2):292–299.

Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. pp. 411–420.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Reese, Michael J, Keven G Ruby and Robert A Pape. 2017. "Days of Action or Restraint? How the Islamic Calendar Impacts Violence." *American Political Science Review* 111(3):439–459.

Ridgeway, Greg, Dan McCaffrey, Andrew Morral, Lane Burgette and Beth Ann Griffin. 2017. "Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package." *Santa Monica, CA: RAND Corporation* .

Robins, James M. 2003. "Semantics of Causal DAG models and the Identification of Direct and Indirect effects." *Highly Structured Stochastic Systems* pp. 70–81.

Robins, James M and Thomas S Richardson. 2010. "Alternative graphical causal models and the identification of direct effects." *Causality and psychopathology: Finding the determinants of disorders and their cures* pp. 103–158.

Steen, Johan, Tom Loeys, Beatrijs Moerkerke and Stijn Vansteelandt. 2017. "Medflex: an R package for Flexible Mediation Analysis using Natural Effect Models." *Journal of Statistical Software* 76(11).

Tomz, Michael R and Jessica L Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107:849–865.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, Tyler J. 2009*a*. "Concerning the Consistency Assumption in Causal Inference." *Epidemiology* 20(6):880–883.

VanderWeele, Tyler J. 2009*b*. "Marginal Structural Models for the Estimation of Direct and Indirect effects." *Epidemiology* 20(1):18–26.

VanderWeele, Tyler J. 2010. "Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects." *Epidemiology (Cambridge, Mass.)* 21(4):540.

VanderWeele, Tyler J and Onyebuchi A Arah. 2011. "Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders." *Epidemiology (Cambridge, Mass.)* 22(1):42–52.

VanderWeele, Tyler J and Stijn Vansteelandt. 2009. "Conceptual Issues Concerning Mediation, Interventions and Composition." *Statistics and its Interface* 2(4):457–468.

VanderWeele, Tyler J, Stijn Vansteelandt and James M Robins. 2014. "Effect Decomposition in the Presence of an Exposure-induced Mediator-outcome Confounder." *Epidemiology* 25:300–306.

Vansteelandt, Stijn. 2009. "Estimating Direct Effects in Cohort and Case–control Studies." *Epidemiology* 20(6):851–860.

Vansteelandt, Stijn, Maarten Bekaert and Theis Lange. 2012. "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects." *Epidemiologic Methods* 1(1):131–158.

Wang, Yue, Maya L Petersen, David Bangsberg and Mark J van der Laan. 2006. "Diagnosing Bias in the Inverse Probability of Treatment Weighted Estimator Resulting from Violation of Experimental Treatment Assignment.".

Wodtke, Geoffrey and Xiang Zhou. 2019. "Effect Decomposition in the Presence of Treatment-induced Confounding: A Regression-with-residuals Approach." *SocArXiv, May 16.* .

Zhou, Xiang and Geoffrey T Wodtke. 2019. "A Regression-with-residuals Method for Estimating Controlled Direct Effects." *Political Analysis* 27(3):360–369.

Zhu, Boliang. 2017. "MNCs, Rents, and Corruption: Evidence from China." *American Journal of Political Science* 61(1):84–99.