Tracing Causal Paths from Experimental and Observational Data*

Xiang Zhou Teppei Yamamoto

Harvard University MIT

Abstract

The study of causal mechanisms abounds in political science, and causal mediation analysis has grown rapidly across different subfields. Yet, conventional methods for analyzing causal mechanisms are difficult to use when the causal effect of interest involves multiple mediators that are potentially causally dependent—a common scenario in political science applications. This article introduces a general framework for tracing causal paths with multiple mediators. In this framework, the total effect of a treatment on an outcome is decomposed into a set of path-specific effects (PSEs). We propose an imputation approach for estimating these PSEs from experimental and observational data, along with a set of bias formulas for conducting sensitivity analysis. We illustrate this approach using an experimental study on issue framing effects and an observational study on the legacy of political violence. An open-source R package, `paths`, is available for implementing the proposed methods.

Keywords— causal inference, causal mediation analysis, path-specific effects, posttreatment confounding

*Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu. The authors benefited from communications with Nate Breznau, Rocio Titiunik, and participants of the 36th Annual Meeting of the Society for Political Methodology and the 2019 Annual Meeting of the American Political Science Association. The authors thank Minh Trinh for his research assistance. Replication files are available in the JOP Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the JOP replication analyst. Supplementary material for this article is available in the online edition.

The study of causal mechanisms abounds in political science. In political psychology, for example, scholars investigate the pathways through which the framing of political issues in mass media and elite communications affects citizens' attitudes and behavior (e.g. Druckman and Nelson 2003; Nelson et al. 1997a; Slothuus 2008). In political economy, a growing body of research examines the mechanisms through which historical events shape contemporary social and political outcomes (e.g., Acharya et al. 2016b; Lupu and Peisakhin 2017; Mazumder 2018). Over the past decade, studies of causal mediation have grown rapidly across different subfields of political science because empirical evaluation of the mechanisms hypothesized to transmit causal effects is central for testing and refining theories of social and political processes (Acharya et al. 2016a; Imai et al. 2011).

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 the so-called average natural direct and indirect effects (Pearl 2001), 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 causal effect of interest involves multiple, potentially overlapping, causal pathways—a common scenario in political 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 (Imai et al. 2010; VanderWeele 2015). Of these assumptions, (iv) is especially restrictive because it requires that there must not be any post-treatment variables that affect both the mediator and outcome, whether they are observed or not.

Consequently, if two mediators are present and one mediator affects both the other mediator and the outcome, the ACME for the second mediator cannot be identified without functional form assumptions (Imai and Yamamoto 2013). To circumvent this problem, empirical studies have often assumed, sometimes implicitly, that different mediators are causally independent (i.e., they do

not affect each other), an assumption that is strong, untestable, and unrealistic in many applications. Moreover, when the causal effect of interest involves multiple mediators that are causally dependent, the causal pathways through those mediators are *not* mutually exclusive, rendering their mediating effects inseparable even conceptually. In fact, the overlapping of causal pathways via different mediators may require us to reformulate and reassess the "competing hypotheses" of underlying processes. The prevailing practice of treating causally dependent mediators as independent can be both methodologically problematic and theoretically inaccurate.

In this article, we show that in the presence of multiple mediators, a more fruitful approach to analyzing causal mechanisms is to trace different causal paths explicitly. Specifically, we make three novel contributions to the methodological toolbox for causal mediation analysis. First, drawing on a previous identification result for path-specific effects (PSEs; Avin et al. 2005), we provide a general framework for effect decomposition with an arbitrary number of mediators. In particular, we provide, for the first time, a general formula that decomposes the total effect of treatment into $K+1$ PSEs — one "direct effect" and K mutually exclusive indirect effects — in the presence of K causally ordered mediators. This is in contrast to the previous literature on PSEs, which has focused on the case of two mediators (e.g., Albert and Nelson 2011; Daniel et al. 2015). The $K+1$ PSEs are nonparametrically identified under the assumption that observed variables can be arranged in a directed acyclic graph (DAG) and, in this DAG, no unobserved confounding exists for any of the treatment-outcome, treatment-mediator, and mediator-outcome relationships (Pearl 2009).

Second, we develop a new method for estimating the PSEs. Our proposed method, based on model-assisted imputation of counterfactual outcomes, holds several distinct advantages over conventional methods for analyzing causal mediation (e.g., Baron and Kenny 1986; Imai et al. 2011). First, it can accommodate either one or multiple mediators, whether different mediators are treated as causally independent, causally dependent, or analyzed as a whole. The proposed approach can therefore be applied to broader empirical settings than are possible with existing approaches. Second, in contrast to the simulation approach developed by Imai et al. (2010), the imputation approach does not require modeling the conditional distributions of the mediators given their antecedent vari-

ables. This is especially appealing because in many political science applications, the mediators of interest are continuous and/or multivariate, making it practically difficult to model their conditional distributions. The imputation approach, instead, involves modeling only the conditional *means* of the outcome variable itself, given treatment, pretreatment confounders, and varying sets of mediators. Estimating conditional means as opposed to distributions is substantially less demanding in terms of both statistical power and the assumptions required, and the analyst needs correct modeling assumptions only for the outcome variable, not for any of the mediators. Moreover, these models can be fit via any method of the analyst's choice, be it linear regression, generalized linear models (GLM), or, as we will illustrate, data-adaptive methods such as Bayesian Additive Regression Trees (BART; Chipman et al. 2010; Hill 2011).

Third, we propose a set of bias formulas for assessing the sensitivity of estimated PSEs to the unconfoundedness assumptions required. Although these assumptions are customary in the mediation literature (VanderWeele 2015), it is never possible to completely rule out the presence of unobserved confounding in many empirical settings (Bullock et al. 2010). To address this limitation, we develop a bias factor approach for conducting sensitivity analysis with regard to unobserved confounding for the mediator-outcome relationships — which may occur in both experimental and observational studies. As an extension of the bias formulas developed by VanderWeele (2010) for the single-mediator setting, our approach provides a set of general-purpose formulas that allow us to calculate potential biases of the estimated PSEs due to unobserved confounding — regardless of the models used to estimate the PSEs.

Taken together, these methodological innovations represent a new, more general framework for analyzing causal mechanisms in empirical political science research. Our framework improves upon existing approaches (e.g. Imai et al., 2011) by allowing multiple mediators, offering a finer decomposition of the treatment effect into multiple PSEs, each corresponding to one of the mediators, and providing a method for sensitivity analysis. Applied researchers can adopt our framework to make richer inferences about how causal effects operate through multiple pathways. To facilitate practice, we offer an open-source R package, `paths`, for implementing all of the proposed methods, which is

available at the Comprehensive R Archive Network (CRAN).

The rest of the paper is organized as follows. For ease of exposition, we start with the case of two causally ordered mediators, for which we present a decomposition of the total effect of treatment into a set of PSEs, outline the assumptions needed for identifying these PSEs, and introduce an imputation approach to estimation. We next generalize the framework for defining, identifying, and estimating PSEs to the setting with an arbitrary number of causally ordered mediators. We then describe the bias factor approach to sensitivity analysis. Finally, we illustrate these methods using several empirical examples where researchers have endeavored to disentangle causal pathways in the presence of multiple causally dependent mediators.

Path-Specific Causal Effects

In political psychology, scholars study how issue framing, i.e., a presenter's deliberate emphasis on certain aspects of a political issue, shapes citizens' attitudes and behavior (Chong and Druckman 2007; Nelson et al. 1997b). An important debate in this literature concerns whether issue framing affects citizens' opinions by altering their beliefs about the issue (hereafter the "belief" mediator) or by changing their perceived importance of different issue-related considerations (hereafter the "importance" mediator) (e.g., Druckman and Nelson 2003; Nelson et al. 1997a; Nelson and Oxley 1999; Slothuus 2008). To assess the relative importance of these two mechanisms, Slothuus (2008) conducted a survey experiment on a sample of 408 Danish students. Specifically, the author examined how two versions of a newspaper article on a social welfare reform bill—one highlighting the reform's purported positive effect on job creation (the "job frame") and the other emphasizing its negative impact on the poor (the "poor frame")—affect the respondent's support for the reform. After randomly assigning respondents to either the job frame or the poor frame, the author asked them a series of five-point-scale questions to measure (a) their beliefs about why some people receive welfare benefits, or who is responsible for the situation of welfare recipients and (b) their perceived importance of competing issue-related considerations (e.g., work incentives versus living conditions among the poor). Finally, the author

measured the outcome variable by asking the respondents whether and to what extent they support the proposed welfare reform.

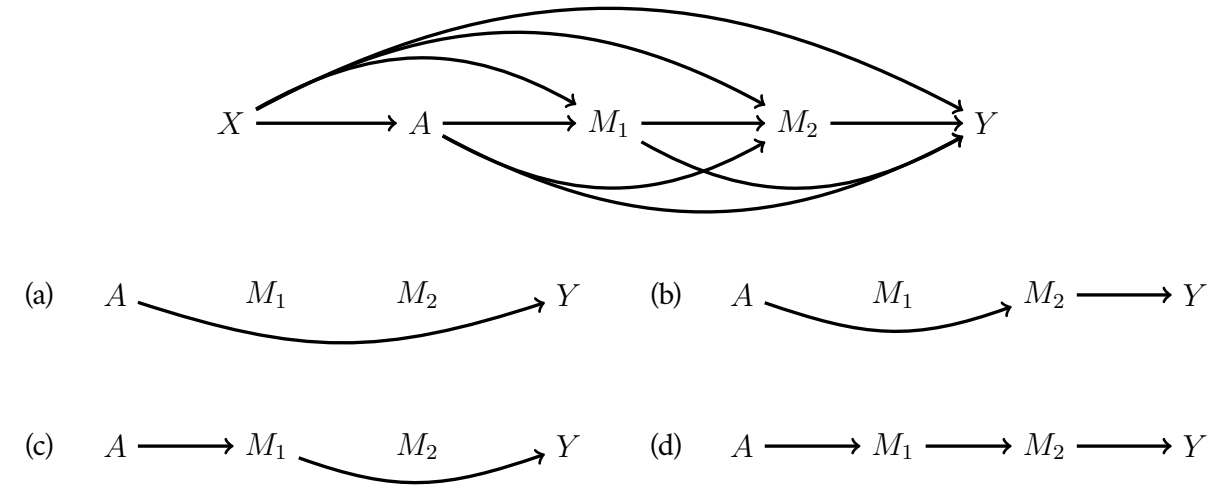
In this study, the author implicitly assumes that the belief mediator and the importance mediator are causally independent. This assumption would be violated if, for example, issue framing induced respondents to modify their beliefs about why some people received welfare benefits, and, in turn, their modified beliefs caused a change in their perceived importance of competing considerations. In fact, this is a major concern in the framing effects literature. As Miller (2007, 711-712) points out on the basis of her experimental study, "individuals use information obtained from the media to evaluate how important issues are," and "when media exposure to an issue causes negative emotional reactions about the issue, increased importance judgments will follow." Moreover, Imai and Yamamoto's (2013, 153) reanalyses of Slothuus's data suggest that the independence assumption is unlikely to hold in this application. If this is the case, the ACME of the importance mediator cannot be nonparametrically identified, since the belief mediator acts as a treatment-induced confounder between the importance mediator and the outcome. Yet, as we will show, we can still identify the strength of the causal path *issue frame→importance→support for welfare reform*, which represents the amount of treatment effect operating via the perceived importance of competing considerations *above and beyond* that operating via the respondent's issue-related beliefs. This quantity is substantively important because it reflects the independent role of the importance mediator in transmitting the framing effect.

Path-Specific Effects

We use A to denote a binary treatment, Y an outcome of interest, and X a vector of observed pretreatment confounders. Although our framework can accommodate an arbitrary number of mediators, for ease of exposition, we first consider the case where two (sets of) mediators, M_1 and M_2, lie on the causal paths from A to Y. We assume that M_1 precedes M_2, such that no component of M_2 can causally affect any component of M_1.[1] A causal DAG that is consistent with the hypothesized rela-

[1]Note that M_1 and M_2 can each consist of multiple variables and that the causal relationships among the component variables can be left unspecified, as long as M_1 causally precedes M_2.

Figure 1: Causal Relationships with Two Causally Ordered Mediators.



Note: A denotes the treatment, Y denotes the outcome of interest, X denotes a vector of pretreatment covariates, and M_1 and M_2 denote two causally ordered mediators. The confounding arcs between X and each of the other nodes are omitted in subgraphs (a)-(d).

tionships between these variables is shown in the top panel of Figure 1. In Slothuus's (2008) study on issue framing effects, A represents the issue frame presented to the respondent, Y represents the respondent's support for the proposed welfare reform, M_1 represents the respondent's beliefs about why some people receive welfare benefits, and M_2 represents the respondent's perceived importance of competing considerations.

In this DAG, four possible paths exist from the treatment to the outcome, as shown in the lower panels of Figure 1: (a) $A \rightarrow Y$; (b) $A \rightarrow M_2 \rightarrow Y$; (c) $A \rightarrow M_1 \rightarrow Y$; and (d) $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$. If the mediators M_1 and M_2 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 M_1 ($A \rightarrow M_1 \rightarrow Y$), the effect operating through M_2 ($A \rightarrow M_2 \rightarrow Y$), and a "direct" effect not operating through M_1 or M_2 ($A \rightarrow Y$) (Imai and Yamamoto, 2013). However, in the general case where M_1 and M_2 are causally dependent, it is not possible to partition the mediating effects of M_1 and M_2 into their respective components, since some of the total effect of A on Y operates through both M_1 and M_2, as represented by the path $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$.

To define the PSEs formally, we use the potential outcomes notation. Specifically, we use

$Y(a, m_1, m_2)$ to denote the potential outcome under treatment status a and mediator values $M_1 = m_1$ and $M_2 = m_2$, $M_2(a, m_1)$ to denote the potential value of the mediator M_2 under treatment status a and mediator value $M_1 = m_1$, and $M_1(a)$ to denote the potential value of the mediator M_1 under treatment status a. This notation allows us to define nested counterfactuals. For example, $Y\big(1, M_1(0), M_2(0, M_1(0))\big)$ represents the potential outcome in the hypothetical scenario where the unit was treated but the mediators M_1 and M_2 were set to values they would have taken had the subject not been treated. Further, if we let $Y(a)$ denote the potential outcome when treatment status is set to a and the mediators M_1 and M_2 take on their "natural" values under treatment status a (i.e., $M_1(a)$ and $M_2(a, M_1(a))$), we have $Y(a) = Y\big(a, M_1(a), M_2(a, M_1(a))\big)$ by definition.

Under the above notation, the average total effect (henceforth ATE) of A on Y can be written as a telescoping sum (VanderWeele et al. 2014):

$$
\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[Y\big(1, M_1(1), M_2(1, M_1(1))\big) - Y\big(0, M_1(0), M_2(0, M_1(0))\big)] \\
&= \underbrace{\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big) - Y\big(0, M_1(0), M_2(0, M_1(0))\big)]}_{A \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big) - Y\big(1, M_1(0), M_2(0, M_1(0))\big)]}_{A \to M_2 \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(1, M_1(1), M_2(1, M_1(1))\big) - Y\big(1, M_1(0), M_2(1, M_1(0))\big)]}_{A \to M_1 \to Y; A \to M_1 \to M_2 \to Y} \\
&\equiv \tau_{A \to Y} + \tau_{A \to M_2 \to Y} + \tau_{A \to M_1 \rightsquigarrow Y}, \qquad (1)
\end{aligned}
$$

The three terms in equation (1) represent the PSEs for causal paths $A \to Y$, $A \to M_2 \to Y$, and $A \to M_1 \rightsquigarrow Y$, respectively, with a straight arrow denoting a single direct path and a squiggly arrow representing a combination of multiple paths.[2] Specifically, the first term ($\tau_{A \to Y}$) corresponds

[2]Equation (1) is not the only way of defining the PSEs for the causal paths $A \to Y$, $A \to M_2 \to Y$, and $A \to M_1 \rightsquigarrow Y$. An alternative decomposition, for example, can be obtained by switching the 0s and 1s in equation (1) and then flipping the signs of both sides. In general, when the treatment and the mediators have an interaction effect on the outcome, the PSEs defined by these alternative decompositions will be different. We focus on equation (1) in the main text and illustrate the above

to the amount of treatment effect if the mediators M_1 and M_2 were set to values they would have taken under treatment status $A = 0$ for each unit, representing the causal path $A \to Y$. The second term ($\tau_{A \to M_2 \to Y}$) corresponds to the amount of treatment effect operating through the mediator M_2 under treatment status $A = 1$ and mediator status $M_1 = M_1(0)$, representing the causal path $A \to M_2 \to Y$. The last term ($\tau_{A \to M_1 \rightsquigarrow Y}$) corresponds to the amount of treatment effect operating through the mediator M_1 under treatment status $A = 1$. It represents the causal path $A \to M_1 \rightsquigarrow Y$, or the combination of the causal paths $A \to M_1 \to Y$ and $A \to M_1 \to M_2 \to Y$.

Although four causal paths exist from A to Y, equation (1) partitions the ATE into only three components: $\tau_{A \to Y}$, $\tau_{A \to M_2 \to Y}$, and $\tau_{A \to M_1 \rightsquigarrow Y}$. In particular, the last component $\tau_{A \to M_1 \rightsquigarrow Y}$ encompasses both the causal path $A \to M_1 \to Y$ and the causal path $A \to M_1 \to M_2 \to Y$. It reflects the overall mediating effect of M_1, some of which may also operate through M_2. By contrast, the component $\tau_{A \to M_2 \to Y}$ captures only the causal path $A \to M_2 \to Y$, but not $A \to M_1 \to M_2 \to Y$. Thus it should not be interpreted as the overall mediating effect of M_2. Instead, it reflects the "independent" mediating effect of M_2, i.e., the mediating effect of M_2 above and beyond that of M_1.

Thus, in the issue framing example, $\tau_{A \to Y}$ reflects the direct effect of issue framing on the respondent's support for welfare reform, i.e., the fraction of the total effect operating neither through the belief mediator nor through the importance mediator; $\tau_{A \to M_2 \to Y}$ reflects the effect of issue framing operating *only* through changing the respondent's perceived importance of competing considerations; and $\tau_{A \to M_1 \rightsquigarrow Y}$ reflects the effect of issue framing operating through changing the respondent's beliefs about the issue, regardless of whether the modified beliefs subsequently change the perceived importance of competing considerations.

Identification

Following Pearl (2009), we use a DAG to denote a nonparametric structural equation model with mutually independent errors. In this framework, the top panel of Figure 1 corresponds to a set of nonparametric structural equations that underlie our key identification assumption: *no confounding*

alternative decomposition in Supporting Information (SI) F.

exists for any of the treatment-mediator, treatment-outcome, and mediator-outcome relationships after conditioning on their antecedent variables (see SI A). This assumption is much stronger than the standard ignorability assumption that researchers often invoke to identify the ATE in observational studies. Unlike the standard ignorability assumption, which stipulates the conditional independence between treatment and potential outcomes, this assumption involves multiple conditional independence relationships, some of which pertain to conditional independence between the so-called "cross-world counterfactuals," such as $Y(a, m_1, m_2) \perp\!\!\!\perp M_1(a_1)|X, A$ for any a, a_1, m_1, m_2. Such cross-world independence relationships will generally be violated when posttreatment confounders are present for any of the mediator-outcome relationships (Richardson and Robins 2013). Thus, in practice, to reduce the bias due to potential posttreatment confounding, we recommend that all observed posttreatment variables be included as components of M_1 or M_2, depending on the hypothesized causal order among these variables. Finally, we note that our identification assumption does not rule out all forms of unobserved confounding for the causal effects of X on its descendants. For example, unobserved variables are permitted (although not shown) in Figure 1 that affect both X and Y.

Under the above assumption, it can be shown that the PSEs defined by equation (1) are nonparametrically identified (Avin et al. 2005). To identify the components of equation (1), it suffices to identify the counterfactual expectation $\mathbb{E}[Y\big(a, M_1(a_1), M_2(a_2, M_1(a_1))\big)]$ for any combination of a, a_1, $a_2 \in \{0, 1\}$. As proved in SI A, this quantity can be written as a function of observed variables:

$$\begin{aligned} &\mathbb{E}[Y\big(a, M_1(a_1), M_2(a_2, M_1(a_1))\big)] \\ &= \iiint \mathbb{E}[Y|x, a, m_1, m_2] f(m_2|x, a_2, m_1) f(m_1|x, a_1) f(x) dm_2 dm_1 dx, \end{aligned} \tag{2}$$

where $f(\cdot)$ denotes a probability density/mass function. This equation generalizes Pearl's (2001) mediation formula to the case of two (sets of) causally dependent mediators (see also Daniel et al. 2015).

Note that the last term in equation (1), i.e., $\tau_{A \to M_1 \rightsquigarrow Y}$, reflects the combination of the causal paths $A \to M_1 \to Y$ and $A \to M_1 \to M_2 \to Y$. Without additional assumptions, the PSEs for the paths $A \to M_1 \to Y$ and $A \to M_1 \to M_2 \to Y$ cannot be separately identified. In the issue framing study,

for example, we can identify the overall mediating effect via the respondent's beliefs about the issue ($A \to M_1 \rightsquigarrow Y$), but we cannot pinpoint how much of this mediating effect further operates through the perceived importance of competing considerations ($A \to M_1 \to M_2 \to Y$). Similarly, we can identify the "independent" mediating effect via the respondent's perceived importance of competing considerations ($A \to M_2 \to Y$), but we cannot gauge the overall effect of the importance mediator, which involves both $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$. Nonetheless, the independent mediating effect is arguably more interesting here because it reflects the effect of the importance mediator above and beyond that of the belief mediator — an effect that would persist even if issue framing did not affect the respondent's beliefs about what had caused the plight of welfare recipients.

Comparison with Existing Approaches

Existing work on causal mediation analysis with multiple mediators has focused on the ACME via each of the mediators, instead of the PSEs. For example, Imai and Yamamoto (2013) consider the following decomposition of the ATE:

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y(1, M_1(1), M_2(0, M_1(0)))] - \mathbb{E}[Y(0, M_1(0), M_2(0, M_1(0)))]}_{A \to Y;\, A \to M_1 \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y(1, M_1(1), M_2(1, M_1(1)))] - \mathbb{E}[Y(1, M_1(1), M_2(0, M_1(0)))]}_{A \to M_2 \to Y;\, A \to M_1 \to M_2 \to Y} \\
&\equiv \text{ADE}_{M_2}(0) + \text{ACME}_{M_2}(1), \qquad (3)
\end{aligned}$$

Here, $\text{ACME}_{M_2}(1)$ represents the amount of treatment effect operating through M_2 (under treatment status $A = 1$), whether the effect also operates through M_1 or not. Similarly, $\text{ADE}_{M_2}(0)$ reflects the amount of treatment effect that does not operate through M_2, regardless of M_1.

The above decomposition is useful when the researcher's substantive interest lies solely in the mediator M_2, whereas the other mediator M_1 is purely a nuisance that needs to be accounted for due to the confounding it causes between M_2 and Y. A limitation of this approach, however, is that neither the ACME nor the ADE for M_2 can be nonparametrically identified because M_1 is a treatment-

Table 1: Path-Specific Effects (PSEs) that Compose the Average Causal Mediation Effects (ACMEs) and Average Direct Effects (ADEs) in the Presence of Two Causally Dependent Mediators.

	ADE for M_2	ACME for M_2
ADE for M_1	PSE for $A \to Y$	PSE for $A \to M_2 \to Y$
ACME for M_1	PSE for $A \to M_1 \to Y$	PSE for $A \to M_1 \to M_2 \to Y$

Note: Under the assumption that the treatment and mediators do not have interaction effects (i.e. the no-interaction assumption; Robins 2003), the PSE for each path is uniquely defined (i.e., they do not depend on the reference levels chosen for the other paths), and each of the ADEs and ACMEs equals the sum of the two component PSEs shown in the same row/column in the table. Without the no-interaction assumption, these relationships still hold, although the rows and the columns correspond to different PSE decompositions. The PSE decomposition defined by equation (1) corresponds to the rows; that is, $\tau_{A\to Y} + \tau_{A\to M_2\to Y}$=ADE$_{M_1}(0)$, and $\tau_{A\to M_1 \rightsquigarrow Y} = \text{ACME}_{M_1}(1)$.

induced confounder of the relationship between M_2 and Y. Moreover, empirical researchers are often in a situation where both M_1 and M_2 are of substantive interest, making it inappropriate to treat the mediator M_1 as purely a nuisance.

In contrast, our proposed approach begins with the following alternative decomposition:

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big) - Y\big(0, M_1(0), M_2(0, M_1(0))\big)]}_{A\to Y; A\to M_2\to Y} \\
&\quad + \underbrace{\mathbb{E}[Y\big(1, M_1(1), M_2(1, M_1(1))\big) - Y\big(1, M_1(0), M_2(1, M_1(0))\big)]}_{A\to M_1\to Y; A\to M_1\to M_2\to Y} \\
&\equiv \text{ADE}_{M_1}(0) + \text{ACME}_{M_1}(1), \qquad (4)
\end{aligned}$$

where the two terms represent the ADE and ACME with respect to M_1, rather than M_2. A comparison of equation (4) with equation (1) reveals that $\text{ACME}_{M_1}(1) = \tau_{A\to M_1 \rightsquigarrow Y}$ and $\text{ADE}_{M_1}(0) = \tau_{A\to Y} + \tau_{A\to M_2\to Y}$. Thus, our proposed approach allows us to estimate the amount of treatment effect that operates through M_1 (i.e., $\text{ACME}_{M_1}(1)$), and, furthermore, to decompose the ADE for M_1 into the effect operating through M_2 but not through M_1 ($\tau_{A\to M_2\to Y}$) and the effect operating neither through M_1 nor through M_2 ($\tau_{A\to Y}$).

Table 1 summarizes how the PSEs relate to the ACMEs and ADEs with respect to M_1 and M_2.

We can see that the PSEs generally represent further decompositions of the ACMEs and ADEs. The table also shows that, if the mediators M_1 and M_2 are causally independent, i.e., if the causal path $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ (bottom right) does not exist, the ACMEs for M_1 and for M_2 will amount to PSEs specific to these mediators. The prevailing practice of treating different mediators as causally independent can therefore be seen as a special case of our approach. Thus, even in applications where the analyst is willing to assume that different mediators are causally independent, our framework for defining, identifying, and estimating PSEs can still be applied, except that the estimated PSEs can now be equivalently interpreted as the overall indirect effects via the corresponding mediators.

Finally, we note that the PSEs are distinct from the controlled direct effect (CDE), an estimand recently advocated for analyzing causal mechanisms in political science (e.g., Acharya et al. 2016a; Zhou and Wodtke 2019). The CDE measures the strength of the causal relationship between a treatment and outcome when a mediator is fixed at a given value for all units. Compared with the ACME, an advantage of the CDE is that it can still be identified in the presence of posttreatment confounders of the mediator-outcome relationship, provided that these confounders are observed. In practice, the CDE is useful in contexts where it is reasonable to entertain a policy intervention that sets the mediator at a given value for all units. However, unlike the ACME and PSEs, the CDE does not directly gauge the strengths of different causal paths from the treatment to the outcome.

Estimating Path-Specific Effects

To date, most estimation methods for causal mediation analysis have focused on the setting involving a single mediator or a set of mediators considered as a whole. 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, and $a, a^* \in \{0, 1\}$. Various estimators have been proposed for this quantity (e.g., Imai et al. 2010; Tchetgen Tchetgen and Shpitser 2012). In particular, Vansteelandt et al. (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 outcome $Y(a, M(a^*))$ for each unit with treatment status a^*, and (c) fitting a model of these imputed counterfactuals conditional on the pretreatment confounders. Albert (2012) proposed a similar method, in which the first two steps are the same and the last step involves an inverse-probability-of-treatment-weighted average of the imputed counterfactuals.

Here, we develop a method for estimating the PSEs by extending these imputation-based methods to the case of potential outcomes involving multiply nested counterfactuals. We start with the setting of two causally ordered mediators, as shown in Figure 1, and discuss the general case of $K(\geq 1)$ causally ordered mediators in the next section.

An Imputation Approach

Consider equation (1). Because the PSEs $\tau_{A\to Y}$, $\tau_{A\to M_2\to Y}$, $\tau_{A\to M_1 \rightsquigarrow Y}$ are governed by four counterfactual means $\mathbb{E}[Y(0)]$, $\mathbb{E}[Y(1)]$, $\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)]$, and $\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)]$, it suffices to estimate each of these latter quantities. 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 assigned, they can be estimated using simple averages of the observed outcome within the control and treatment groups.

Using the mediation formula (2), the latter two quantities, $\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)]$ and $\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)]$, can be written as

$$\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y|X, A=1, M_1, M_2]|X, A=0\big]\Big] \tag{5}$$

$$\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y|X, A=1, M_1]|X, A=0\big]\Big]. \tag{6}$$

A proof of these equations is given in SI B. Thus, to evaluate these nested counterfactuals, we need only estimate (a) the conditional means $\mathbb{E}[Y|X, A=1, M_1, M_2]$ and $\mathbb{E}[Y|X, A=1, M_1]$, 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 (5) and (6) can be

estimated using their sample analogs.

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

$$\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)] = \mathbb{E}\left[\mathbb{E}[Y|X, A=1, M_1, M_2]\frac{\Pr[A=0]}{\Pr[A=0|X]}\Big|A=0\right] \quad (7)$$

$$\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)] = \mathbb{E}\left[\mathbb{E}[Y|X, A=1, M_1]\frac{\Pr[A=0]}{\Pr[A=0|X]}\Big|A=0\right]. \quad (8)$$

These equations suggest that to evaluate the nested counterfactuals, we need only estimate $\mathbb{E}[Y|X, A=1, M_1, M_2]$, $\mathbb{E}[Y|X, A=1, M_1]$, and the probability ratio $\Pr[A=0]/\Pr[A=0|X]$. After these estimates are obtained, the outer expectation in equations (7) and (8) can be estimated using their sample analogs.

Hence, equations (5-6) and (7-8) suggest two different routes to estimating the nested counterfactuals $\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)]$ and $\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)]$. They can be seen as extensions of Vansteelandt et al.'s (2012) and Albert's (2012) imputation-based estimators for the ACME to the estimation of PSEs, respectively. Since the first procedure involves only model-based imputation and the second procedure involves both imputation and inverse probability weighting, we refer to them as a "pure imputation estimator" and an "imputation-based weighting estimator," respectively.

An important advantage of our proposed estimators over existing approaches to causal mediation (e.g., Imai et al. 2010) is that they do not require estimating the conditional densities/probabilities of the mediators. Our approach therefore obviates the problem of high instability and model sensitivity in the common empirical setting where the mediators M_1 and M_2 are multivariate and/or continuous. Moreover, the proposed approach only requires the analyst to correctly specify models for the outcome, not for any of the mediators. This will likely reduce the possibility of model misspecification, since researchers often have better substantive understandings of the generative process for the outcome variable itself than for the mediators. Below, we provide a step-by-step guide to the implementation of these estimators in experimental and observational studies.

Implementation

First, consider the experimental setting where treatment is randomly assigned. In this case, because treatment status A is independent of the pretreatment confounders X, both equations (5-6) and equations (7-8) reduce to

$$\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)] = \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, M_1, M_2]|A = 0\big]$$
$$\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)] = \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, M_1]|A = 0\big].$$

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

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

2. Fit an outcome model conditional on the treatment A, the mediators M_1 and M_2, and the pretreatment confounders X. For the control units, impute their counterfactual outcome $Y\big(1, M_1(0), M_2(0, M_1(0))\big)$ by setting $A = 1$ (while using their observed values of X, M_1, and M_2). The average of these imputed counterfactuals constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)]$.

3. Fit an outcome model conditional on the treatment A, the mediator M_1, and the pretreatment confounders X. For the control units, impute their counterfactual outcome $Y\big(1, M_1(0), M_2(1, M_1(0))\big)$ by setting $A = 1$ (while using their observed values of X, M_1). The average of these imputed counterfactuals constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)]$.

4. Calculate the PSEs as defined in equation (1).

In practice, to reduce model dependence, data-adaptive/machine learning methods can be used to fit the outcome models in steps 2 and 3. This can be useful for mitigating bias due to model misspecification, especially when nonlinear or interaction effects are likely to exist (Glynn 2012). Approximate

standard errors and confidence intervals can be constructed by bootstrapping steps 1-4.

In observational studies, the pure imputation estimator (equations 5-6) and the imputation-based weighting estimator (equations 7-8) 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 $\hat{\mathbb{E}}[Y|X, A = 0]$ and $\hat{\mathbb{E}}[Y|X, A = 1]$ among all units, respectively.

2. Fit an outcome model conditional on the treatment A, the mediators M_1 and M_2, and the pretreatment confounders X. For the untreated units, impute their counterfactual outcome $Y\big(1, M_1(0), M_2(0, M_1(0))\big)$ by setting $A = 1$ (while using their observed values of X, M_1, and M_2).

3. Fit a model of the imputed counterfactual $\hat{Y}\big(1, M_1(0), M_2(0, M_1(0))\big)$ conditional on X among the untreated units, and obtain model-based predictions for all units. The average of these predictions constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)]$.

4. Fit an outcome model conditional on the treatment A, the mediator M_1, and the pretreatment confounders X. For the untreated units, impute their counterfactual outcome $Y\big(1, M_1(0), M_2(1, M_1(0))\big)$ by setting $A = 1$ (while using their observed values of X and M_1).

5. Fit a model of the imputed counterfactual $\hat{Y}\big(1, M_1(0), M_2(1, M_1(0))\big)$ conditional on X among the untreated units, and obtain model-based predictions for all units. The average of these predictions constitutes an estimate of the counterfactual mean $\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)]$.

6. Calculate the PSEs as defined in equation (1).

The imputation-based weighting estimator requires an estimate of the probability ratio $\Pr[A = 0]/\Pr[A = 0|X]$. To that end, we can first estimate the numerator $\Pr[A = 0]$ using its sample analog and the denominator $\Pr[A = 0|X]$ using a propensity score model for the treatment. Then, repeat the above procedure while replacing steps 3 and 5 with the following steps, each of which utilizes an inverse-probability weighted average instead of model-based predictions:

3*. Estimate $\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)]$ using a weighted average of the imputed counterfactuals $\hat{Y}\big(1, M_1(0), M_2(0, M_1(0))\big)$ among the untreated units, with weight $\widehat{\Pr}[A = 0]/\widehat{\Pr}[A = 0|X]$.

5*. Estimate $\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)]$ using a weighted average of the imputed counterfactuals $\hat{Y}\big(1, M_1(0), M_2(1, M_1(0))\big)$ among the untreated units, with weight $\widehat{\Pr}[A = 0]/\widehat{\Pr}[A = 0|X]$.

To reduce model dependence, data-adaptive/machine learning methods can be used to fit the outcome models, and, for the imputation-based weighting estimator, also the propensity score model. Approximate standard errors and confidence intervals can be constructed by bootstrapping steps 1-6.

Alternative Estimation Methods

In statistics and epidemiology, several alternative methods have been proposed to estimate PSEs. VanderWeele et al. (2014) proposed a weighting estimator that involves estimating the conditional densities/probabilities of the mediators M_1 and M_2 given their antecedent variables. This estimator, however, is difficult to use when either or both of the mediators is multivariate or continuous, in which case estimates of the conditional density/probability functions $f(m_2|x, a, m_1)$ and $f(m_1|x, a)$ tend to be unstable and highly sensitive to model misspecification (Kang and Schafer 2007). Moreover, even if models for these conditional densities/probabilities are correctly specified, weighting estimators are often inefficient and susceptible to large finite sample biases (Cole and Hernán 2008; Zhou and Wodtke 2020). Miles et al. (2017) proposed a maximum likelihood estimator that is generally more efficient than the weighting estimator. However, like the weighting estimator, the maximum

likelihood estimator also involves estimating the conditional densities/probabilities of the mediators, making it difficult to use in the presence of multivariate/continuous mediators.

For a specific PSE in the two-mediator setting, Miles et al. (2020) developed a semiparametric estimator based on the efficient influence function of the estimand. Compared with the weighting, imputation, and maximum likelihood estimators, this semiparametric estimator is more robust to model misspecification in that it remains consistent even if some of the treatment/mediator/outcome models on which it depends are misspecified. Moreover, when data-adaptive methods, combined with sample splitting, are used to fit the nuisance functions, theoretically valid standard errors can be constructed from the sample variance of the estimated influence function (Zheng and van der Laan 2011; Chernozhukov et al. 2018). In related work, we have extended this approach for more general PSEs in settings with more than two mediators (Zhou 2022).

Generalization to $K(\geq 1)$ Causally Ordered Mediators

So far, we have assumed that two mediators lie on the causal paths from A to Y. The definition, identification, and estimation of PSEs can be generalized to the setting where the treatment effect operates through K causally ordered (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 component of M_j can causally affect any component of 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 definition.

The ATE of A on Y can now be decomposed as

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y\big(1, \mathcal{M}_K(0)\big) - Y(0)]}_{A \to Y} + \sum_{k=1}^{K} \underbrace{\mathbb{E}[Y\big(1, \mathcal{M}_{k-1}(0)\big) - Y\big(1, \mathcal{M}_k(0)\big)]}_{A \to M_k \rightsquigarrow Y} \\ &= \tau_{A \to Y} + \sum_{k=1}^{K} \tau_{A \to M_k \rightsquigarrow Y}. \end{aligned} \tag{9}$$

We assume that the variables $A, M_1, \dots M_K, Y$ follow a DAG that encodes a nonparametric struc-

tural equation model with mutually independent errors, such that no unobserved confounding exists for any of the treatment-mediator, treatment-outcome, and mediator-outcome relationships.

To identify the components of equation (9), it suffices to identify the counterfactual expectations $\mathbb{E}[Y(0)]$, $\mathbb{E}[Y(1)]$, and $\mathbb{E}[Y\big(1, \mathcal{M}_k(0)\big)]$ for all $k \in \{1, \dots K\}$. Similar to the two-mediator setting, these counterfactual expectations can be expressed as functions of observed variables:

$$\mathbb{E}[Y\big(1, \mathcal{M}_k(0)\big)] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y|X, A = 1, \mathcal{M}_k]|X, A = 0\big]\Big] \tag{10}$$

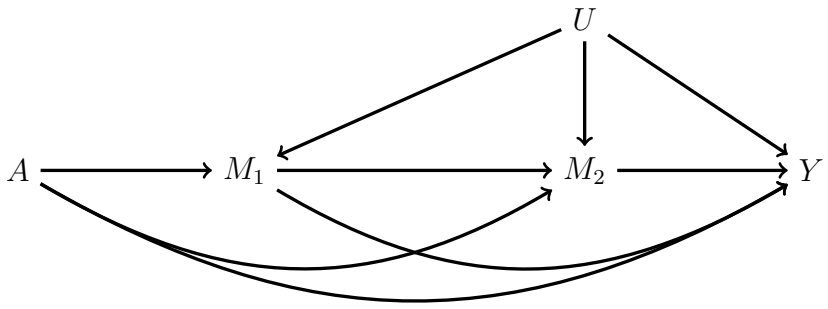
$$= \mathbb{E}\big[\mathbb{E}[Y|X, A = 1, \mathcal{M}_k]\frac{\Pr[A = 0]}{\Pr[A = 0|X]}|A = 0\big]. \tag{11}$$

Equations (10) and (11) suggest a pure imputation estimator and an imputation-based weighting estimator, respectively, for the PSEs defined in equation (9). The algorithms for implementing these estimators are detailed in SI C. In the *Empirical Illustrations* Section, we illustrate the case of three causally ordered mediators ($K = 3$) with an empirical example on the legacy of political violence.

Sensitivity Analysis for Unobserved Confounding

The identification of PSEs is premised on a nonparametric structural equation model in which no unobserved confounding exists for any of the treatment-outcome, treatment-mediator, and mediator-outcome relationships. In observational studies where treatment is not randomly assigned, all of these assumptions must be scrutinized. If any are violated, estimates of PSEs will likely be biased. In experimental studies where treatment is randomly assigned, the assumptions of no unobserved treatment-outcome and treatment-mediator confounding are met by design, but it remains possible that some of the mediator-outcome relationships are confounded by unobserved factors. To address this concern, we develop a bias factor approach to sensitivity analysis that allows us to assess the degree to which estimates of PSEs are robust to unobserved confounding of the mediator-outcome relationships. This approach can be seen as an extension of the bias formulas developed by VanderWeele (2010) to the setting of multiple causally dependent mediators. For ease of exposition, we focus on the case of two causally ordered mediators in this section and discuss the general case of $K(\geq 1)$

Figure 2: Causal Relationships with Two Causally Ordered Mediators where Unobserved Confounding Exists for the Relationship between Mediators $\{M_1, M_2\}$ and outcome Y.



Note: A denotes the treatment, Y denotes the outcome, M_j denotes mediator j. Baseline covariates X are kept implicit.

causally ordered mediators in SI D.

Suppose there exists an unobserved confounder that affects both the mediators (M_1, M_2) and the outcome Y, but not the treatment. Figure 2 shows a causal diagram reflecting the relationships between these variables, where the baseline covariates X are kept implicit. In this case, because no unobserved confounding exists for the treatment-outcome relationship, the ATE is still identified, and their estimates are not subject to confounding bias. We now assess the biases for the PSEs via M_1 and via M_2. Following VanderWeele (2010), we make three simplifying assumptions: (a) U is binary; (b) the average "effect" of U on Y, conditional on baseline covariates X, the treatment A, and the mediator set $\mathcal{M}_k = \{M_1, \ldots M_k\}$ (where $k \in \{1, 2\}$), is constant, which we denote by γ_k; and (c) the difference in the prevalence of U between treated and untreated units, conditional on baseline covariates X and the mediator set $\mathcal{M}_k$ (where $k \in \{1, 2\}$), is constant, which we denote by η_k. Then, as shown in SI D, estimates of the direct and path-specific effects without adjusting for U are subject to the following biases:

$$\text{Bias}[\tau_{A \to Y}] = \gamma_2 \eta_2; \tag{12}$$

$$\text{Bias}[\tau_{A \to M_1 \rightsquigarrow Y}] = -\gamma_1 \eta_1; \tag{13}$$

$$\text{Bias}[\tau_{A \to M_2 \to Y}] = \gamma_1 \eta_1 - \gamma_2 \eta_2. \tag{14}$$

These formulas (12)-(14) allow us to construct a range of bias-adjusted estimates for $\tau_{A \to Y}$, $\tau_{A \to M_1 \rightsquigarrow Y}$, and $\tau_{A \to M_2 \to Y}$ across potential values of (γ_1, γ_2) and (η_1, η_2). In practice, we may focus on estimands that are of particular relevance to the research question. For example, if we are primarily interested in the robustness of the estimated PSE via M_1, i.e, $\hat{\tau}_{A \to M_1 \rightsquigarrow Y}$, we can identify the values of γ_1 and η_1 that would suffice to reduce it to zero. Alternatively, if we are primarily interested in the robustness of the estimated direct effect, we can identify the values of γ_2 and η_2 that would suffice to reduce $\hat{\tau}_{A \to Y}$ to zero. In applications, we can also use observed covariates to suggest plausible values for the sensitivity parameters. For example, if we have an observed binary confounder $Z \in X$, we can fit a linear model of Y on X, A, and $\mathcal{M}_k$, whose coefficient on Z will provide a plausible value of γ_k. In the meantime, we can fit a linear model of Z on A, $\mathcal{M}_k$, and other components of X, whose coefficient on A will provide a plausible value of η_k. By combining these plausible values of γ_k and η_k, we can assess the amount of bias that would result if an unobserved variable "worked exactly like" Z in confounding the mediator-outcome relationships. In the next section, we illustrate these techniques with two empirical examples.

Although the bias formulas (12)-(14) are derived under the assumption that U affects both M_1 and M_2, they are still applicable in the special case where U does not affect M_1. In this case, it can be shown that $\eta_1 = 0$ (see SI D), leading to a simplification of equations (13)-(14): $\text{Bias}[\tau_{A \to M_1 \rightsquigarrow Y}] = 0$ and $\text{Bias}[\tau_{A \to M_2 \to Y}] = -\gamma_2\eta_2$. The former result is expected because when U does not affect M_1, no unobserved confounding exists for the M_1-Y relationship, leading to unbiased estimates of the PSE $\tau_{A \to M_1 \rightsquigarrow Y}$.

A common limitation to sensitivity analysis methods for unobserved confounding is the reliance on simplifying assumptions about the exact form of confounding, which are required for the sake of interpretability (e.g., Imbens 2003). Our proposed method is no exception. First, the unobserved confounder U is assumed to be a pretreatment variable. Thus the bias formulas cannot be used to assess the sensitivity of estimated PSEs to unobserved posttreatment confounders or, for that matter, to mismeasured mediators. For example, in the issue framing study, the bias formulas cannot be used to assess bias due to measurement error when the measured belief and importance variables are noisy

indicators of some true but latent values of beliefs and importance. Second, U is assumed to be binary. Thus the bias formulas do not directly apply to cases where unobserved confounders are known to be continuous or multivariate. Finally, by assuming that both γ_k and η_k are constant, we stipulate that the conditional expectation $\mathbb{E}[Y|X, A, \mathcal{M}_k, U]$ depends on $(X, A, \mathcal{M}_k)$ and U additively, and that the conditional probability $\Pr[U = 1|X, A, \mathcal{M}_k]$ depends on $(X, \mathcal{M}_k)$ and A additively. Given the stringency of these assumptions, the bias formulas (12)-(14) should best be viewed as an approximation of the true biases that would result from unobserved confounding.[3]

Empirical Illustrations

We illustrate the proposed methods for estimation and sensitivity analysis by first reanalyzing Slothuus's (2008) data on issue framing effects. We then revisit an observational study on the multigenerational effects of political violence (Lupu and Peisakhin 2017). In SI G, we demonstrate the utility of our framework with an experimental study by Tomz and Weeks (2013), where the authors have attempted to isolate the mediating effect of morality in the democratic peace.

Issue Framing Effects

Using a survey experiment on a sample of Danish students, Slothuus (2008) found that individuals are substantially more supportive of a proposed welfare reform if they are exposed to a newspaper article that highlights its positive effect on job creation (the job frame) rather than one emphasizing its negative effect on the poor (the poor frame). To analyze the causal mechanisms underlying this effect, the author used a series of five-point-scale questions to tap (a) the respondents' beliefs about why some people receive welfare benefits (the belief mediator) and (b) their perceived importance of five competing considerations directly related to welfare policy (the importance mediator). The author then conducted a mediation analysis under the assumption that the belief mediator and the impor-

[3] In SI E, we conduct a simulation study to investigate the performance of this approximation under plausible scenarios. The results suggest that the approximation is excellent under these scenarios.

tance mediator are causally independent. However, as noted previously, respondents' beliefs about welfare recipients likely influence their perceived importance of competing issue-related considerations. In the following analysis, we allow the two mediators to be causally dependent. Following the literature (Imai and Yamamoto 2013; Miller 2007), we treat respondents' beliefs about the issue as causally prior to their perceived importance of competing considerations. Under this assumption, the pathways that transmit the framing effect can be represented by a DAG akin to the top panel of Figure 1.

In this DAG, the outcome, Y, is a measure of support for the proposed welfare reform on a seven-point scale; treatment, A, denotes whether the respondent receives the job frame rather than the poor frame; the mediator M_1 includes measures of the respondent's beliefs about why some people receive welfare benefits, or who is responsible for those people's situation; the mediator M_2 includes the respondent's ratings on the importance of five competing considerations related to welfare policy; finally, the pretreatment covariates X include measures of gender, education, political interest, ideology, political knowledge, and extremity of political values.[4] We control for a set of pretreatment covariates because, although treatment is randomly assigned, the mediator-outcome relationships may still be confounded by the respondent's baseline characteristics.

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 average support for the proposed welfare reform (measured on a seven-point scale) is 4.3 among respondents exposed to the job frame and 3.16 among those exposed to the poor frame. The total effect of treatment, therefore, is about 1.14.

We estimate the PSEs for the paths $A \rightarrow Y$, $A \rightarrow M_1 \rightsquigarrow Y$, and $A \rightarrow M_2 \rightarrow Y$ using the imputation approach described earlier. To allow for nonlinear and interaction effects, we use BART to fit the outcome models conditional on treatment, the pretreatment covariates, and varying sets of mediators (namely, $\{M_1, M_2\}$ and $\{M_1\}$). The results are shown in Table 2. The estimated PSE via the belief mediator ($A \rightarrow M_1 \rightsquigarrow Y$) is 0.24 (95% CI: [-0.02, 0.52]), suggesting that the respondent's

[4]Detailed definitions of these variables are given in Slothuus (2008).

Table 2: Estimates of Total and Path-Specific Effects of Issue Framing on Policy Support.

	Estimate
Average total effect (ATE)	1.15 [0.61, 1.65]
Through the belief mediator ($A \rightarrow M_1 \rightsquigarrow Y$)	0.24 [-0.02, 0.52]
Through the importance mediator ($A \rightarrow M_2 \rightarrow Y$)	0.18 [0.01, 0.36]
Direct effect ($A \rightarrow Y$)	0.72 [0.32, 1.08]

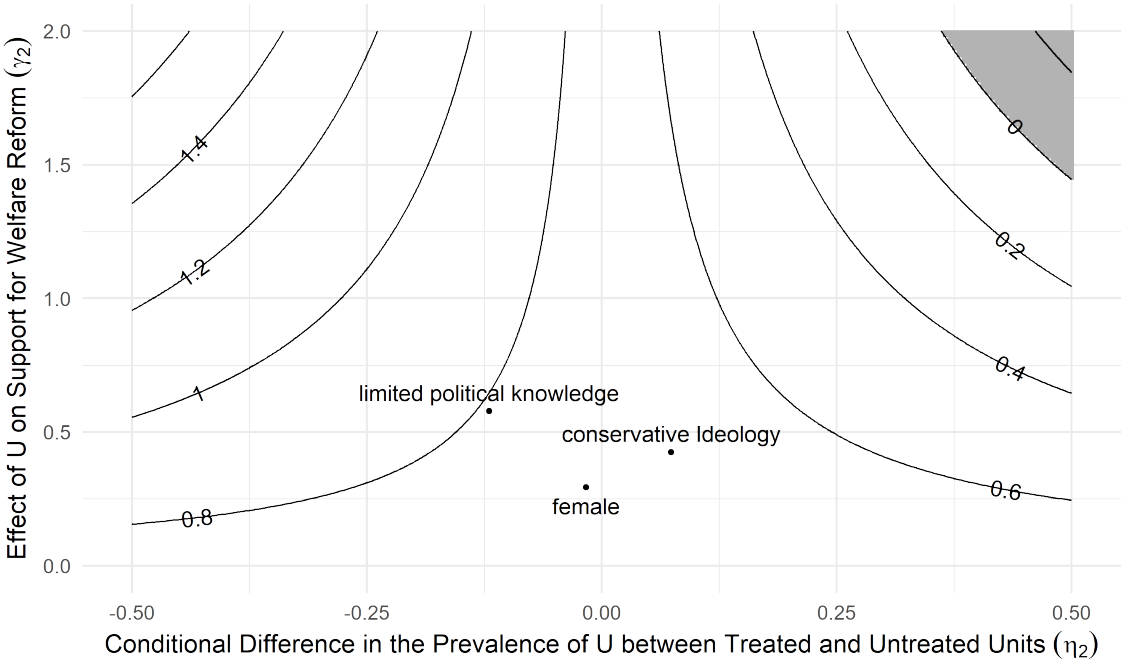
Note: Numbers in brackets represent 95% bootstrapped confidence intervals (1,000 iterations).

beliefs about the causes of the situation of welfare recipients have a relatively minor and statistically insignificant mediating effect. The estimated PSE via the importance mediator ($A \rightarrow M_2 \rightarrow Y$) is 0.18 (95% CI: [0.01, 0.36]), suggesting that the perceived importance of competing considerations plays an independent, albeit small, role in transmitting the effect of issue framing on policy support. Finally, we find that over half of the total effect appears to be "direct," i.e., operating neither through the belief mediator nor through the importance mediator.

We now conduct a sensitivity analysis for the direct effect of issue framing on policy support. Suppose there exists a binary unobserved confounder U that affects respondents' beliefs about the issue, perceived importance of issue-related considerations, as well as their support for welfare reform. Equation (12) indicates that in this scenario, the estimated direct effect is subject to a bias of $\gamma_2\eta_2$, where γ_2 denotes the average effect of U on policy support (Y) conditional on treatment (A), the belief and importance mediators (M_1 and M_2), and the baseline covariates (X), and η_2 denotes the difference in the prevalence of U between treated and untreated units conditional on the belief and importance mediators (M_1 and M_2) and the baseline covariates (X).

To obtain some intuition as to the signs of γ_2 and η_2, let us consider U as a dummy variable indicating middle- or upper-class background, which might lead to stronger support for the proposed welfare reform, i.e., $\gamma_2 > 0$. Since treatment is randomly assigned in this study, the prevalence of U should be similar between treated and untreated units. However, because both middle/upper-class background (U) and the job frame (A) are supposed to affect beliefs about the issue (M_1) and perceived importance of competing considerations (M_2), the conditional association between A and U given

Figure 3: Bias-adjusted Estimates of the Direct Effect of Issue Framing on Policy Support.



Note: The contours represent the bias-adjusted estimates of the direct effect ($\tau_{A\to Y}$) plotted as a function of γ_2 and η_2. The grey area shows the values of γ_2 and η_2 that would reverse the sign of the estimated $\tau_{A\to Y}$. The annotated points represent the γ_2 and η_2 values that would result if the unobserved variable U "worked exactly like" one of the observed covariates in its confounding effect on the mediator-outcome relationships.

M_1, M_2, and X can deviate from zero. Specifically, because M_1 and M_2 are both colliders of A and U, the conditional association between A and U might be negative — especially if the effects of U and A on the mediators are in the same direction. In this scenario, the bias $\gamma_2\eta_2$ would be negative, implying an *underestimate* of the direct effect. From this perspective, our finding that most of the framing effect does not operate through the belief mediator or the importance mediator appears robust.

We can also use observed binary covariates to obtain a range of plausible values for the sensitivity parameters γ_2 and η_2. Here, we consider three such variables — gender, right-wing ideology, and limited political knowledge, where right-wing ideology and limited political knowledge are dummy variables obtained by dichotomizing the original measures of ideology and political knowledge at their medians. We then use the procedures described in the preceding section to compute the values of γ_2 and η_2 that would result if the unobserved variable U "worked exactly like" each of these covariates in its confounding effect. Figure 3 shows the contours of bias-adjusted estimates of the direct effect

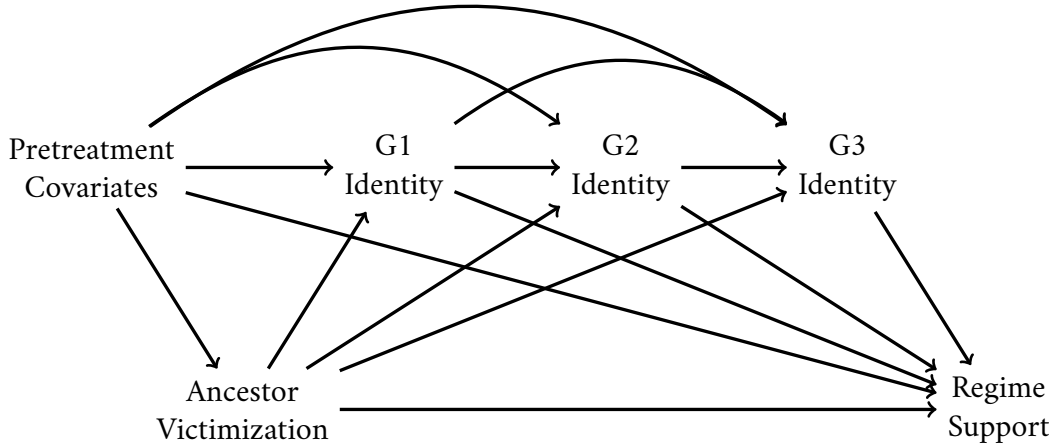
at different values of γ_2 and η_2, as well as those corresponding to an unobserved variable that mimics gender, right-wing ideology, and limited political knowledge in its confounding effect. We can see that the original estimate (0.72) can be explained away by unobserved confounding only when both γ_2 and η_2 are positive and much larger than their plausible values suggested by these observed covariates.

The Legacy of Political Violence

We now illustrate the imputation approach for tracing causal paths from observational data. 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, these authors 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 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" (Lupu and Peisakhin 2017, 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 intergenerational pathways that transmit the legacy of political violence, the authors 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 is potentially problematic, however, because descendants' political identities are likely shaped by the political identities of their parents and grandparents, which might also have a direct effect on descendant political attitudes and behavior. In other words, the identities of first- and second-generation respondents are posttreatment confounders of the mediator-outcome relationship, i.e., the relationship between descendants' identities and their political attitudes and behavior, implying that the ACME via descendants' political identities cannot be nonparametrically identified.

Figure 4: Causal Pathways from Ancestor Victimization to Descendants' Regime Support.



In contrast to the authors' mediation analyses that focused on the political identity of the descendant as the only mediator, we treat the political identities of first-, second-, and third-generation respondents as three causally ordered 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 4. 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, and mediator-outcome relationships.

We then estimate the PSEs as defined by equation (9), 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

Table 3: Estimates of Total and Path-Specific Effects of Ancestor Victimization on Support for Russia's Annexation of Crimea.

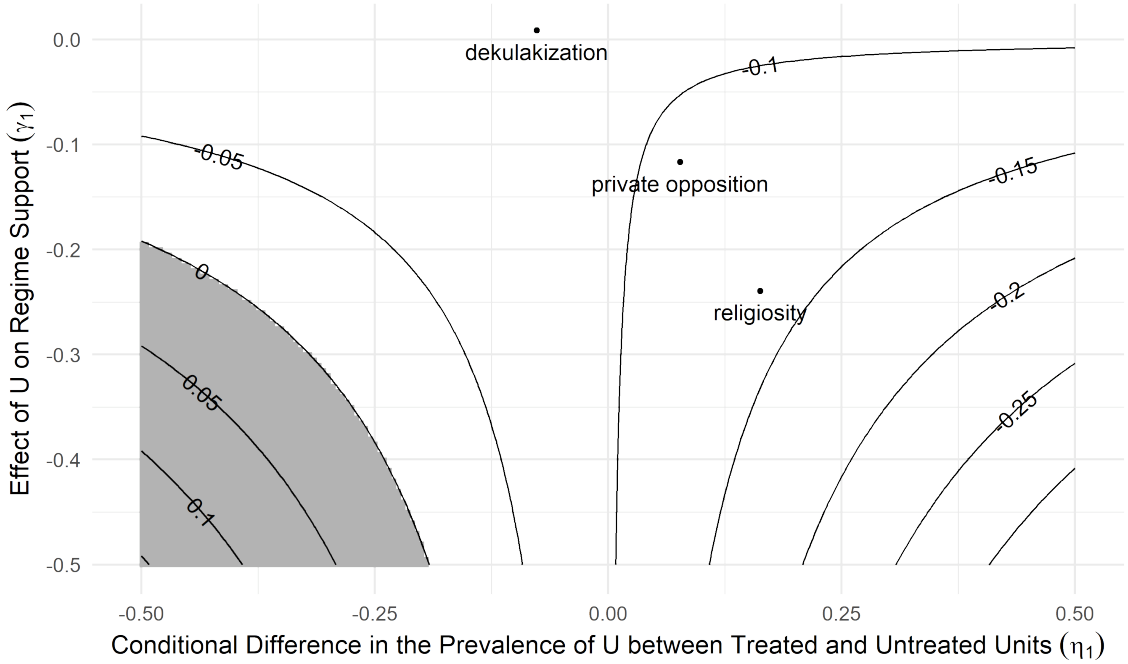
	Pure imputation estimator	Imputation-based weighting estimator
Average total effect (ATE)	-0.20 [-0.30, -0.11]	-0.20 [-0.30, -0.11]
Through G1 identity ($A \rightarrow M_1 \rightsquigarrow Y$)	-0.10 [-0.15, -0.06]	-0.12 [-0.18, -0.07]
Through G2 identity ($A \rightarrow M_2 \rightsquigarrow Y$)	-0.02 [-0.06, 0.02]	-0.02 [-0.06, 0.03]
Through G3 identity ($A \rightarrow M_3 \rightarrow Y$)	-0.03 [-0.07, 0.00]	-0.03 [-0.07, 0.01]
Direct effect ($A \rightarrow Y$)	-0.05 [-0.12, 0.03]	-0.04 [-0.12, 0.05]

Note: Numbers in brackets represent 95% bootstrapped confidence intervals (1,000 iterations).

imputation-based weighting estimator, we estimate all outcome models using BART and estimate the propensity score model using gradient boosting machines that are calibrated to maximize covariate balance (McCaffrey et al. 2004; Ridgeway et al. 2017). The results, as shown in Table 3, are similar between the two estimators. 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 (from 0.64 to 0.44) on the probability scale. By the pure imputation estimator, the direct effect is only about -0.05, meaning that most of the total effect operates through the political identities of first-, second-, and third-generation respondents. The bulk of the indirect effect appears to be transmitted through the political identities of grandparents ("via G1 identity"), rather than through the political identities of second- and third-generation respondents directly ("via G2 identity" and "via G3 identity"). This finding suggests 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. This is a key theoretical hypothesis of Lupu and Peisakhin (2017). However, it was not tested in the authors' implicit mediation analysis, which considered only the role of the descendant's political identity (G3 identity).

To assess the robustness of the above finding to unobserved confounding of the mediator-outcome relationships, we apply the bias formulas introduced in the preceding section for the PSE via G1 identity ($\tau_{A \rightarrow M_1 \rightsquigarrow Y}$). Suppose there exists a binary unobserved confounder U (e.g., presence

Figure 5: Bias-adjusted Estimates of the Path-Specific Effect of Ancestor Victimization on Regime Support via G1 Identity.



Note: The contours represent the bias-adjusted estimates of the PSE via G1 identity ($\tau_{A \to M_1 \rightsquigarrow Y}$) plotted as a function of γ_1 and η_1 (with the unadjusted estimate computed from the pure imputation estimator). See the note for Figure 3 for the interpretation of other elements of the graph.

of some personality trait in the first-generation respondent) that affects both the political identities of first-, second-, and third-generation respondents and regime support among the grandchildren. Equation (13) indicates that in this scenario, the estimated PSE via G1 identity suffers a bias of $-\gamma_1\eta_1$, where γ_1 denotes the effect of U on regime support (Y) conditional on ancestor victimization (A), G1 identity (M_1), and the baseline covariates (X), and η_1 denotes the difference in the prevalence of U between treated and untreated units conditional on G1 identity (M_1) and the baseline covariates (X). To be more concrete, let us consider U as a personality trait of the G1 respondent that facilitates in-group solidarity, which would suggest a negative effect of U on regime support, i.e., $\gamma_1 < 0$. The sign of η_1 is less clear. If both violent victimization (A) and the unobserved personality trait (U) had had a positive effect on G1 identity (M_1), the association between A and U conditional on M_1, a collider between A and U, might be negative. In this case, $-\gamma_1\eta_1$ will be negative, suggesting an overestimate of the (negative) PSE via G1 identity.

Figure 5 shows the contours of bias-adjusted estimates of the PSE via G1 identity at different values of γ_1 and η_1. In addition, it shows the values of the γ_1 and η_1 that would result if the unobserved variable U "worked exactly like" one of three observed binary covariates: whether the G1 respondent had close relatives subject to dekulakization (*dekulakization*), whether the G1 respondent's close relatives privately opposed Soviet authorities (*private opposition*), whether the G1 respondent's family considered it very important to follow Islamic customs and traditions while in deportation (*religiosity*). We can see that the original estimate (-0.1) is quite robust, as it can be attributed entirely to unobserved confounding only when both γ_1 and η_1 are sizable (e.g., when $\gamma_1 = \eta_1 = -0.32$) and far from their plausible values suggested by these observed covariates.

Concluding Remarks

Despite a growing interest in the study of causal mechanisms in political science, conventional methods for causal mediation analysis are difficult to use when the causal effect of interest operates through multiple causally dependent mediators. In particular, the ACME cannot be nonparametrically identified if the mediator-outcome relationship is confounded by posttreatment variables, even if these variables are observed. In this article, we introduced a general framework for tracing causal paths with multiple mediators. In this framework, the total effect of a treatment on an outcome is decomposed into a set of path-specific effects (PSEs). These PSEs, unlike the ACMEs of individual mediators, are nonparametrically identified under a set of unconfoundedness assumptions.

We then described an imputation approach for estimating these PSEs from experimental and observational data. In contrast to conventional methods for analyzing causal mediation, 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 data-adaptive methods such as GBM and BART. Therefore, minimal modeling assumptions are needed to implement this approach, and dif-

ferent models of the expected outcome can be used to check the robustness of results. In SI F, we illustrate this point by showing that for our two empirical examples, estimates of the PSEs are similar whether we use GLM, GBM, or BART to fit the outcome models.

The identification of PSEs is premised on a set of potentially strong assumptions, which require that all relevant confounders of the treatment-outcome, treatment-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-outcome relationships can still be confounded by unobserved factors. As we have shown, in cases where some of these assumptions are questionable, a set of general-purpose bias formulas can be used to assess the robustness of conclusions. To facilitate implementation, we offer an open-source R package, `paths`, for implementing the proposed methods for estimation and sensitivity analysis, which is available from Github and CRAN. In addition, in SI H, we provide R code illustrating the use of `paths` with our empirical examples.

References

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

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016b. "The Political Legacy of American Slavery." *The Journal of Politics* 78:621–641.

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

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

Avin, Chen, Ilya Shpitser, and Judea Pearl. 2005. "Identifiability of Path-specific Effects." In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 357–363. Morgan Kaufmann Publishers Inc.

Baron, Reuben M and David A Kenny. 1986. "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51:1173.

Bullock, John G, Donald P Green, and Shang E Ha. 2010. "Yes, but what's the mechanism?(don't expect an easy answer)." *Journal of personality and social psychology* 98:550.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21:C1–C68.

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

Chong, Dennis and James N Druckman. 2007. "Framing Theory." *Annual Review of Political Science* 10:103–126.

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

Daniel, RM, BL De Stavola, SN Cousens, and Stijn Vansteelandt. 2015. "Causal Mediation Analysis with Multiple Mediators." *Biometrics* 71:1–14.

Druckman, James N and Kjersten R Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47:729–745.

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

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

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* 105:765–789.

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

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21:141–171.

Imbens, Guido W. 2003. "Sensitivity to exogeneity assumptions in program evaluation." *American Economic Review* 93:126–132.

Kang, Joseph DY and Joseph L Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22:523–539.

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

Mazumder, Soumyajit. 2018. "The Persistent Effect of US Civil Rights Protests on Political Attitudes." *American Journal of Political Science* 62:922–935.

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* 9:403.

Miles, Caleb H, Ilya Shpitser, Phyllis Kanki, Seema Meloni, and Eric J Tchetgen Tchetgen. 2017. "Quantifying an Adherence Path-Specific Effect of Antiretroviral Therapy in the Nigeria Pepfar Program." *Journal of the American Statistical Association* 112:1443–1452.

Miles, Caleb H, Ilya Shpitser, Phyllis Kanki, Seema Meloni, and Eric J Tchetgen Tchetgen. 2020. "On Semiparametric Estimation of a Path-Specific Effect in the Presence of Mediator-Outcome Confounding." *Biometrika* 107:159–172.

Miller, Joanne M. 2007. "Examining the Mediators of Agenda Setting: A New Experimental Paradigm Reveals the Role of Emotions." *Political Psychology* 28:689–717.

Nelson, Thomas E, Rosalee A Clawson, and Zoe M Oxley. 1997a. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91:567–583.

Nelson, Thomas E and Zoe M Oxley. 1999. "Issue Framing Effects on Belief Importance and Opinion." *The Journal of Politics* 61:1040–1067.

Nelson, Thomas E, Zoe M Oxley, and Rosalee A Clawson. 1997b. "Toward a Psychology of Framing Effects." *Political Behavior* 19:221–246.

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

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

Richardson, Thomas S and James M Robins. 2013. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality." *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128:2013.

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.

Slothuus, Rune. 2008. "More than Weighting Cognitive Importance: A Dual-process Model of Issue Framing Effects." *Political Psychology* 29:1–28.

Tchetgen Tchetgen, Eric J and Ilya Shpitser. 2012. "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness, and Sensitivity Analysis." *Annals of Statistics* 40:1816.

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

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

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

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

Zheng, Wenjing and Mark J van der Laan. 2011. "Cross-validated Targeted Minimum-loss-based Estimation." In *Targeted Learning*, pp. 459–474. New York, NY: Springer.

Zhou, Xiang. 2022. "Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

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

Zhou, Xiang and Geoffrey T. Wodtke. 2020. "Residual Balancing: A Method of Constructing Weights for Marginal Structural Models." *Political Analysis* 28:487–506.

Biographical Statements

Xiang Zhou is an Associate Professor of Sociology at Harvard University, Cambridge, MA 02138. Teppei Yamamoto is an Associate Professor of Political Science at Massachusetts Institute of Technology (MIT), Cambridge, MA 02139.

Supporting Information

A Proof of Equation (2)

We interpret a DAG as representing a nonparametric structural equation model with mutually independent errors (Pearl 2009). Thus the DAG in the top panel of Figure 1 corresponds to the following nonparametric structural equations:

$$
\begin{aligned}
X &= f_X(\epsilon_X), \\
A &= f_A(X, \epsilon_A), \\
M_1 &= f_{M_1}(X, A, \epsilon_{M_1}), \\
M_2 &= f_{M_2}(X, A, M_1, \epsilon_{M_2}), \\
Y &= f_Y(X, A, M_1, M_2, \epsilon_Y),
\end{aligned}
$$

where the error terms ϵ_X, ϵ_A, ϵ_{M_1}, ϵ_{M_2}, and ϵ_Y are mutually independent but otherwise arbitrarily distributed. The potential outcomes defined in the main text can thus be written as

$$
\begin{aligned}
M_1(a) &= f_{M_1}(X, a, \epsilon_{M_1}), \\
M_2(a, m_1) &= f_{M_2}(X, a, m_1, \epsilon_{M_2}), \\
Y(a, m_1, m_2) &= f_Y(X, a, m_1, m_2, \epsilon_Y).
\end{aligned}
$$

From the above equations and the mutual independence of the error terms, we have the following conditional independence relationships: (a) $M_1(a_1) \perp\!\!\!\perp M_2(a_2, m_1)|X$; (b) $Y(a, m_1, m_2) \perp\!\!\!\perp (M_1(a_1), M_2(a_2, m_1))|X$; (c) $M_1(a) \perp\!\!\!\perp A|X$; (d) $M_2(a, m_1) \perp\!\!\!\perp (A, M_1)|X$; (e) $Y(a, m_1, m_2) \perp\!\!\!\perp$

$(A, M_1, M_2)|X$. Thus

$$
\begin{aligned}
&\mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_1)))|X = x] \\
&= \int \mathbb{E}[Y(a, m_1, M_2(a_2, M_1(a_1)))|X = x, M_1(a_1) = m_1] f_{M_1(a_1)|X=x}(m_1) dm_1 \\
&= \iint \mathbb{E}[Y(a, m_1, m_2)|X = x, M_1(a_1) = m_1, M_2(a_2, m_1) = m_2] \\
&\quad f_{M_1(a_1)|X=x}(m_1) f_{M_2(a_2,m_1)|X=x}(m_2) dm_1 dm_2 \quad \text{by (a)} \\
&= \iint \mathbb{E}[Y(a, m_1, m_2)|X = x] f_{M_1(a_1)|X=x}(m_1) f_{M_2(a_2,m_1)|X=x}(m_2) dm_1 dm_2 \quad \text{by (b)} \\
&= \iint \mathbb{E}[Y(a, m_1, m_2)|X = x] f_{M_1(a_1)|X=x,A=a_1}(m_1) f_{M_2(a_2,m_1)|X=x,A=a_2,M_1=m_1}(m_2) dm_1 dm_2 \quad \text{by (c) and (d)} \\
&= \iint \mathbb{E}[Y(a, m_1, m_2)|X = x, A = a, M_1 = m_1, M_2 = m_2] f(m_1|x, a_1) f(m_2|x, a_2, m_1) dm_1 dm_2 \quad \text{by (e)} \\
&= \iint \mathbb{E}[Y|x, a, m_1, m_2] f(m_1|x, a_1) f(m_2|x, a_2, m_1) dm_1 dm_2 \qquad (15)
\end{aligned}
$$

Marginalizing the above expression over $f(x)$ yields equation (2).

B Proofs of Equations (5)-(8)

Let us first consider equations (5) and (6). By equation (15), we have

$$
\begin{aligned}
&\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)|X = x] \\
=&\iint \mathbb{E}[Y|x, A = 1, m_1, m_2] f(m_1|x, A = 0) f(m_2|x, A = 0, m_1) dm_1 dm_2 \\
=&\iint \mathbb{E}[Y|x, A = 1, m_1, m_2] f(m_1, m_2|x, A = 0) dm_1 dm_2 \\
=&\mathbb{E}\big[\mathbb{E}[Y|x, A = 1, M_1, M_2]|x, A = 0\big].
\end{aligned}
$$

Marginalizing the above expression over $f(x)$ yields equation (5). Similarly,

$$
\begin{aligned}
&\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)|X = x] \\
=&\iint \mathbb{E}[Y|x, A = 1, m_1, m_2] f(m_1|x, A = 0) f(m_2|x, A = 1, m_1) dm_1 dm_2 \\
=&\int \mathbb{E}[Y|x, A = 1, m_1] f(m_1|x, A = 0) dm_1 \\
=&\mathbb{E}\big[\mathbb{E}[Y|x, A = 1, M_1]|x, A = 0\big].
\end{aligned}
$$

Here, the second equality derives from the fact that $\int \mathbb{E}[Y|x, A = 1, m_1, m_2] f(m_2|x, A = 1, m_1) dm_2 = \mathbb{E}[Y|x, A = 1, m_1]$. Marginalizing the above expression over $f(x)$ yields equation

(6). Now, consider equations (7) and (8). By the mediation formula (2), we have

$$
\begin{aligned}
&\mathbb{E}[Y\big(1, M_1(0), M_2(0, M_1(0))\big)] \\
=&\iiint \mathbb{E}[Y|x, A=1, m_1, m_2] f(m_1|x, A=0) f(m_2|x, A=0, m_1) f(x) dm_1 dm_2 dx \\
=&\iiint \mathbb{E}[Y|x, A=1, m_1, m_2] f(m_1, m_2|x, A=0) f(x) dm_1 dm_2 dx \\
=&\iiint \mathbb{E}[Y|x, A=1, m_1, m_2] f(m_1, m_2, x|A=0) \frac{f(x)}{f(x|A=0)} dm_1 dm_2 dx \\
=&\iiint \mathbb{E}[Y|x, A=1, m_1, m_2] f(m_1, m_2, x|A=0) \frac{f(x)}{\frac{\Pr[A=0|X=x]f(x)}{\Pr[A=0]}} dm_1 dm_2 dx \quad \text{(via Bayes' rule)} \\
=&\iiint \mathbb{E}[Y|x, A=1, m_1, m_2] f(m_1, m_2, x|A=0) \frac{\Pr[A=0]}{\Pr[A=0|X=x]} dm_1 dm_2 dx \\
=&\mathbb{E}\big[\mathbb{E}[Y|X, A=1, M_1, M_2] \frac{\Pr[A=0]}{\Pr[A=0|X]} | A=0\big].
\end{aligned}
$$

Here, the 4th line is due to the fact that $f(m_1, m_2|x, A=0) = f(m_1, m_2, x|A=0)/f(x|A=0)$, and the 5th line is due to the fact that $f(x|A=0) = \frac{\Pr[A=0|X=x]f(x)}{\Pr[A=0]}$ (Bayes' rule). Similarly,

$$
\begin{aligned}
&\mathbb{E}[Y\big(1, M_1(0), M_2(1, M_1(0))\big)] \\
=&\iiint \mathbb{E}[Y|x, A=1, m_1, m_2] f(m_1|x, A=0) f(m_2|x, A=1, m_1) f(x) dm_1 dm_2 dx \\
=&\iint \mathbb{E}[Y|x, A=1, m_1] f(m_1|x, A=0) f(x) dm_1 dx \\
=&\iint \mathbb{E}[Y|x, A=1, m_1] f(m_1, x|A=0) \frac{f(x)}{f(x|A=0)} dm_1 dx \\
=&\iint \mathbb{E}[Y|x, A=1, m_1] f(m_1, x|A=0) \frac{f(x)}{\frac{\Pr[A=0|X=x]f(x)}{\Pr[A=0]}} dm_1 dx \quad \text{(via Bayes' rule)} \\
=&\iint \mathbb{E}[Y|x, A=1, m_1] f(m_1, x|A=0) \frac{\Pr[A=0]}{\Pr[A=0|X=x]} dm_1 dx \\
=&\mathbb{E}\big[\mathbb{E}[Y|X, A=1, M_1] \frac{\Pr[A=0]}{\Pr[A=0|X]} | A=0\big].
\end{aligned}
$$

C Algorithms for Implementing the Imputation Approach with K Causally Ordered Mediators

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 $\hat{\mathbb{E}}[Y|X, A = 0]$ and $\hat{\mathbb{E}}[Y|X, A = 1]$ 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 untreated 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 the 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 PSEs as defined in equation (9).

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 $\Pr[A = 0]$ using its sample analog and $\Pr[A = 0|X]$ using a propensity score model for the treatment.

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

(a) Fit an outcome model conditional on the treatment A, the mediators $\mathcal{M}_k$, and the pre-treatment confounders X. For the untreated 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 $\widehat{\Pr}[A = 0]/\widehat{\Pr}[A = 0|X]$.

3. Calculate the PSEs as defined in equation (9).

In experimental studies, step (1) can be simplified because $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ can be estimated using simple averages of the observed outcome within the control and treatment groups. In the meantime, the inverse-probability weights in step 2(b) are 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.

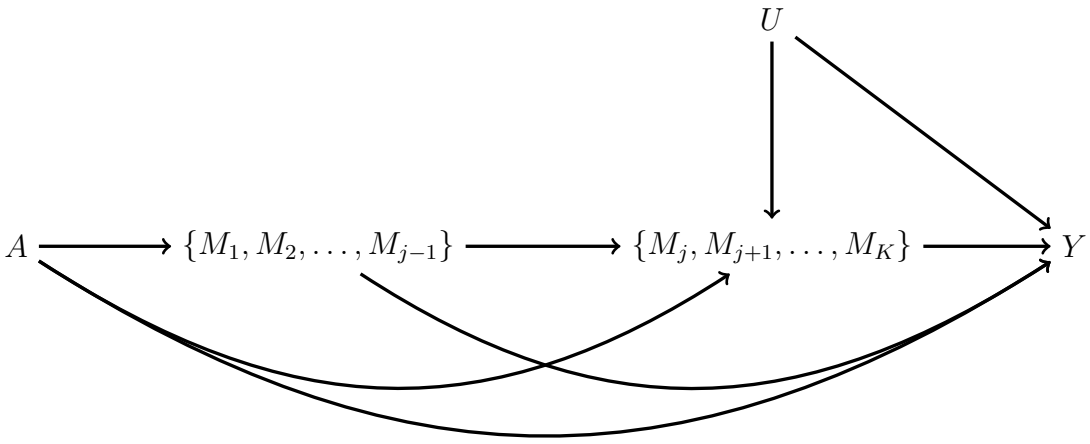


Figure D1: Causal Relationships with K Causally Ordered Mediators where Unobserved Confounding Exists for the Relationship between Mediators $\{M_j, \ldots M_K\}$ and outcome Y.

Note: A denotes the treatment, Y denotes the outcome, M_j denotes mediator j. Baseline Covariates X are kept implicit.

D Sensitivity Analysis with $K(\geq 1)$ Causally Ordered Mediators

Suppose that the treatment effect operates through K causally ordered mediators $M_1, M_2, \ldots M_K$, and that there exists an unobserved confounder that affects both the outcome Y and the mediators $\{M_j, M_{j+1}, \ldots M_K\}$, but not mediators $\{M_1, M_2, \ldots M_{j-1}\}$. Figure D1 shows a DAG reflecting the relationships between these variables, where the baseline covariates X are kept implicit. In this case, because no unobserved confounding exists for any of the mediators preceding M_j, the PSEs via $M_1, M_2, \ldots M_{j-1}$ are still identified, and their imputation-based estimates are not subject to confounding bias. We now assess the biases for the PSEs via $M_j, M_{j+1}, \ldots M_K$. As in the case of two mediators, we make three simplifying assumptions : (a) U is binary; (b) the average "effect" of U on Y, conditional on baseline covariates X, treatment A, and mediator set $\mathcal{M}_k = \{M_1, M_2, \ldots M_k\}$, is constant; and (c) the difference in the prevalence of U between treated and untreated units, conditional on baseline covariates X and the mediator set $\mathcal{M}_k$, is constant. Denote the average effect of U on Y as γ_k and the conditional difference in the prevalence of U between treated and untreated units

as η_k:

$$\gamma_k = \mathbb{E}[Y|X, A, \mathcal{M}_k, U = 1] - \mathbb{E}[Y|X, A, \mathcal{M}_k, U = 0];$$
$$\eta_k = \Pr[U = 1|X, A = 1, \mathcal{M}_k] - \Pr[U = 1|X, A = 0, \mathcal{M}_k].$$

Then estimates of the direct and path-specific effects without adjusting for U are subject to the following biases:

$$\text{Bias}[\tau_{A \to Y}(a^*)] = \gamma_K \eta_K; \tag{16}$$
$$\text{Bias}[\tau_{A \to M_j \rightsquigarrow Y}(a)] = -\gamma_j \eta_j; \tag{17}$$
$$\text{Bias}[\tau_{A \to M_k \rightsquigarrow Y}(a)] = \gamma_{k-1}\eta_{k-1} - \gamma_k \eta_k, \quad \text{for any } k > j. \tag{18}$$

These formulas can be seen as a generalization of the bias formulas presented in the main text. As in the case of two mediators, the simplicity of these bias formulas rests on a set of strong and restrictive assumptions, such as the absence of interaction effects implied by the constancy of γ_k and η_k.

D.1 Proof of Bias Formulas (16)-(18)

Consider the DAG shown in Figure D1, where the baseline covariates X, which may affect any of the variables in $\{A, U, M_1 \dots M_K, Y\}$, are kept implicit. To see how the PSEs are connected to the average direct effect (ADE) and average causal mediation effect (ACME), let us consider $\mathcal{M}_k = \{M_1 \dots M_k\}$ as a whole, where $k \in \{1, \dots K\}$. The ADE and ACME for $\mathcal{M}_k$ are

$$\text{ADE}_k(0) = \tau_{A \to Y} + \sum_{l=k+1}^{K} \tau_{A \to M_l \rightsquigarrow Y};$$
$$\text{ACME}_k(1) = \sum_{l=1}^{k} \tau_{A \to M_l \rightsquigarrow Y}.$$

Hence the PSEs can be written as

$$\tau_{A \to Y} = \text{ADE}_K(0);$$
$$\tau_{A \to M_k \rightsquigarrow Y} = \text{ACME}_k(1) - \text{ACME}_{k-1}(1).$$

Under the three simplifying assumptions outlined in the main text, VanderWeele (2010) shows that estimates of the ADE and ACME without adjusting for U are biased by $\gamma_k\eta_k$ and by $-\gamma_k\eta_k$, respectively, where

$$\gamma_k = \mathbb{E}[Y|X, A, \mathcal{M}_k, U = 1] - \mathbb{E}[Y|X, A, \mathcal{M}_k, U = 0];$$
$$\eta_k = \Pr[U = 1|X, A = 1, \mathcal{M}_k] - \Pr[U = 1|X, A = 0, \mathcal{M}_k].$$

Thus the bias factors for the PSEs can be written as

$$\text{Bias}[\tau_{A \to Y}] = \gamma_K\eta_K; \tag{19}$$
$$\text{Bias}[\tau_{A \to M_k \rightsquigarrow Y}] = \gamma_{k-1}\eta_{k-1} - \gamma_k\eta_k. \tag{20}$$

Because the DAG in Figure 2 encodes a nonparametric structural equation model with independent errors, it implies $A \perp\!\!\!\perp U|X, \mathcal{M}_k$ for any $k < j$. Thus we have

$$\begin{aligned}\eta_k &= \Pr[U = 1|X, A = 1, \mathcal{M}_k] - \Pr[U = 1|X, A = 0, \mathcal{M}_k] \\ &= \Pr[U = 1|X, \mathcal{M}_k] - \Pr[U = 1|X, \mathcal{M}_k] \\ &= 0. \end{aligned} \tag{21}$$

It follows from equations (20-21) that

$$\text{Bias}[\tau_{A \to M_k \rightsquigarrow Y}] = 0, \quad \text{for any } k < j.$$

E A Simulation Study on Bias Formulas (12)-(14)

In this section, we conduct a simulation study to assess the performance of the bias formulas (12)-(14) when some of the underlying assumptions fail to hold. We consider a binary treatment A, a continuous outcome Y, two causally ordered mediators M_1 and M_2, and four pretreatment covariates X_1, X_2, X_3, X_4 generated from the following model:

$$
\begin{aligned}
(U_1, U_2, U_3, U_{XY}) &\sim N(0, I_4) \\
U_{M_1M_2Y} &\sim \text{Bernoulli}\big(0.5\big) \\
X_j &\sim N((U_1, U_2, U_3, U_{XY})\beta_{X_j}, 1) \quad j = 1, 2, 3, 4 \\
A &\sim \text{Bernoulli}\big(\text{logit}^{-1}[(1, X_1, X_2, X_3, X_4)\beta_A]\big) \\
M_1 &\sim N\big((1, X_1, X_2, X_3, X_4, A)\beta_{M_1} + U_{M_1M_2Y}\alpha_{M_1}, 1\big) \\
M_2 &\sim N\big((1, X_1, X_2, X_3, X_4, A, M_1)\beta_{M_2} + U_{M_1M_2Y}\alpha_{M_2}, 1\big) \\
Y &\sim N\big((1, U_{XY}, X_1, X_2, X_3, X_4, A, M_1, M_2)\beta_Y + U_{M_1M_2Y}\alpha_Y, 1\big).
\end{aligned}
$$

The coefficients $\beta_{X_j} (1 \leq j \leq 4)$ and β_Y are drawn from Uniform$[-1, 1]$, the coefficients β_A are drawn from Uniform$[-0.5, 0.5]$, and the coefficients β_{M_1} and β_{M_2} are drawn from Uniform$[0, 0.5]$.

Specifically,

$$\begin{aligned}
\beta_{X_1} &= (0.77, -0.86, 0.35, 0.88) \\
\beta_{X_2} &= (-0.99, -0.72, -0.1, 0.54) \\
\beta_{X_3} &= (-0.74, 0.1, 0.91, 0.46) \\
\beta_{X_4} &= (-0.21, -0.43, -0.21, -0.7) \\
\beta_{A} &= (-0.36, -0.08, -0.06, 0.4, -0.14) \\
\beta_{M_1} &= (0, 0.3, 0.42, 0.48, 0.28, 0.41) \\
\beta_{M_2} &= (0.04, 0.2, 0.09, 0.12, 0.39, 0.34, 0.24) \\
\beta_{Y} &= (-0.27, -0.1, 0.25, 0.2, -0.08, 0.78, 0.76, -0.4, 0.96).
\end{aligned}$$

The unobserved variable U_{XY} confounds only the X-Y relationship and thus does not pose an identification threat. However, the unobserved variable $U_{M_1M_2Y}$ confounds the M_1-Y and M_2-Y relationships, which can lead to biased estimates of the corresponding PSEs. These confounding effects are governed by the α_{M_1}, α_{M_2}, and α_Y parameters.

In the above setup, we can show that the parameters γ_1 and γ_2 are indeed constant, but the parameters η_1 and η_2 are not, because the conditional probability $\Pr[U = 1|X, A = 1, \mathcal{M}_k]$ corresponds to a logit, rather than linear, function of X, A, and $\mathcal{M}_k$. Thus, we expect the bias formulas (12)-(14) to provide only an approximation of the true biases. To investigate the quality of this approximation at varying degrees of confounding, we draw 100 triplets of $(\alpha_{M_1}, \alpha_{M_2}, \alpha_Y)$ from independent uniform distributions over $[-1, 1]$, yielding 100 data-generating processes (DGP). Then, for each DGP, we generate 1,000 Monte Carlo samples of size 2,000, and, for each sample, obtain pure imputation estimates of the PSEs $\tau_{A\to M_1 \rightsquigarrow Y}$, $\tau_{A\to M_2\to Y}$, and $\tau_{A\to Y}$ using appropriate outcome models. For each DGP and each PSE, we average the 1,000 estimates and then subtract the true value of the corresponding PSE, yielding what we call the *true biases* of our estimated PSEs.

For each Monte Carlo sample of a given DGP, we also fit a linear model of Y on X, A, $\mathcal{M}_k$, and $U_{M_1M_2Y}$, whose coefficient on $U_{M_1M_2Y}$ constitutes an estimate of γ_k, and a linear model of $U_{M_1M_2Y}$

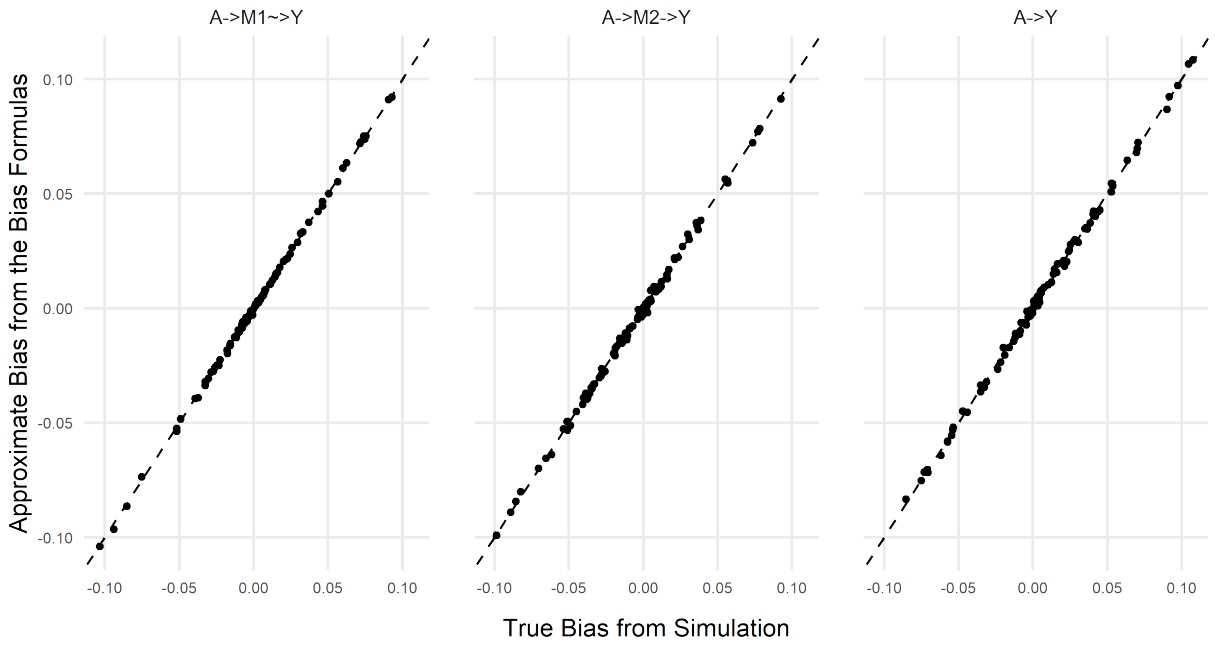


Figure E1: True Biases from Simulation versus Approximate Biases from Equations (12)-(14) in the Presence of Unobserved Confounding of the M_1-Y and M_2-Y Relationships.

on X, A, and $\mathcal{M}_k$, whose coefficient on A constitutes an estimate of η_k (despite the fact that η_k is non-constant and thus ill-defined under our DGPs). Then, for each DGP, we evaluate the "true values" of γ_k and η_k by averaging their estimates over the 1,000 samples. We then apply these true values of γ_k and η_k to the bias formulas (12)-(14), yielding what we call the *approximate biases* of our estimated PSEs.

Figure E1 shows how the approximate bias varies with the true bias across the 100 DGPs for each PSE of interest, with the 45-degree lines representing perfect alignment. We can see that the bias formulas provide excellent approximations of the true biases for all of the three PSEs.

F Empirical Results from Alternative Decompositions and Models

Equation (9) is not the only way of defining the PSEs for the causal paths $A \to Y$, $A \to M_k \rightsquigarrow Y$ ($1 \leq k \leq K-1$), and $A \to M_K \to Y$. In particular, switching the 0s and 1s in equation (9) and then flipping the signs of both sides yields

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] &= \underbrace{\mathbb{E}[Y(1) - Y\big(0, \mathcal{M}_K(1)\big)]}_{A \to Y} + \sum_{k=1}^{K} \underbrace{\mathbb{E}[Y\big(0, \mathcal{M}_k(1)\big) - Y\big(0, \mathcal{M}_{k-1}(1)\big)]}_{A \to M_k \rightsquigarrow Y} \\ &= \tau^*_{A \to Y} + \sum_{k=1}^{K} \tau^*_{A \to M_k \rightsquigarrow Y}. \end{aligned} \tag{22}$$

In the R package `paths`, we call equations (9) and (22) Type I decomposition and Type II decomposition, respectively. Figures F1 and F2 show results for our two empirical examples in the main text under both types of decomposition and three different methods for fitting the outcome models: Generalized Linear Models (GLM), Gradient Boosting Machines (GBM), and Bayesian Additive Regression Trees (BART). We can see that estimates of PSEs for these two examples are substantively similar across different specifications.

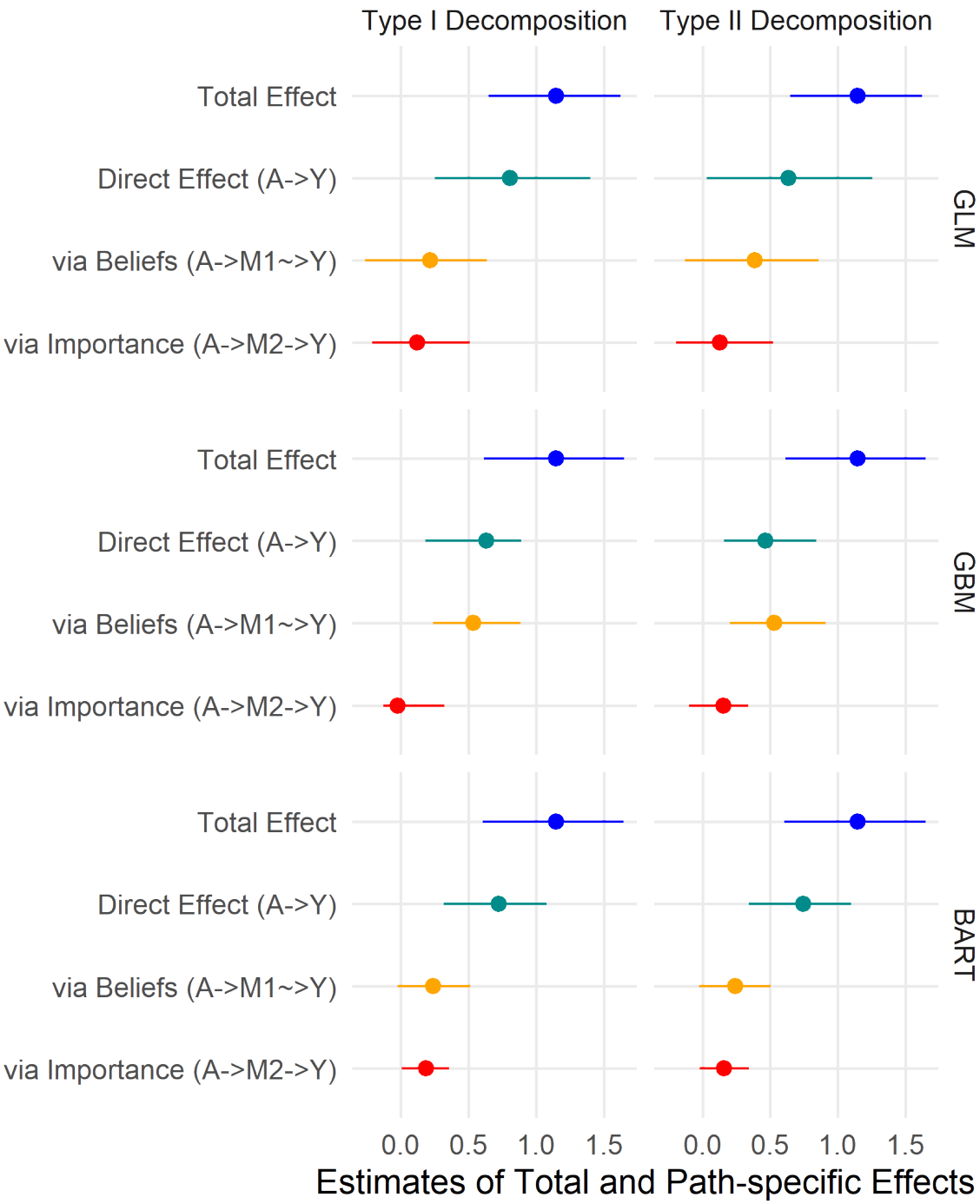


Figure F1: Alternative Estimates of Total and Path-Specific Effects of Issue Framing on Policy Support. Note: GLM = Generalized Linear Model; GBM = Gradient Boosting Machines; BART = Bayesian Additive Regression Trees. Error ranges correspond to 95% bootstrapped confidence intervals (1,000 iterations).

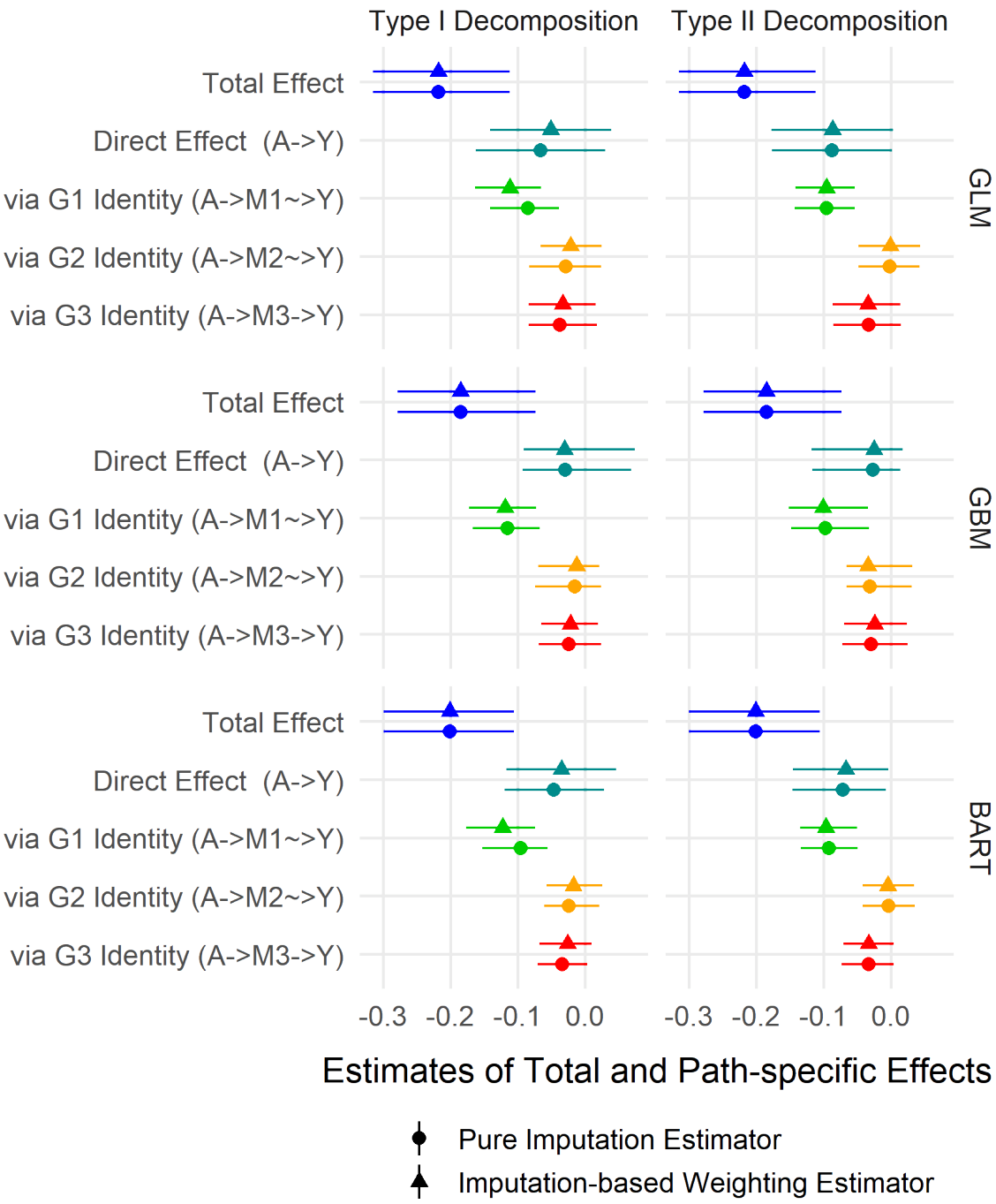


Figure F2: Alternative Estimates of Total and Path-Specific Effects of Ancestor Victimization on Support for Russia's Annexation of Crimea.

Note: GLM = Generalized Linear Model; GBM = Gradient Boosting Machines; BART = Bayesian Additive Regression Trees. Error ranges correspond to 95% bootstrapped confidence intervals (1,000 iterations).

G Morality and the Democratic Peace

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 rarely 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 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 with the assumption 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. If so, results from the authors' mediation analysis will likely be biased estimates of the mediating effects of *morality*. In fact, the authors recognized this possibility and addressed it through a "more complicated analysis" (2013, 860):

> But did people regard preventive strikes as morally wrong because they thought the target posed little threat, the attack would involve significant costs, and/or military action would fail? To answer this question, we carried out a more complicated analysis

> in which we modeled morality not only as an independent force but also as a potential consequence of the other mediators. Having estimated this more complicated model, we credited morality as a mediator only to the extent that democracy changed perceptions of morality directly. Where democracy influenced morality indirectly—by altering other mediators that, in turn, affected morality—we allocated credit to the other mediators and not to morality itself. Even with this conservative method of scoring, morality mediated more than 10% of the total effect of democracy on support for war.

Clearly, the authors aimed to isolate the mediating effect of morality above and beyond that of *threat*, *cost*, and *success,* that is, the PSE for the causal path $democracy \rightarrow morality \rightarrow support\, for\, war$, although they did not explicitly define this estimand or describe this "more complicated analysis" in much detail. In what follows, we apply our proposed methodology to estimate this PSE.

Following Tomz and Weeks (2013), we assume the mediators *threat*, *cost*, and *success* are causally prior to *morality*. To simplify our analysis, we group these mediators as a whole, forming a vector-valued mediator reflecting the respondent's beliefs about *the costs and benefits of war*. The causal mechanisms underlying the effect of democracy can then 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 M_1 include measures of the respondent's beliefs about the costs and benefits of war; the mediator M_2 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-outcome relationships may still be confounded by baseline covariates in these data.

Because treatment is randomly assigned in this study, we first estimate $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ using sample 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

Table G1: Estimates of Total and Path-Specific Effects of Democracy on Public Opposition to War.

	Estimate
Average total effect (ATE)	0.112 [0.056, 0.164]
Through costs and benefits ($A \rightarrow M_1 \rightsquigarrow Y$)	0.039 [0.014, 0.062]
Through morality ($A \rightarrow M_2 \rightarrow Y$)	0.016 [0.001, 0.033]
Direct effect ($A \rightarrow Y$)	0.058 [0.019, 0.096]

Note: Numbers in brackets represent 95% bootstrapped confidence intervals (1,000 iterations).

an autocracy and 38.8% when it is a democracy. Therefore the ATE is about 11.2%.

We estimate the PSEs for the paths $A \rightarrow Y$, $A \rightarrow M_1 \rightsquigarrow Y$, and $A \rightarrow M_2 \rightarrow Y$ using the imputation approach described in the main text. To allow for nonlinear and interaction effects, we use BART to fit the outcome models conditional on treatment, the pretreatment covariates, and varying sets of mediators (namely, $\{M_1, M_2\}$ and $\{M_1\}$). The results are shown in Table 2. We can see that taken together, perceived costs and benefits of war and perceived morality of war explain about half of the total effect. Of the mediated effect, about 70% ($0.039/(0.039 + 0.016)$) operates through the respondent's beliefs about the costs and benefits of war, and the remaining 30% appears to operate independently via the respondent's perceived morality of war. These findings are broadly consistent with those reported by Tomz and Weeks (2013). Nonetheless, it is our framework for tracing causal paths that offers a precise definition and a rigorous assessment of the mediating role of morality in the democratic peace.

H Illustration of the R package `paths`

The following R code illustrates the use of the R package `paths` for our two empirical examples.

```
# install.packages("paths")
library(paths)
library(gbm)
#########################################################
# Example 1: Issue Framing Effects
#########################################################
# variable names
x <- c("gender1", "educ1", "polint1", "ideo1", "know1", "value1")
a <- "ttt"
m1 <- c("W1", "W2")
m2 <- c("M1","M2","M3","M4","M5")
y <- "Y"
m <- list(m1, m2)
# formulas
form_m0 <- as.formula(paste0(y, "~", a))
form_m1 <- as.formula(paste0(y, "~", paste0(c(x, a, m1), collapse = "+")))
form_m2 <- as.formula(paste0(y, "~", paste0(c(x, a, m1, m2), collapse = "+")))
# baseline model for overall treatment effect
lm_m0 <- lm(form_m0, data = welfare)
# GBM outcome models
gbm_m1 <- gbm(form_m1, data = welfare, distribution = "gaussian",
              interaction.depth = 3)
gbm_m2 <- gbm(form_m2, data = welfare, distribution = "gaussian",
```

```
              interaction.depth = 3)
gbm_ymodels <- list(lm_m0, gbm_m1, gbm_m2)
# causal paths analysis
welfare_paths <- paths(a, y, m, models = gbm_ymodels,
                       data = welfare, nboot = 250)
# summarize results
summary(welfare_paths)
#######################################################
# Example 2: The Legacy of Political Violence
#######################################################
# K=3 causally ordered mediators
m1 <- c("trust_g1", "victim_g1", "fear_g1")
m2 <- c("trust_g2", "victim_g2", "fear_g2")
m3 <- c("trust_g3", "victim_g3", "fear_g3")
mediators <- list(m1, m2, m3)
# outcome model formulas
formula_m0 <- annex ~ kulak + prosoviet_pre + religiosity_pre + land_pre +
  orchard_pre + animals_pre + carriage_pre + otherprop_pre + violence
formula_m1 <- update(formula_m0,    ~ . + trust_g1 + victim_g1 + fear_g1)
formula_m2 <- update(formula_m1,    ~ . + trust_g2 + victim_g2 + fear_g2)
formula_m3 <- update(formula_m2,    ~ . + trust_g3 + victim_g3 + fear_g3)
# outcome models
gbm_m0 <- gbm(formula_m0, data = tatar, distribution = "bernoulli",
              interaction.depth = 3)
gbm_m1 <- gbm(formula_m1, data = tatar, distribution = "bernoulli",
              interaction.depth = 3)
gbm_m2 <- gbm(formula_m2, data = tatar, distribution = "bernoulli",
```

```
              interaction.depth = 3)
gbm_m3 <- gbm(formula_m3, data = tatar, distribution = "bernoulli",
              interaction.depth = 3)
gbm_ymodels <- list(gbm_m0, gbm_m1, gbm_m2, gbm_m3)
# causal paths analysis using gbm
tatar_paths <- paths(a = "violence", y = "annex", m = mediators,
  gbm_ymodels, data = tatar, nboot = 250)
# summarize results
summary(tatar_paths)
# propensity score model via gbm
formula_ps <- violence ~ kulak + prosoviet_pre + religiosity_pre + land_pre +
  orchard_pre + animals_pre + carriage_pre + otherprop_pre
gbm_ps <- gbm(formula_ps, data = tatar, distribution = "bernoulli",
              interaction.depth = 3)
# causal paths analysis using both the pure imputation estimator and
# the imputation-based weighting estimator
tatar_paths2 <- paths(a = "violence", y = "annex", m = mediators,
  ps_model = gbm_ps, gbm_ymodels, data = tatar, nboot = 250)
# plotting PSEs
plot(tatar_paths2, mediator_names = c("G1 identity", "G2 identity", "G3 identity"),
     estimator = "both")
# sensitivity analysis for the path-specific effect via M1
sens_paths <- sens(tatar_paths, confounded = "M1", estimand = "via M1",
  gamma_values = - seq(0, 0.5, 0.002), eta_values = seq(-0.5, 0.5, 0.002))
plot(sens_paths)
```